VIDEO CLASSIFICATION USING SPATIAL-TEMPORAL FEATURES AND PCA

Li-Qun Xu and Yongmin Li

BTexact Technologies, Adastral Park, Ipswich IP5 2GT, UK
Email: (li-qun.xu, yongmin.li)@bt.com

ABSTRACT

We investigate the problem of automated video classification by analysing the low-level audio-visual signal patterns along the time course in a holistic manner. Five popular TV broadcast genre are studied including sports, cartoon, news, commercial and music. A novel statistically based approach is proposed comprising two important ingredients designed for implicit semantic content characterisation and class identities modelling. First, a spatial-temporal audio-visual "super" feature vector is computed, capturing crucial clip-level video structure information inherent in a video genre. Second, the feature vector is further processed using Principal Component Analysis to reduce the spatial-temporal redundancy while exploiting the correlations between feature elements, which give rise to a compact representation for effective probabilistic modelling of each video genre. Extensive experiments are conducted assessing various aspects of the approach and their influence on the overall system performance.

1. INTRODUCTION

Automated video genre classification from an input video stream is becoming of increased significance in multimedia information processing. With the advent of digital TV broadcasts of several hundred channels and the availability of large digital video libraries, there are compelling needs for the provision of such a solution to help end users search, choose or verify a desired programme based on the semantic content thereof. On the one hand, such a solution, when provided as a real-time component in a set-top box or car radio, can be used to automatically select a desired broadcast programme for a user. On the other hand, as part of a video indexing and retrieval system it can automatically classify an incoming video file according to its content, thus providing some of the desired metadata information and enabling fast video browsing and retrieval.

Conventional approaches for video classification tend to adopt a step-by-step heuristic strategy [3][5]. They usually proceed by first extracting certain low-level visual and/or audio features, from which analysis is made to build the so-called intermediate-level signatures, style attributes etc that are likely to be specific to certain video class. Finally the class label is hypothesised and verified using a precompiled knowledge-based heuristic rules or some learning methods. In the same fashion but with more insight into the knowledge of a film making, Truong et al. [12] analysed a set of computational features derived from cinematic editing effects, motion and colours in videos. Experiments show that the trends of these features for different genre are distinctive, and good for video classification.

Based on the observation that most TV programs focus on human activities and human face is of primary interest, yet another recent approach [2] tried to extract face trajectories and track text appeared on the screen. The changes in these tracked "objects" are then modelled using statistical/dynamic models with a view to differentiating individual video genre identities.

Recently, a statistically based video genre modelling approach has been introduced [9], focusing on *direct* analysis of the relationship between the probabilistic distribution of low-level features and the associated genre identity. The Gaussian Mixture Model (GMM) was used to model the class-based probabilistic distribution of audio and/or visual feature vectors in a high-dimensional feature space. These features are computed directly from successive short segments of audio and/or visual signals of a video sequence, accounting for, e.g. 46 ms audio information or 640 ms visual information [8][9] albeit in a simplified representation, respectively. No explicit or sensible temporal information of the video stream at *a segmental level* is incorporated except that the acoustic feature used has built into it some short-term transitional changes. The assumption that the successive feature vectors from the source video sequence are largely independent of each other is only a simplification in most cases. Besides, a problem with the GMM used in this manner is the "curse of dimensionality", as it demands a large amount of *training data* to handle data in a very high dimensional space beyond a few tens.

Meanwhile, subspace data analysis methods have been used widely in image processing [11][4], and recently in the area of content-based video indexing and retrieval [10]. Sahouria and Zakhor [10] used PCA to reduce the dimensionality of features (colour histograms, motion vectors) of video frames for the purpose of detecting scene changes and characterising an entire sport video by motion activities, respectively.

In this paper we propose to combine the strength of the statistically based signal modelling approach and the effective redundancy reduction capability of PCA for the purpose of video classification. In section 2, the computation of spatial-temporal audio-visual features inherent in a video class is described. In section 3, the system modules for video content modelling and classification are discussed, which is followed in Section 4 with extensive experiments on aspects of the system performance and results of five TV broadcast genre classifications. The paper concludes in Section 5.

2. FEATURE EXTRACTION

In this section we outline our method how to acquire compact spatial-temporal feature vectors that potentially encapsulate the generic semantic content of a video.

2.1. Low-level audio/visual features

For the video domain under study there is a rich set of descriptors that we can exploit for content / semantics characterisation. The feature vectors can assume either a visual mode or an acoustic (audio) mode, or indeed the combined audio-visual mode. Figure 1 shows the process involved to compute a spatial-temporal audio-visual "super" feature vector.

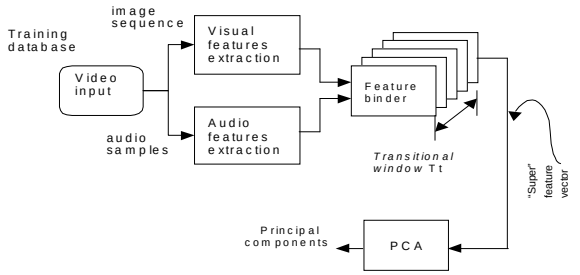

Figure 1: The computation of "super" feature vectors and subsequent condensation via PCA.

Audio features: Various audio-based features, on a short-term (tens of milliseconds) or clip (seconds) scale, have been suggested for video content analysis, either in an audio only or in a joint audio-visual mode [5][7]. In the current work we set to use only the short-term spectrum of audio signals - the 14 Mel-frequency cepstrum coefficient (MFCC) features. It is important to note that unlike the work e.g. in [9] we made no efforts in computing short-term *dynamic* features, or characterising a meaningful longer time variations at this stage. However, it will soon be clear that the concatenation of these raw features over a defined transitional window along the time course to generate a very high-dimensional "super" feature vector and the subsequent PCA analysis applied have captured properly the desired dynamic properties in a concise manner.

Visual features: We set to use the MPEG-7 compliant low-level content descriptors [6] on colour and textures as part of the visual content description feature, which include Scalable Colour (16), Colour Layout (12) and Homogenous Texture (32). Also, the mean and standard deviation (2 parameters) computed from the magnitudes of MPEG video motion vectors are added. This gives rise to a 62-dimensional feature vector in describing the imaging content for each video frame.

Audio-visual features: Either the visual or audio features discussed above can be used alone for video content analysis and classification, though it is more beneficial to use them in conjunction, taking advantage of the complementary and richer expressive and discriminating power of the combined audio-visual feature. As the visual and audio signals are captured at different rate, e.g., 25 fps for visual signal in a PAL video, and 22.5kHz for audio signal, to synchronise the audio/visual features, a *transitional* window with a length, e.g., $T_t = 1000$ ms, is adopted. All the audio/visual features falling within this transitional window are alternatively concatenated, resulting in a "super" temporal feature vector characterising each one second long video clip. It is important to note that, as well as synchronising the audio/visual information; the transitional window also captures the short-term and clip-level dynamic information of the video, which is more important to embedding the semantics of the content.

2.2. Analysing audio-visual feature using PCA

One problem with the concatenated "super" feature vector over the transitional window is its very high dimensionality. It reaches 2138 following the figures given above (62x25 for visual and 14x42 for audio). Apparently it is not feasible to perform any computation and analysis within such a high dimensional space. Furthermore, there exists considerable amount of spatial redundancy in the feature vector, e.g. correlation between different audio/visual features, as well as the temporal redundancy exhibited in adjacent audio/video frames.

To obtain a compact and low-dimensional representation of the super feature vector, we need to explore valid subspace. For such purpose PCA is performed on a training database, which includes the video data of all classes to be identified. Given N d-dimensional "super" feature vectors $\{\mathbf{x}_i, i = 1,2,\cdots,N\}$, the mean vector and covariance matrix are given, respectively, by

$$\bar{\mathbf{i}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i \quad \text{and} \quad \mathbf{C} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \bar{\mathbf{i}})(\mathbf{x}_i - \bar{\mathbf{i}})^T .$$

Then the principal components assume the first P significant eigenvectors of $\mathbf{C}$, *i.e.* $\{\mathbf{v}_i, i = 1,2,\cdots,P\}$. By constructing the eigenmatrix $\mathbf{U} = [\mathbf{v}_1\mathbf{v}_2\cdots\mathbf{v}_P]$ of $d \times P$ dimension, an arbitrary d-dimensional original feature vector $\mathbf{x}$ can be represented as a new P-dimensional compact vector, $\mathbf{y} = \mathbf{U}^T(\mathbf{x} - \bar{\mathbf{i}})$, with $P << N$ and $P << d$.

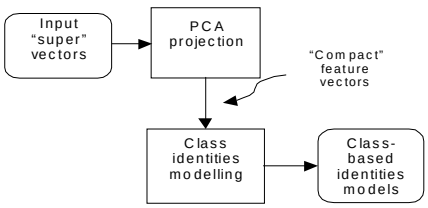

Figure 2: The learning of class-based identity models.

3. VIDEO MODELLING AND CLASSIFICATION

3.1. The genre-based identity modelling

The schematic diagram of the video class-identities learning module is shown in Figure 2. The input sequence of training samples ("super" feature vectors) computed over a transitional window T_t is first subject to a PCA projection to extract the low-dimensional "compact" feature vectors. The statistical distribution of these feature vectors for each intended genre is then modelled using an appropriate modelling technique.

In the current studies we choose to use the GMM, the probability of a feature vector $\mathbf{y}$ belonging to a pre-defined video class is expressed as $p(\mathbf{y}) = \sum_{j=1}^{M} p(\mathbf{y} \mid j)P(j)$, where M is the number of mixture components, $P(j)$ is the prior of the jth component, and $p(\mathbf{y}|j)$ is the Gaussian density function of the jth mixture component.

The Expectation-Maximisation (EM) algorithm can be adopted to estimate the parameters of a GMM, which iterates between E-step and M-step until no significant improvement for

the overall likelihood of the all the training data. A good reference of the training algorithm can be found in [1].

3.2. The temporal video classification module

Figure 3 shows the diagram of the video genre classification module. A test video stream first undergoes the same "super" feature vector extraction process to produce a sequence of spatial-temporal audio-visual feature vectors. The sequence of features within a pre-defined *decision* window T_d is projected successively onto the PCA space, resulting in the compact feature vectors. This sequence of PCA vectors is subsequently fed to the class-based models learned before; the class model that matches the sequence best in terms of a metric is declared to be the class label of the current test video stream falling within the decision window. The choice of an appropriate similarity measure depends on the class-based identities models adopted.

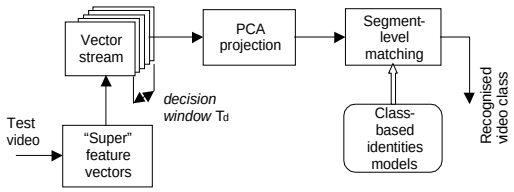

Figure 3: The temporal video genre classification module

As shown in Figure 3, one important parameter of the module is the *decision window* length T_d, which defines the elapsed time when an answer is required of the genre of the video programme being queried. More discussions on the impact of this on system performance will be given in experiments Section 4.2.

4. EXPERIMENTAL STUDIES

We have applied the approach described above to the problem of video genre classification. Five popular TV broadcast genres are considered including sports, cartoon, news, commercial, and music. There is a total of 1 hour of recordings per genre, which are continuous sequences of typically 5 minutes each, a little less for commercials and music genre. For the experiments the data is split into 0.5 hour for test and train, respectively. Note that the recordings are not validated, labelled or edited in any manner other than to confirm that they do belong to the said genre for the interval of recording. This means that sub-sequences can contain material that could well be classified as other genre, e.g. a football sub-sequence within a news sequence. This is less favourable to our assessment results but reflects the real-world application scenarios.

4.1. The dimension of the PCA vectors

By projecting a concatenated "super" feature vector onto the pre-trained PCA space, we obtain a compact PCA feature vector. It is then important to know how many principal components that is proper for the compact feature in order to represent the essential structure information for video classification.

We experiment with *P* the dimension of the PCA vectors, from 10 to 100 at an increment of 10. The classification accuracy on the testing data set is shown in Figure 5, using a different decision window length $T_d = 10$, 20, 30 and 40 seconds. It is interesting to note that the performance shows only a marginal improvement with the increase of the PCA dimension. In other words, good classification accuracy can be achieved with a considerably low dimension of PCA vectors.

4.2. The decision window length T_d

It has been shown that, by using the transitional window, the dynamic variations of a video can be captured up to seconds in time. It is noted that, by acquiring the transitional features, we intend to formulate a minimal representation of the semantic video content, though it is not possible to obtain a perfect discriminatory feature vector for video classification at this level. Thus judging by the likelihood of a GMM w.r.t. individual transitional feature vectors, a segment of a cartoon video may look like a music video. However, if we leave the classifiers to cover for a relatively longer time, they would be more likely to make a right decision. In fact this is the significance of the decision window. We have experimented over the whole test data set with the decision window length ranging from 2 to 40 seconds at 2 seconds interval. The results given in Figure 5 show a constant increase in classification accuracy with a longer decision window length. An average correct rate of about 86.5% is achieved at $T_d = 40$ seconds.

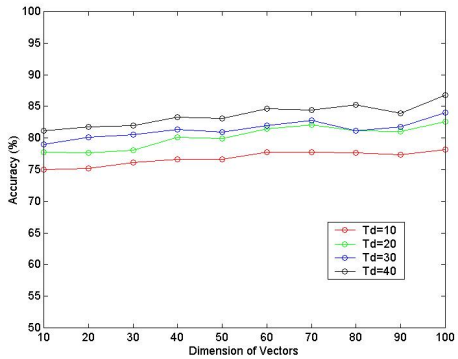

Figure 4: The impact of the dimension of the PCA vectors on overall genre classification accuracy with the choice of different decision window length.

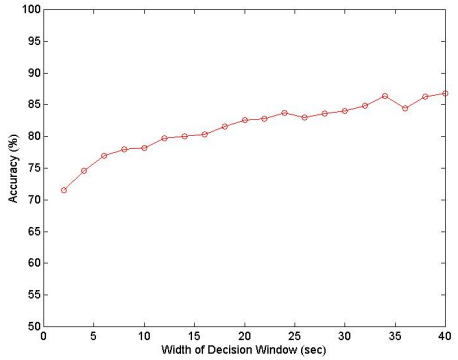

Figure 5: The impact of the decision window length T_d (in second) on overall video genre classification performance.

4.3. Model complexity

The model complexity of a GMM, or the number of mixture Gaussian components, is an important factor for a model's learning and generalisation capability. We examine the performance of video classification over the whole test data set with different choices of component number (from M=1 to 19 with an increment of 2). The results are shown in Figure 6. It can be observed that an approximately constant accuracy is achieved when the number of components is above 3. This demonstrates that, for the video genre classification problem of this scale (5 genres), a considerably small number of mixture components, for example, 3 to 10, can be sufficient for a satisfactory performance.

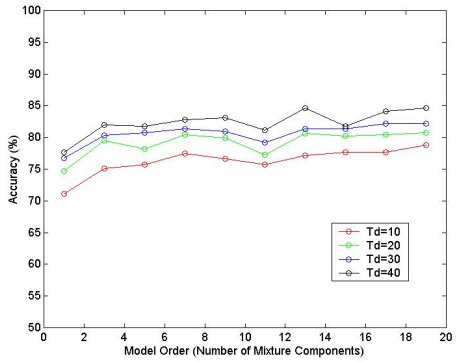

Figure 6: The impact of GMM complexity on the classification performance under different T_d .

4.4. Classification results on individual genres

Finally, to demonstrate the inter-dependence among semantic contents belonging to these five video genres, we compute the confusion matrix of the classification results. Table 1 shows one of the matrices, given that $T_t = 230$ ms, M=9, P=100, and $T_d = 40$ seconds. It is interesting to note that the sports genre has the best classification results and least confusion with others owing to its specific characteristics. 17.1% of the cartoon videos have been misclassified as commercial, which reflects the similarity between the two genres. Another important reason is that some commercial videos were collected from cartoon channels, which advertise children's toys, food etc using animated production. As expected, 15.8% of the music videos have been misclassified as commercials since there is usually music background in commercial videos, and visual scenes are experiencing dynamic changes.

Table 1: The confusion matrix for the five video genres

Genre	Sports	Cart.	News	Com.	Music
Sports	96.7	0.0	3.3	0.0	0.0
Cart.	0.0	79.5	0.0	17.1	3.4
News	0.0	4.6	87.5	4.5	3.4
Com.	0.0	6.8	3.5	89.7	0.0
Music	0.0	0.0	2.6	15.8	81.6

5. CONCLUSION

We have presented in this paper a novel approach to automatic video content categorisation at the highest semantic level. Three key contributions have been made. First, to integrate audio-visual features for content description in an effortless and natural concatenation; Second, to embed proper temporal dynamics within a transitional window to obtain a segment level "super" feature vector; Third, to apply PCA to remove the spatial-temporal redundancy in the low-level audio-visual descriptors, deriving an effective compact feature vector. The use of statistical models to learn the properties of a video genre from these feature vectors is then becoming a standard practice. A video database of non-edited TV broadcast programme containing five popular genres namely sports, cartoon, news, commercial, and music are tested. An average correct classification rate of 86.5% has been achieved, given a 40 seconds decision window. Future work will be devoted to studying an even large collection of video database, investigating open-set video genre verification, and semantic event detection for particular genres e.g. sports.

REFERENCES

[1] C.M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, 1995.

[2] N. Dimitrova, L. Agnihotri and G. Wei, "Video classification using object tracking," *International Journal of Image and Graphics*, Vol.1, No.3, July 2001.

[3] S. Fischer, R. Lienhart, and W. Effelsberg, "Automatic recognition of film genres," *Proc. of ACM Multimedia'*95.

[4] Y. Li et al. "Recognising trajectories of facial identities using Kernel Discriminant Analysis," *Proceedings of British Machine Vision Conference*, pp 613-622, September 2001.

[5] Z. Liu, Y. Wang, and T. Chen, "Audio feature extraction and analysis for scene segmentation and classification," *Journal of VLSI Signal Processing Systems*, pp 61-79, October 1998.

[6] B. S. Manjunath, J.-R. Ohm, V. V. Vasudevan, and A. Yamada, Color and texture descriptors, *IEEE Trans. on Circuits and Systems for Video Technology*, Vol.11, No. 6, June 2001.

[7] S. Pfeiffer, S. Fischer and W. Effelsberg, "Automatic audio content analysis," *Proceedings of 4th ACM Multimedia Conference*, Nov. 1996, pp 21-30.

[8] M.J. Roach, J.S.D. Mason, and M. Pawlewski, "Video genre classification using dynamics," *Proc of ICASSP'2001*.

[9] M.J. Roach, J.S.D. Mason, L.-Q. Xu, "Classification of non-edited broadcast video using holistic low-level features**,"** *Proceedings of* IWDC'2002, Capri, Italy, Sept. 2002.

[10] E. Sahouria and A. Zakhor, "Content analysis of video using principal components," *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 9, No. 8, Dec 1999.

[11] D.L. Swets and J. Weng, "Using discriminant eigenfeatures for image retrieval," *IEEE Trans. on PAMI*, **8**(8), pp 831-836, August 1996.

[12] B.T. Truong and C. Dorai, "Automatic genre identification for content-based video categorization," *Proc. of ICPR' 2000*, pp 230-33.