VideoCube: a novel tool for video mining and classification

Jia-Yu Pan and Christos Faloutsos

Computer Science Department, Carnegie Mellon University

Abstract. We propose a new tool to classify a video clip into one of n given classes (e.g., "news", "commercials", etc). The first novelty of our approach is a method to *automatically* derive a "vocabulary" from each class of video clips, using the powerful method of "Independent Component Analysis" (ICA). Second, the method is *unified* which works on both video and audio information and gives vocabulary describes not only the still images, but also the motion, as well as the audio part. Furthermore, this vocabulary is *natural* that it is closely related to human perceptual processing. More specifically, every class of video clips gives a list of "basis functions", which can compress its members very well.

Once we represent video clips in "vocabularies", we can do classification and pattern discovery. For the classification of a video clip, we propose to use compression: we test which of the "vocabularies" can compress the video clip best, and we assign it to the corresponding class.

For data mining, we inspect the basis functions of each video genre class, and thus we can figure out whether the given class has, e.g., fast motions/transitions, more harmonic audio, etc. Experiments on real data of 62 news and 43 commercials show that our method achieves overall ≈81% accuracy.

1 Introduction

Video classification is useful for organizing and segmenting video clips of digital video libraries [25]. It also facilitates browsing and content-based retrieval that only certain genres of video clips are returned to the users according to their demands.

The process of classifying video clips into several predefined genre classes (such as news and commercials) usually involves two steps: First, building model of each genre class from training video clips of that class. Second, classifying video clips of unknown genre by comparing them to class models. The design decisions in the first step includes: *How do we represent a video clip? What kind of features are used?* At the second step we have to decide *how do we determine which class a video clip belongs to? What kind of similarity function is used to determine the closeness between a video clip and a class?*

In this paper we propose a novel video classification method with the following characteristics:

- Feature extraction is done *automatically*, as opposed to handpicking the features.
- The method is *unified* that it deals with both visual and auditory information, and captures both spatial and temporal characteristics.
- The extracted features are "*natural*", in the sense that they are closely related to the human perceptual processing.

This paper is organized as follow: In section 2, we give a brief survey on the previous work on video classification. In section 3, we introduce the independence component analysis and its relationship with human perceptual processing. In section 4, we describe our proposed method for video classification. Section 5 shows the experimental results and gives discussions. We conclude in section 6.

2 Survey

Previous work had tried different visual/auditory features derived from video clips for classification. There are studies based on pixel-domain visual information (color histograms) [14, 26], and transform (compressed) domain visual information [6]. Other kinds of meta-data, such as motion vectors [20] and faces [4], have also been used. On the other hand, time-domain and frequency-domain auditory features have also been used on classifying video genre [15, 19].Recent studies [5, 22] also combined visual and audio features for video classification. One common problem is that features are usually hand-picked, whose applicability relies on the experience of the researchers, and their understanding on the deploying domain.

Another orthogonal design issue on feature selection is whether the features are global or local [21]. With increased complexity, the latter provides class models that are more accurate and with higher representative power [27].

Several approaches have been used to construct representation of genre classes, namely, statistical (Gaussian) modeling [19], hidden Markov model [4, 17] and other general classifier representations, such as decision trees or neural networks.

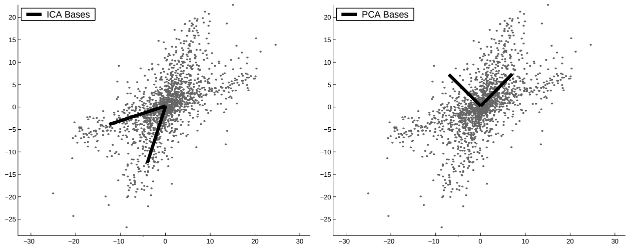
Previous results on video classification used different features and modeling approaches to classify different sets of genre classes:

- Roach'01 [19] used audio features on 5 classes: sport, cartoon, news, commercial and music. They achieved ≈76% accuracy on classification.
- Truong'00 [23] used visual statics, dynamics, and editing features on 5 classes: sport, cartoon, news, commercials and music videos. They reported an accuracy of ≈80%.
- Liu'98 [15] used audio features on 3 classes: commercial, report(news and weather) and sport. Their classification accuracy is 93%.

It is difficult to compare the performance of these results, because of the different sets of genre classes and the different collections of video clips they used.

3 Independent component analysis

Independent component analysis (ICA) [9, 10], like principal component analysis (PCA), has been proven a useful tool for finding structure in data. Both techniques represent the given multivariate data set by a linear coordinate system. Unlike PCA, which gives orthogonal coordinates (or bases) and is based on the second-order statistics (covariance) of the data, ICA is more generalized and gives non-orthogonal bases determined by the second- and higher-order statistics of the data set [12]. Figure 1 demonstrates the difference between ICA and PCA. In this case, ICA captures the non-orthogonal underlying components of the data set which PCA misses.



(a) ICA bases

(b) PCA bases

Fig. 1. ICA bases and PCA bases PCA fails to capture the underlied components of the data set, while ICA does.

Specifically, let an observed data point be $\mathbf{x} = (x_1, x_2, \ldots, x_m)^T$ a zero-mean m-dimensional random vector. Then under the assumption of ICA,

$$\mathbf{x} = \mathbf{A}\mathbf{s},$$

where $\mathbf{A}$ is a m-by-n matrix (usually $n < m$), and $\mathbf{s} = (s_1, s_2, \ldots, s_n)^T$ is a n-dimensional vector of source variables. s_i's are called the *independent components*. $\mathbf{A}$ is called the *mixing matrix* and its columns are the *bases* that are used to construct the observed $\mathbf{x}$. Given a set of data points $\mathbf{x}$'s, ICA finds the matrix $\mathbf{A}$ and the vector $\mathbf{s}$ which satisfies the above equation, subjected to some optimization constraints on $\mathbf{s}$ such as maximizing non-gaussianity (or kurtosis, or independence of s_i's) [10].

3.1 ICA and human perceptual processing

The mechanism of how human perceptual system represents things we see and hear has long been an intriguing topic. Barlow [1] proposed that neurons perform redundancy reduction and make up a factorial code for the input data, i.e. a representation with independent components, which is supported by recent studies on efficient natural image encoding, either from the direction of sparse coding [18] (maximizing redundancy reduction), or from the direction of independent components [3]. These experimental results show that human perceptual processing is based on *independent features which encode the input signals efficiently.* The independent components, either of visual or auditory signals, are generally filters resembles wavelet (Gabor) filters, which are oriented, localized in space (or time), bandpass in the frequency domain [3, 13] and resembles the receptive fields of neurons in mammals' cortex.

Analysis have also been extended to the spatial-temporal domain [24] where the independent components of natural image sequences (video clips) and color images [7] are examined. The results are qualitatively similar to those of the static and gray-scale images and are again closely related to the results from human perceptual studies.

Due to the fact that human perceptual processing based on independent components of signals which ICA are able to compute, ICA has been used in applications such as face recognition [2], object modeling [27] and speech recognition [11], and achieved better or comparable performance than conventional approaches based on hand-picked features.

4 Proposed Method

Next, we describe our proposed method, including an unified, automatic feature extraction method and the classification process based on these features. We will first introduce the features we extracted, which are obtained from ICA, how effective these features are in representing the video genres, and then how they are used on classification.

4.1 Features

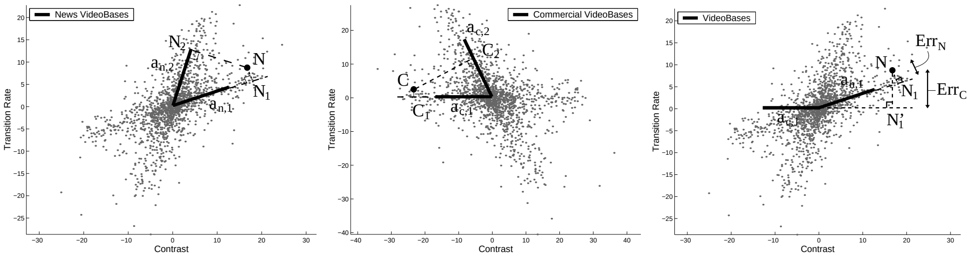
We want to extracted visual and auditory features that can capture both spatial and temporal characteristics of a video genre. Visual features are derived based on pixel information. We said two pixels are **spatially adjacent** if they are adjacent to each other on the same video frame, and two pixels are **temporally adjacent** if they are at the same position of adjacent frames. For example, pixels (2,1) and (2,2) on frame 1 are spatially adjacent, and pixel (2,1) on frame 1 and pixel (2,1) on frame 2 are temporally adjacent. To consider the spatial and temporal information at once, spatially and temporally adjacent pixels of a video clip are grouped into cubes as basic processing units.

Definition 1. ***(VideoCube)*** *We denote the pixel located at position (x,y) on frame t as pixel (x,y,t). A* ***n-by-n-by-n cube****, located at pixel (x,y,t) consists all pixels (i,j,k) where i=x,...,(x+n-1), and j=y,...,(y+n-1), and k=t,...,(t+n-1). We called such cubes* ***VideoCubes****, which incorporate both spatial and temporal pixel information of the video frames.*

Definition 2. ***(VideoBasis) VideoBases*** *are spatial-temporal features of a genre class extracted by performing ICA on a set of n-by-n-by-n VideoCubes sampled randomly from the training video clips. They are effectively the basis functions obtained from ICA. VideoBases can be considered as the commonly used* ***vocabulary*** *for describing the visual content of a genre. In other words, clips of the news genre share a set of VideoBases which is different from the set shared by commercial clips.*

Figure 5 shows several VideoBases of news stories and commercials. These basis functions are similar to moving Gabor filters [24]. VideoBases of commercials have greater chopping effects along the time axis, which is because that more activities happen inside commercials.

With the vocabulary and using an idea similar to the vector space model in the traditional information retrieval, the visual content of a clip is rendered as a linear combination of the VideoBases of its genre. Figure 2 illustrates the intuition behind the idea. Two fictitious data sets representing VideoCubes (samples of visual content) from news and commercial clips are shown, where each point is a VideoCube. Every VideoCube can be represented as a linear combination of VideoBases. The combination is most efficient (most coefficients are small and closed to 0) when a VideoCube is represented (encoded) by the correct VideoBases, i.e., when a news VideoCube is encoded by the news VideoBases. This is due to the sparseness coding property of ICA, by which our VideoBases are computed.



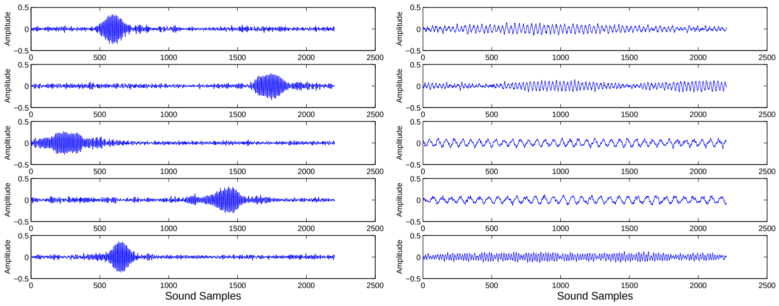
(a) News (b) Commercial (c) Classification

Fig. 2. Representation of news and commercial clips **N** and **C** are VideoCubes of a news clip and a commercial clip, respectively. $\mathbf{a_{N_1}}$ and $\mathbf{a_{N_2}}$ are VideoBases of news, and $\mathbf{c_{N_1}}$ and $\mathbf{c_{N_2}}$ are those of commercial. (a) $\mathbf{N} = \mathbf{N_1} + \mathbf{N_2} = s_{N_1}\mathbf{a_{N_1}} + s_{N_2}\mathbf{a_{N_2}}$ and (b) $\mathbf{C} = \mathbf{C_1} + \mathbf{C_2} = s_{C_1}\mathbf{a_{C_1}} + s_{C_2}\mathbf{a_{C_2}}$. ($s_{N_i}$'s and s_{N_i}'s are weighting scalars.) (c) Only one basis from each genre ($\mathbf{a_{N_1}}$ and $\mathbf{a_{C_1}}$) are kept. VideoCube N is best represented by $\mathbf{a_{N_1}}$. That is, reconstruction error Err_N is less than that when represented by $\mathbf{a_{C_1}}$ (Err_C).

Similar to the visual part, we use ICA and extract auditory vocabulary of different genres.

Definition 3. ***(AudioBasis) AudioBases*** *are auditory features of a genre class extracted by performing ICA on a set of audio segments of duration d seconds, randomly sampled from the training clips. In the following, when both AudioBasis and VideoBasis are mentioned together, they are called* ***AV-Bases****.*

Figure 3 shows some of the AudioBases of news and commercial clips. The different appearances of the AudioBases of the two genres reveal the different characteristics between news and commercials, suggesting its capability of classifying the two genres.



(a) AudioBases (news) (b) AudioBases (commercials)

Fig. 3. AudioBases of news and commercial clips AudioBases reveal the characteristic differences between the two genres.

ICA provides a way to automatically extract features which, as shown later, capture the essentials of a genre. As opposed to other feature extraction studies [5, 16], which require intensive human involvement in understanding the data set and trial-and-error processes to find features with good expressive power.

4.2 Classification

Our classification method is based on the idea of "compression", that is, the set of AV-Bases which compresses a video clip most efficiently are that of the genre class to which the clip belongs. To measure the "goodness" of compression, we compare the reconstruction errors of a video clip when it is encoded by the AV-Bases of different genre classes. The set of AV-Bases which gives the smallest reconstruction error is the one that compresses the clip best, and the clip is classified as the genre associated with that set. Figure 2(c) illustrates the idea of our classification approach. The existence of the reconstruction error comes from the fact that we are using fewer AV-Bases than the dimension of the data samples, and therefore can not fully reconstruct the content of a clip.

Figure 4 gives the algorithm (**VCube**) using the VideoBases of the genre classes to classify a video clip. Using AudioBases to classify clips based on their audio information is similar, with the processing unit changes from VideoCubes to audio segments of duration d seconds (e.g., d=0.5).

```
Input: Video track of a clip; VideoBases of $G$ genre classes
Output: Genre class of the clip

1. Pixels of I-frames are collected, and are formed into non-overlapped n-by-n-by-n VideoCubes. (e.g., n=12)
2. Initialize the sum of reconstruction error of class $c_i$, $sumErr_i$, $i = 1, \dots, G$.
3. For each VideoCubes ($vc$),
    3.1 Encode $vc$ with the VideoBases of each genre class ($c_i, i = 1, \dots, G$).
    3.2 Compute the reconstruction error ($err_i, i = 1, \dots, G$).
    3.3 $sumErr_i = sumErr_i + err_i$
4. Return $c_k$, where $k = \underset{i}{argmin}\ sumErr_i$.
```

Fig. 4. Classification algorithm (VCube)

To summarize, our proposed method has the following characteristics:

- Feature (VideoBases and AudioBases) extraction is automatic and considers both spatial and temporal properties at once.
- The proposed feature extraction method is an unified approach, which is applicable to extract either visual or auditory features.
- VideoBases and AudioBases reveals the essential characteristics of a genre class, and are closely related to the neural signals used in the human perceptual process.

4.3 Capture local activities

A video clip with intense local activities will be mistakenly considered to have activities throughout the scene, if scene changes at different positions are not quantified separately. To better capture the local activities in the scene, we model the local scene changes separately, rather than summing all activities in the video frames as a global effect.

We divide the video frame into 9 rectangular regions of equal size (in 3-by-3 arrangement) and extract the VideoBases (visual vocabulary) for each region separately. In other words, each genre has now 9 sets of VideoBases, one for each region. For news clips, activities are more concentrated on the certain regions, while activities are spread out among the regions for commercial clips.

The classification algorithm **VCube_P** ("P" stands for "Position") which consider the local effects at different position is similar to algorithm **VCube** (Figure 4). The only difference is the step 3.1 and is modified as

```
    3.1 Encode $vc$ with the VideoBases ($VB_{i,r}$) of each genre class ($c_i$) and each region $r$, where $i = 1, \dots, G$, and $r = 1, \dots, 9$.
```

5 Experimental Result

In this section, we describe our experiments on classifying news and commercial video clips. Specifically, we discuss the property of the ICA bases of the two classes and report our classification result.

5.1 Data Set

We divide our collection of video clips into two sets: training set and testing set. The training set contains 8 news clips and 6 commercial clips, each is about 30 seconds long. The testing set contains 62 news clips and 43 commercial clips, each is about 18 seconds long. The training set are used for constructing VideoBases and AudioBases of the two classes. We use the FastICA [8] package from the Helsinki University of Technology for ICA processing.

Video frames are divided into 9 rectangular regions of equal size, in a 3-by-3 matrix-like arrangement. As a result, each genre class has 9 sets of VideoBases (one for each of the 9 regions) and 1 set of AudioBases. VideoBases are derived from 12-by-12-by-12 randomly sampled VideoCubes from the training clips. The color information of each pixel is separated into 3 (Red,Green,Blue) channels. In our experience, VideoBases on any of the three channels give us similar result and in the following we will use only the VideoBases from the channel Red.

AudioBases are derived from 0.5-second long audio segments sampled randomly from the training clips. The audio segments are down-sampled by a factor of 10 as a trade-off between the data size and coverage period. This is because we want to extract auditory features of longer periods to better capture auditory characteristics, but under the encoding frequency of 44.1kHz, a 0.5-second long audio segment without down-sampling has too much data and will hinder the subsequent computation.

The number of video clips in the training set is not an important factor on the quality of the derived ICA bases. What really matters is the amount of data samples used for extracting features. In our training phase, 10,000 12x12x12 VideoCubes and 7,000 0.5-second audio segments are used, and 160 VideoBases (per region) and 60 AudioBases are extracted for each genre class.

5.2 Rule discovery

Figure 5 and 6 show several VideoBases and AudioBases. Since the VideoBases are extracted from 12x12x12 VideoCubes, they are also cubes of the same size (12x12x12). And since VideoCubes consist of pixels in both spatial and temporal dimensions, Videobases can be viewed as spatial-temporal filter (stacks of spatial filters at sequential time steps) which captures spatial-temporal characteristics at once. In the following, we call the spatial filter of a VideoBasis at each time point a **slice**. In Figure 5, each row is a VideoBasis and its slices are arranged in their temporal order, from left to right.

The VideoBases of the two genre classes, news and commercial, reflect the major differences between these two classes.

Observation 1 ***VideoBasis*** *VideoBases for news have clearer edge-like patterns (shown in the slices of the VideoBases) and have less transition as time moves on, i.e., few differences on the patterns from slice to slice. On the other hand, the VideoBases for commercial are more noisy (specific patterns are less obvious) and have greater transition/chopping between slices. The properties that these bases reflect about news and commercial clips generally agree with our perceiving of news stories and commercials.*

For example, in Figure 5(a), the $\mathbf{a_4}$ VideoBasis of news shows a (white) -45^o edge moving first from bottom-left to upper-right (slice 1 to 4) and then reverse (slice 5 to 8) and then reverse again (slice 9 to 12). In Figure 5(b), the $\mathbf{a_4}$ VideoBasis of commercial shows a big transition between slice 6 and 7, where slices 1 to 6 have some random patterns and slice 7 to 12 have a clear edge pattern moving downward.

Figure 6 shows the AudoBases we extracted. These bases also coincide with our common sense about the sounds occur in news stories and commercials.

Observation 2 ***AudioBasis*** *The AudioBases of news stories contain amplitude envelopes that are localized in time. The waveforms of these bases are intermediate between those of pure harmonic sound and pure non-harmonic sound, and resembles those of human speech. Efficient coding of human speech contains both harmonic (for vowels) and non-harmonic (for consonants) sounds [13]. This agrees with our knowledge that the most frequent sound in news stories is human speech. As for the AudioBases for commercials, the waveforms are more harmonic and are similar to those of animal vocalizations and natural sounds [13]. This characteristics of AudioBases can be attributed to the more dominant of music than speech in commercials.*

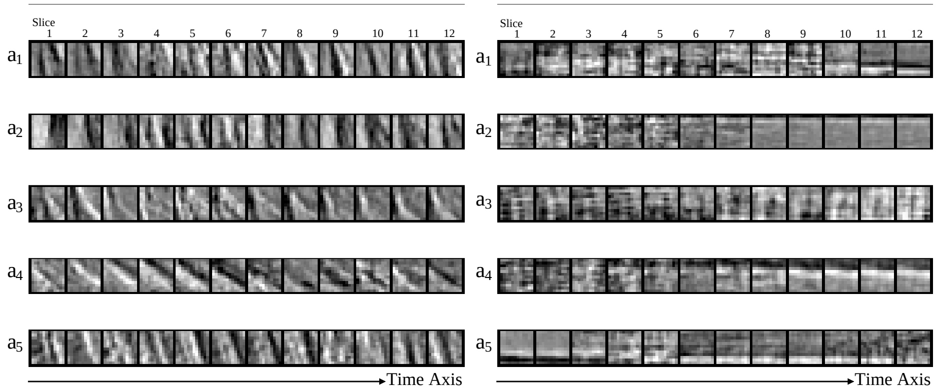
Observation 3 ***Cross-media rule*** *Comparing the discovery from VideoBases and AudioBases, we found that*

- *static scene usually corresponds to human voice, and*
- *dynamic scene usually corresponds to natural sounds (music, animal vocalization).*

5.3 Classification

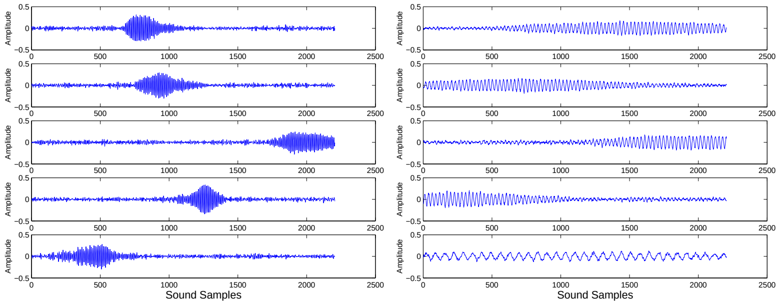
The VideoBases and AudioBases we extracted give efficient coding of a video genre, due to the sparse coding property of ICA. In other words, a video clip is encoded most efficiently when it is encoded by the (AV-)bases of its genre. We classify a video clip to a genre class whose AV-Bases give the least reconstruction error (Figure 2(c)).

In this study, VideoBases and AudioBases are used for classification separately. That is, we do 2 kinds of classification: one based on VideoBases and the other based on AudioBases. Table 5.3 lists our result on classifying a testing set of 62 news and 43 commercials into two genre classes (news and commercial).



(a) VideoBases (news) (b) VideoBases (commercial)

Fig. 5. VideoBases Each VideoBasis contains spatial-temporal information and is arranged as a stack of slices at sequential time points. Slices of a VideoBasis are arranged in a row from left to right in their temporal order. (a) Bases of news have clearer Gabor edge detector pattern shown within slices, and the pattern slightly moves along the time axis. (b) Patterns of commercial VideoBases are not as clear as those of news bases, but the changes along the time axis are more significant. (VideoBases shown here are from the central region of the 9 regions.)



(a) AudioBases (news) (b) AudioBases (commercial)

Fig. 6. AudioBases AudioBases capture the sound characteristics of different video genres: (a) bases for news resemble waveforms of human speech (mix of harmonic and non-harmonic sounds), (b) bases for commercials resemble waveforms of natural sound and animal vocalization (harmonic sound).

In our experiments, classification using either VideoBases or AudioBases gives similar results. Commercial clips are classified with higher accuracy, while news is classified less accurately. This is because news stories often contain field coverages which have more (background) motions and faster transitions, and these confuse the classification process using VideoBases. There are also back-

Class	Total	VideoBases			AudioBases		
		News	Commercial	Accuracy	News	Commercial	Accuracy
News	62	45	17	0.726	46	16	0.742
Commercial	43	3	40	0.930	4	39	0.907

Table 1. Classification result using VideoBases and using AudioBases

ground sounds along with the speech sounds of the foreground reporters in these field coverages which deteriorate the classification accuracy using AudioBases. Overall, we achieve a classification accuracy around 81% which is comparable to previous classification results (Section 2).

6 Conclusion

In this paper, we proposed VideoCube, a novel method for video classification. Its contributions are:

1. An **unified** approach to incorporate spatial and temporal information, and works on both video and audio information.
2. An **automatic** feature extraction method base on *independent component analysis (ICA)*.
3. The extracted features (*VideoBases* and *AudioBases*) are **natural** that they are closely related to those used in human perceptual processing.

VideoBases and AudioBases successfully capture the major characteristics of video content. They found the following patterns:

- *News* In news reports, less activity (motion) and editing transitions are present, and the major sound is human speech.
- *Commercial* In commercials, more activities and fast editing transitions are presented, and the major sounds are natural sound, music and animal vocalization.
- *Cross-media rule* Static scenes correspond to human speech and dynamic scenes correspond to natural sounds.

Our experiments on classifying 62 news and 43 commercial clips using either VideoBases or AudioBases achieved good classification accuracy: around 73% for news and 92% for commercials. Overall accuracy is around 81%, which is comparable to those of previous work.

References

1. Horace B. Barlow. Unsupervised learning. *Neural Computation*, (1):295–311, 1989.
2. Marian Stewart Bartlett, H. Martin Lades, and Terrence J. Sejnowski. Independent component representations for face recognition. *Proceedings of SPIE; Conference on Human Vision and Electronic Imaging III*, January 1998.

3. Anthony J. Bell and Terrence J. Sejnowski. The "independent components" of natural scenes are edge filters. *Vision Research*, (37):3327–3338, 1997.
4. Nevenka Dimitrova, Lalitha Agnihotri, and Gang Wei. Video classification based on hmm using text and faces. *ACM Multimedia*, 2000.
5. Stephan Fischer, Rainer Lienhart, and Wolfgang Effelsberg. Automatic recognition of film genres. *The 3rd ACM International Multimedia Conference and Exhibition*, 1995.
6. Andreas Girgensohn and Jonathan Foote. Video classification using transform coefficients. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 6.
7. Patrik O. Hoyer and Aapo Hyvarinen. A probabilistic framework for the adaptation and comparison of image codes. *J. Opt. Soc. of Am. A: Optics, Image Science, and Vision*, March 1999.
8. Aapo Hyvarinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 1999.
9. Aapo Hyvarinen. Survey on independent component analysis. *Neural Computing Surveys*, 2:94–128, 1999.
10. Aapo Hyvarinen. Independent component analysis: Algorithms and applications. *Neural Networks*, 13(4-5):411–430, 2000.
11. Jong-Hwan Lee, Ho-Young Jung, Te-Won Lee, and Soo-Young Lee. Speech feature extraction using independent component analysis. *International Conference on Acoustics, Speech, and Signal Processing*, in press, June 2000.
12. Te-Won Lee, Mark Girolami, Anthony J. Bell, and Terrence J. Sejnowski. A unifying information-theoretic framework for independent component analysis. *International Journal on Mathematical and Computer Models*, in press, 1999.
13. Michael S. Lewicki. Efficient coding of natural sounds. *http://www.cs.berkeley.edu/~lwalk/seminar/lewicki-NN5407.pdf*, 2001.
14. Rainer Lienhart, Christoph Kuhmunch, and Wolfgang Effelsberg. On the detection and recognition of television commercials. *Proceedings of the International Conference on Multimedia Computing and Systems*, pages 509–516, 1996.
15. Zhu Liu, Jincheng Huang, and Yao Wang. Classification of tv programs based on audio information using hidden markov model. *Proc. of 1998 IEEE Second Workshop on Multimedia Signal Processing (MMSP'98)*, pages 27–31, December 1998.
16. Zhu Liu, Yao Wang, and Tsuhan Chen. Audio feature extraction and analysis for scene classification. *Journal of VLSI Signal Processing*, Special issue on multimedia signal processing:61–79, October 1998.
17. Cheng Lu, James Au, and Mark S. Drew. Classification of summarized videos using hidden markov models on compressed chromaticity signatures. *ACM Multimedia*, 2001.
18. Bruno A. Olshausen and David J. Field. Wavelet-like receptive fields emerge from a network that learns sparse codes for natural images. *Nature*, (381):607–609, 1996.
19. Matthew J. Roach and John S. Mason. Video genre classification using audio. *EuroSpeech*, 2001.
20. Matthew J. Roach, John S. Mason, and Mark Pawlewski. Video genre classification using dynamics. *Int. Conf. on Acoustics, Speech and Signal Processing*, 2001.
21. Kim Shearer, Chitra Dorai, and Svetha Venkatesh. Local color analysis for scene break detection applied to tv commercials recognition. *Proc. 3rd. Intl. Conf. on Visual Information and Information Systems (VISUAL'99)*, pages 237–244, June 1999.

22. Kim Shearer, Chitra Dorai, and Svetha Venkatesh. Incorporating domain knowledge with video and voice data analysis in news broadcasts. *IEEE KDD 2000*, Multimedia Data Mining Workshop, August 2000.
23. Ba Tu Truong, Svetha Venkatesh, and Chitra Dorai. Automatic genre identification for content-based video categorization. *International Conference Pattern Recognition*, 4:230–233, 2000.
24. J. H. van Hateren and D. L. Ruderman. Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. *Proc. Royal Society Lond. B*, (265):2315–2320, 1998.
25. Howard Wactlar, Michael Christel, Y. Gong, and A. Hauptmann. Lessons learned from the creation and deployment of a terabyte digital video library. *IEEE Computer*, 32(2):66–73, February 1999.
26. Ramin Zabih, Justin Miller, and Kevin Mai. A feature-based algorithm for detecting and classifying scene breaks. *Proceedings of ACM Multimedia*, pages 189–200, November 1995.
27. Xiang Sean Zhou, Baback Moghaddam, and Thomas S. Huang. Ica-based probabilistic local appearance models. *IEEE International Conference on Image Processing (ICIP)*, October 2001.