ACCV2002: The 5th Asian Conference on Computer Vision, 23--25 January 2002, Melbourne, Australia.

Locating Anchor Shots in Compression Domain Based on Neural Networks

W. Q. Wang[1] L.Y. Qing[2] Y. Fu[1] W. Gao[1,2,3]
[1](Institute of Computing Technology, Chinese Academy of Sciences, BeiJing, 100080)
[2](Computer Department, School of Graduate, Chinese Academy of Sciences, Beijing, 100039)
[3](Department of Computer Science, Harbin Institute of Technology, 150001)
E-mail: {wqwang, lyqing, yfu, wgao}@ict.ac.cn

Abstract

*Anchor shots are important elements in news video***,** *and locating them accurately and thoroughly is crucial to parse news video. The paper presents a novel approach, using neural networks, to detect anchor clips. Firstly, a background model is constructed through neural networks learning. Then, the trained neural networks classify frames in news video into two classes, i.e. anchor frames and non-anchor frames. At last, based on repeatability and dispersing of anchor shots on the temporal axis, false declarations in the outputs of neural networks are filtered out by clustering. The evaluation experiments, on nine days of news videos, demonstrate the approach is a fast, effective one, with the recall 98.2% and the accuracy 100%..*

1. Introduction

With the increasing digital video available, effective organization, retrieval and browse of video documents become one of the most active research areas. Large volume and lack of structure of video data bring much trouble to the management of them. Therefore it is necessary to structure automatically video documents to support content-based video retrieval and browse. But it is not yet well resolved to extract high-level semantic structure for general video. TV news, due to its inherent structural characteristic and the strong demand of viewers to fast search news items of their interest, has attracted much attention in the research area of video retrieval. An anchor shot is an important structure marker in news video, which usually introduces and concludes a news item. Therefore locating anchor shots accurately and without missing can provide a powerful clue to extract news items automatically.

Among the existing video parsing systems for news video[1][2][3], anchor shot detection is one of key modules. Zhang et.al. [1] exploit image analysis techniques to identify anchor shots. For each shot generated by a shot segmentation module, the parser exploits histogram and pair-wise pixel metrics to find candidates of anchor shots, and further verifies them based on region models. They used two days' SBC News as the test data, and reported the accuracy of detecting anchor shots achieved above 95%, with the assumption that shot segmentation had the accuracy 100%. Qi et.al. [3] integrate audio and visual cues to detect anchorperson shots. After audio segments characterized by different speakers and shots represented by key frames are clustered, anchorperson candidates are selected based on the following heuristics, that the proportion of the anchorperson speech/image is higher and the distribution is more disperse. Then **AND** operation between anchorperson shots and speech segments is done to identify the true ones. The accuracy close to 98％ is reported.

The paper presents a novel approach, using neural networks to detect anchor clips. The approach uses multiple neural networks to memorize and model different backgrounds in anchor shots. During detection, the trained neural networks apply inherent association capability to decide whether a frame is an anchorperson frame or not. The approach has many good characteristics. For instance, the whole detection process can be executed online, and does not involve shot segmentation. The algorithm operates in compression domain. Therefore the computation cost is very low.

The rest of the paper is organized as follows. In section 2, a detailed description of our approach is presented. Section 3 gives the results of evaluation experiments. Section 4 concludes the paper.

2. Description of our algorithm

Anchor shots are common structural elements in TV news. In news program produced by different broadcast corporations, a frame in an anchor shot usually involves two objects, anchorperson and background, as in Fig. 1. Anchorpersons may differ, or be in different dresses on different days, but the whole background or certain a region in it usually keeps unchanged for a long period. Due to the learning and memory features of neural networks, the background model can be constructed and embedded in neural networks through training. Then the system uses the trained neural networks to classify frames in news video into two classes, anchor frames and non-anchor frames, so as to locate the corresponding anchor shots.

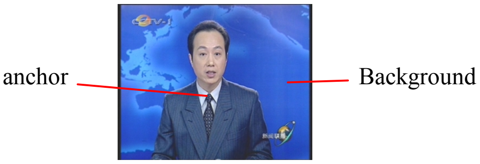

Fig. 1 A typical anchorperson frame in CCTV news

2.1. Constructing background model for anchor shots

2.1.1. Feature extraction. In MPEG video streams, each frame image is divided into 16*16 pixel blocks, called as macroblocks. Each macroblock contains a number of 8*8 pixel blocks. Applying the algorithm in [5][6], we can extract a DC image for each frame in MPEG compressed video with minimal decoding. In DCT domain, a DC coefficient is equal to the average of the pixels in that block; therefore the DC image contains most information of the original frame. Through interactively choosing an appropriate region in the background as a feature region, feature vectors are formed by the DC coefficients in the region and used as the training samples of neural networks. The system does not choose the whole background as the feature region, since it will result in too high dimension of the feature vector and too long training time. To eliminate the influence of change of light conditions on the model stability on different days, only DCs of the components Cb and Cr are extracted to form the feature vector.

2.1.2. Training Neural Networks. Define a feature vector $N = \{b_1, r_1, b_2, r_2, ..., b_i, r_i, ..., b_k, r_k\}$, $i = 1,2,...,k$, where b_i, r_i are the DCs of the components Cb and Cr of the i^{th} block in the feature region. Before training, we normalize the input features N. In the normalization process, each component b_i or r_i is divided by 255. After that, a new feature vector is generated, which is defined as $N' = \{b_1', r_1', b_2', r_2', ..., b_i', r_i', ..., b_k', r_k'\}$. Each component of the vector N' belongs to the same range [0, 1]. For anchorperson frames, the desired output of neural networks is 1; for non-anchorperson frames, the desired output of neural networks is 0. The type of neural networks in the system is BP network. In training, samples within the training set are input to the neural networks in the random order with each epoch. In some news video, more than one anchor preside the program and the corresponding backgrounds differ to some extent. Therefore the system can consist of multiple networks and each network memorizes a class of background.

2.2. Locating anchor shots using neural networks

Suppose the system consists of multiple BP networks, defined as $T_1, T_2, \ldots, T_M$. For news program on each day, the following algorithm is used to locate anchor shots.

① Initialize each component of priority vector $P = \{P_1, P_2, ..., P_M\}$ as zero, and set current frame Fn to the sequence number of the first I frame

② Extract the feature vector $N_c = \{b_1, r_1, b_2, r_2, ..., b_i, r_i, ..., b_k, r_k\}$ from the current frame Fn.

③ Choose an appropriate neural network to accept the input of the feature vector N_c using the following method. First all the networks are sorted according to the priorities before the input of N_c, let $P_{q1} \ge P_{q2} \ge ,..., \ge P_{ql}$, and then T_{q1} will be chosen as the current network, and go to ④. If all neural networks have been used, the system will declare that the frame Fn is not an anchorperson frame and go to ⑤.

④ After N_c enters the network T_{q1}, the corresponding output O_c is generated. If $O_c \ge \delta$ holds, the system considers the current frame as an anchorperson frame, where δ （$0 < \delta < 1$） is a predefined threshold. Additionally, $P_c = P_c + 1$ and continue ⑤. If $O_c \ge \delta$ does not hold, go to ③

⑤ If the current frame is an anchorperson frame, and previous detection generated the same declarations in W_s consecutive I frames, and the $W+1^{\text{th}}$ I frame prior to the current frame is not an anchorperson frame, the system will declare an anchor shot appearance event. The corresponding event start point s is set to $Fn - W \times goplength$, where $goplength$ represents the GOP length of the MPEG stream. If the current frame is not an anchorperson frame, and previous detection generated the same declarations in W_s consecutive I frames, and the $W+1^{\text{th}}$ I frame prior to the current frame is an anchorperson frame, the system will declare an anchor shot disappearance event. The corresponding event end point e is set to $Fn - (W+1) \times goplength$. The resulting anchor shot clip is outputted.

⑥ $Fn = Fn + goplength$. If the video comes to the end, the algorithm terminates; else go to ②.

In the above procedure, the priority vector is introduced to improve the efficiency of the system, which make the system heuristically choose the neural network with maximal probability to calculate first. At the same time, consistency inspection is imposed in step ⑤, to prevent noise disturbance from some natural scenes which have similar color with that of the feature region.

2.3. Refinement of the output of neural networks

Since anchor shots in news video appear repeatedly and distribute dispersedly, the feature is exploited to refine the output of neural networks in section 2.2. Define anchor shots clips generated by neural networks as $< AS_i, AE_i >$, $i = 0,1,2,....,N$, where AS_i, AE_i are the sequence number of start frame and end frame of clips. We choose key frames for each clip. For instance, the system simply chooses the middle frame of a clip as representation frame, and we define the notation $K_0, K_1,..., K_N$ to represent them. The corresponding sequence numbers of the representation frames rank in ascending order. First, all representation frames are clustered and each one gets a class label using the below procedure.

①Initialization. Set $C = 1, L[i] = 0$, for $i = 0,1,2,....,N$

② $g = \min\limits_{L[i]=0}\{i\}$

③ $L[g]$=C

④ For $k = g+1,....,N$, if $L[k]$ =0 holds, and $Dist(H(K_k), H(K_g)) \le \hbar$, Set $L[k] =$ C , where $H(K_k), H(K_g)$ are chrominance histograms of the representation frame K_k, K_g , $Dist(x, y)$ is the function of calculating Euclidean distance between two vectors, $\hbar$ is a small threshold.

⑤ $C = C + 1$

⑥ If $L[i] \neq 0$ holds for arbitrary $i = 0,1,2,....,N$, the procedure exits; else go to step ②.

The above computation classifies all representation frames into $C-1$ classes. The component $L[i]$ of vector L holds the class of the frame K_i . Define the histogram of vector L as $Hist(L)$ ={ $h_1, h_2, ..., h_{c-1}$ }, where h_r , $r = 1,2,....,c-1$ is the number of the components in L whose value is r . If $h_r < \eta$ holds, where η is a threshold, set all $L[i] \Leftarrow -1$， if $L[i] = r$ ($i = 0,1,2,....,N$) holds. For the case $h_r \ge \eta$, define $a = \min\limits_{L[i]=r}\{i\}$, $b = \max\limits_{L[i]=r}\{i\}$, if d(K_a, K_b)< D holds, where d(K_a, K_b) is the distance between frame K_a and frame K_b on the temporal axis and D is a large threshold , then set all $L[i]$, whose value is r , to -1， $i = 0,1,2,....,N$. In the last step, the system checks the values of $L[i]$ ($i = 0,1,2,....,N$), if $L[i] \neq -1$ holds, the clip outputted by neural networks and containing the representation frame K_i will be declared as an anchor shot clip. Otherwise, the clip will be filtered out as a false one.

3. Evaluation experiment

We have implemented the algorithm described in section 2 and use CCTV news program to evaluate the approach. The data set is formed by fifteen days of CCTV news recorded in April and May, 2001. Six days' data forms a training set and the remaining data is used to test. The system consists of two neural networks, and they have the same topology. Both are made up of three layers, with 160 nodes in the input layer, 80 nodes in the hidden layer and 1 node in the output layer. Since a post-process after neural networks detection can effectively filter out false clips but can not get back those missed anchor shot clips, it is desirable that the output of neural networks has a high recall. The experiment results show the trained networks not only have a high recall but also locate anchor shot clips accurately when W = 3 and δ = 0.83. Tab. 1 lists the related experiment results generated only through neural networks detection. According to Tab. 1, the recall R and the accuracy C are calculated out as follows:

$$R = \frac{S - U}{S} = \frac{114 - 0}{114} = 100\ \%$$

$$C = \frac{D - E}{D} = \frac{129 - 15}{129} = 88.4\ \%$$

Tab. 1 the Experiment results only using neural networks with W =3 and δ =0.83

Programs	Frames	S	D	E	U
CCTV0419	49930	15	15	0	0
CCTV0427	44807	14	17	3	0
CCTV0429	44819	9	11	2	0
CCTV0509	44783	12	16	4	0
CCTV0510	44735	14	16	2	0
CCTV0511	44807	12	13	1	0
CCTV0516	44483	12	12	0	0
CCTV0518	43871	11	13	2	0
CCTV0523	44740	15	16	1	0
Total	406975	114	129	15	0

Note: S, number of anchor shots manually identified by watching the programs; D, anchor shots identified by the system; E, anchor shots falsely identified by the system; U, anchor shots missed by the system.

For the tested data, the output of neural networks has a very ideal recall 100%. Further process of the output using histogram analysis will generate the resulting anchor shot clips. The results are tabulated in Tab.2, and the corresponding parameters chosen are $\hbar$ =0.18, η =3, D=5000. From tab.2, the recall and accuracy of the system are derived:

$$R = \frac{S - U}{S} = \frac{114 - 2}{114} = 98.2\%$$

$$C = \frac{D - E}{D} = \frac{112 - 0}{112} = 100\ \%$$

In the test experiment, the system filters out all the false declarations by histogram analysis. It demonstrates the post-process addressed in section 2.3 has a strong filtering ability and is very significant in the system. Therefore we are inclined to require the output of neural networks to reach a high recall when we choose parameters for neural networks. Additional experiments show a higher value chosen for δ will produce a higher accuracy but the recall will lower to some extent.

Tab. 2 the results after histogram analysis further filters the results in tab.1

Programs	Frames	S	D	E	U
CCTV0419	49930	15	15	0	0
CCTV0427	44807	14	14	0	0
CCTV0429	44819	9	9	0	0
CCTV0509	44783	12	12	0	0
CCTV0510	44735	14	12	0	2
CCTV0511	44807	12	12	0	0
CCTV0516	44483	12	12	0	0
CCTV0518	43871	11	11	0	0
CCTV0523	44740	15	15	0	0
Total	406975	114	112	0	2

Note: S, number of anchor shots manually identified by watching the programs; D, anchor shots identified by the system; E, anchor shots falsely identified by the system; U, anchor shots missed by the system.

In the test experiment, two true anchor shots in the stream CCTV0510 are filtered out wrongly. Further analysis finds they form a new class, since the distances between them and the class identified by human eyes exceed $\hbar$ =0.18. And the new class contains so few elements that $h_r \geq \eta$ does not hold. We feel such case is very rare.

4. Conclusion

An anchor shot is an important event in TV news. It is a very useful indicator of news content structure, therefore fast and accurate identification of them is crucial for automatically parsing TV news. The paper presents a new approach, based on neural networks to resolve the problem. Since the approach operates in compression domain, the computation cost is low, and it supports the feature of on-line detection. The approach is tested on 4.5 hours of news video and recall 98.2% and accuracy 100% are achieved. The experiment results show the approach is valid and effective. But the system need to construct background models for different TV stations through neural networks, and it is not easy for an ordinary user to extract features and control learning process. So Simplifying the process and developing more convenient human-machine interface will be one of our efforts in the future.

Acknowledgment

This work has been supported by National Science Foundation of China (contract number 69789301), National Hi-Tech Development Programme of China (contract number 863-306-ZD03-01-2), and 100 Outstanding Scientists foundation of Chinese Academy of Sciences.

5. References

[1] H. J. Zhang, S.Y. Tan, S. W. Smoliar and Y. Gong, "Automatic Parsing and Indexing of News Video", Multimedia Systems, 2: 256-266, 1995

[2] C. Y. Low, Q. Tian, and H.J. Zhang, "An Automatic News Video Parsing , Indexing and Browsing System," Proc. ACM Multimedia 96, Boston, MA,PP.425-426, Nov. 1996

[3] W. Qi, L. Gu, H. Jiang, X.R. Chen and H.J. Zhang, " Integrating Visual, Audio and Text Analysis for News Video", IEEE ICIP-2000, Vancouver, Canada, Sept., 2000.

[4] N.Y. Zhang, P.F. Yan, "Neural Networks and Fuzzy Control", Original Edition published by Tsinghua University Publisher, 1998.

[5] B.L. Yeo and B. Liu, "Rapid Scene Analysis on Compressed Videos", in IEEE Transactions on Circuits and Systems for Video Technology, Vol. 5, No. 6, pp. 533-544, December 1995

[6] J. Song and B. L. Yeo, "Spatially Reduced Image Extraction from MPEG-2 Video: Fast Algorithms and Applications", in Storage and Retrieval for Image and Video Database VI, Vol. SPIE 3321, Jan. 1998. pp. 93-107