NEURAL NET BASED SCENE CHANGE DETECTION FOR VIDEO CLASSIFICATION

R.K. Mallikarjuna Rao,
K. R. Ramakrishnan,
N. Balakrishnan,
Multimedia Systems Lab,
Supercomputer Education and
Research Centre,
Indian Institute of Science

S.H.Srinivasan
Institute of Robotics and
Intelligent Systems,
Bangalore
India

Abstract - A new technique for scene change detection using neural networks is proposed to operate on compressed video data obtained with only minimal decoding of video. In this paper various types of neural network structures have been trained and tested on different varieties of video clips. The neural networks are shown to be robust for camera activities and object movements. They gave comparatively good results over the conventional techniques. All the algorithms are tested on MPEG-1 video streams.

INTRODUCTION

The basic step towards the development of a video database management system is to segment the video into fundamental parts namely video shots. Video segmentation is done by detecting the breaks that occur in the video and this detection process is called scene change detection*Scene breaks occur in the video due to various edit effects.

Editing effects introduce additional set of images into the video at particular locations and they will smoothen the transition between two video shots. Various edit effects are:

- **Fade** Here a video shot gradually changes to a back ground (fade in) or a video shot gradually comes out from the background (fade out).
- **Dissolve** Here two video shots will be involved and the transition from one to another occurs smoothly. It is a simultaneous operation of both fade in and fade out.
- **Wipe** Here a line moves across the screen and the video shot comes behind that line.

0-7803-5610-1/99/$10.00 © 1999

As we see above scene breaks occur either abruptly or gradually depending on whether edit effects are there in the video.

RELATED WORK

The existing work in the area of scene change detection can be classified into a) Uncompressed domain approaches b) Compressed domain approaches c) Model based approaches.

Abrupt scene change detection is easy compared to that of gradual scene changes. Many techniques have been reported in the literature [1], [4], [2]. Twin comparison approach has been proposed [3] for gradual scene changes. In this they use two thresholds, one to identify possible location of the start of the transition and the second threshold to mark the end of the transition, when the cumulative measure exceeds the second threshold.

Model based approaches proposed in [5], use explicit production models of the video to get the location of the various scene changes. Though these methods are quite accurate, they will work only on uncompressed data.

Compressed domain approaches have been reported in [6], [7], [3]. They have used DCT coefficients and motion vector counts to identify the location of the scene breaks.

NEURAL NET BASED VIDEO SEGMENTATION

Scene change detection methods can also be classified as top-down or data driven approaches [1], [2], [3] and bottom-up or model driven approaches [5]. Data driven approaches extract some parameters from the data (compressed or uncompressed) and by applying suitable thresholds, try to detect scene breaks. They are not accurate and robust, as it is not possible to set global thresholds. Model driven approaches, on the other hand, use the mathematical models of all the editing effects and they will declare a scene break, when the extracted parameters follow any of these models. These techniques are more accurate than data driven approaches. But the biggest disadvantage of model driven approaches is that, they can be applied to uncompressed data only, as the mathematical models are available for pixel data. When we extract some parameters from the compressed domain with minimal decoding, there are no models available to fit this data. We need an intelligent system which can approximate the underlying model with the parameters extracted from the compressed domain. In this paper, application of neural networks in this perspective is investigated. We have experimented with feed forward network, radial basis function network and recurrent network for scene change

detection. For comparing the performance same training data has been used with all the neural networks.

TRAINING DATA PREPARATION

The essential idea behind compressed domain processing is to avoid the expensive overhead of decoding up to pixel level. We have used the following parameters from the compressed video, which can be obtained with only minimal decoding of the video streams.

- *Number of Intra Coded Blocks:* This parameter gives the total number of macro blocks which are coded in intra-mode. For an I-Frame all the macro blocks will be intra coded. For P or B frames this parameter varies from almost zero to maximum number of macro blocks in the picture.
- *Number of Threshold Exceeded Intra Coded Blocks:* This parameter is obtained by averaging the DCT difference of two corresponding blocks belonging to present and past I-Frames and declaring a block as "changed" or "unchanged" depending on whether average exceeds a threshold. The count of changed blocks gives the parameter.
- *Number of Macro Blocks with Forward Prediction:* This parameter measures the total number of macro blocks which are motion compensated with forward prediction (backward in time). This parameter for I-frames will be zero.
- *Number of Macro Blocks with Backward Prediction:* This i parameter measures the total number of macro blocks which are motion compensated with backward prediction (backward in time). This parameter will be zero for both I and P types of frames.
- *Number of Macro Blocks with Bidirectional Prediction:* This parameter measures the total number of macro blocks which are bi-directionally predicted (both forward and backward in time). This parameter will also be zero for both I and P types of frames.
- *Frame Type:* This indicates the type of the frame. I-frame is given number '1', P-frame is given number '2', and B-frame is given number '3'.

MPEG-I decoder code is modified to extract these parameters. The training data preparation consists of decoding the video once up to the DCT coefficient level and extracting the parameters.

FEATURE SET WITH CONTEXT

The individual frame's characteristics are not sufficient to label a frame, as belonging to a particular class of scene transition. Though it is intuitively clear that neighborhood is very important in deciding the class of a particular frame, it has been experimentally observed that, neural networks are not converging with only a single frame data. To account for this a window is made consisting of some number of past and future frame data. As there are 6 features per each frame, if n is the window size, the total number of features for a particular frame is (2n+1)x6. The window size n can be varied, to increase the number of features for a frame. After the extraction of parameters during decoding is over, the training data is prepared by using the window.

NEURAL NETWORK OUTPUT CLASS

The output of the neural net will mark each frame as belonging to an abrupt change or a gradual change or no change. In the case of gradual transition, there will be many sub classes like fades, dissolves, wipes etc. For the video classification purposes, it is sufficient to detect where the transition has taken place, rather than the exact type of transition that has occurred. Hence for all the gradual transitions, only one class is given. About 164 sets of training samples, containing the data from video frame with various effects like abrupt changes, gradual changes, no changes are considered for training all the neural nets. Though the number of training samples of each type are not equal, it is observed that training is good. MATLAB neural net toolbox is used for all the networks.

DETAILS OF THE NEURAL NETWORKS

Radial Basis Neural Network

Number of Hidden layers	= 1
Number of Hidden neurons	= 164
Input data size	= 164 × 30
output data size	= 164 × 3
Spread of radial basis functions	= 1.0

Feed Forward Neural Network

Number of Hidden layers	= 1
Hidden layer activation function	= tansigmoid
output layer activation function	=linear
Number of hidden layer neurons	=30
Input data size	= 164 × 30
Output data size	= 164 × 3
Learning rate	= 0.01
Momentum constant	=0.9
Error reached	= 0.252
Number of epochs	= 30000

Recurrent Neural Networks

Hidden layer activation function	= tansigmoid
output layer activation function	= linear
Sum-squared error goal	= 0.01
Momentum constant	= 0.95
Error reached	= 0.45
Number of epochs	= 30000

SIMULATION RESULTS

Various video clips with different scene changes have been used for testing the trained neural network. Five clips have been considered with various scene changes and one clip is prepared with no scene changes, but with camera motions like, zoom in, zoom out, pan etc. This clip is considered for checking the robustness the neural network for the camera activities. The following table gives a comparative study of the performance of all neural networks with conventional techniques.

Clip Name	Total No. of Frames	Actual No of Frames with scene changes	Correctly Detected/Detection Accuracy			
			with Con.	with F.F.Net	with Radial	with Recurrent
Nasa	1707	31	4/95.70	16/96.60	3/97.47	12/89.3
Moon	1577	93	20/91.80	31/91.90	23/94.46	34/87.95
Crick	582	7	3/81.78	3/93.81	1/97.05	2/95.87
Moh	1000	0	0/63.40	0/68.50	0/99.98	0/76.20
Echo	169	5	1/92.89	1/85.79	1/96.96	1/84.61

SUMMARY

In this paper neural networks have been shown to be efficient for scene change detection. Different neural network structures, namely Feed forward, Radial basis function and Recurrent neural networks have been studied. A comparison with conventional methods is also presented. A judicious combination of the three neural network structures is likely to yield more accurate results.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the financial support from IBM Solutions Research Centre, New Delhi.

REFERENCES

1. Nagasaka A, Tanaka Y (1991), "Automatic Video Indexing and Full Video Search for Object Appearances". *Knuth E, Wegner L (eds) Second Working Conference on Visual Database Systems (Budapest, Hungary).* pp 119-133.

2. Shiuh-Sheng Yu, Jinn-Rong Liou, and Wen-Chin Chen. "Computational Similarity Based on Chromatic Barycenter Algorithm". *IEEE Transactions on Consumer Electronics*, Vol. 42, No. 2, May 1996.

3. Zhang H. J, Kankanahalli A, Somilar S. W. " Automatic Parsing of Full Motion Video", *Multimedia Systems*, Vol. 1, pp 10-28, 1993.

4. Zabih R, Miller J, Mai K, "Feature Based Algorithm for Detecting and Classifying Scene Breaks", *Fourth ACM Conference on Multimedia*, pp 189-200, 1995.

5. Arun Hampapur, Ramesh Jain and Terry Weymouth, "Digital Video Segmentation", *ACM Multimedia*, 1994.

6. Yeo B. L, "Efficient Processing of Compressed Images and Video", *PhD Thesis Princeton University*, NJ.

7. Meng J, Juan Y, Chang S. F, "Scene Change Detection in an MPEG-Compressed Video Data",*SPIE Digital Video Compression*, pp 14-25.

8. Zhang H. J, Low C. Y, Gong Y and Smoliar S. W, "Video Parsing Using Compressed Data", *SPIE Image and Video Processing*, pp 142-149.