Content-based Retrieval of Video Shot Using the Improved Nearest Feature Line Method

Li Zhao[2*], Wei Qi[1], S. Z. Li[1], S. Q. Yang[2], H. J. Zhang[1]

[2] Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

[1]Microsoft Research China, Zhichun Road, Beijing 100080, China

ABSTRACT

The shot based classification and retrieval is very important for video database organization and access. In this paper we present a new approach 'Nearest Feature Line – NFL' used in shot retrieval. We look key-frames in shot as feature points to represent the shot in feature space. Lines connecting the feature points are further used to approximate the variations in the whole shot. The similarity between the query image and the shots in video database are measured by calculating the distance between the query image and the feature lines in feature space. To make it more suitable to video data, we improved the original NFL method by adding constrains on the feature lines. Experimental results show that our improved NFL method is better than the traditional classification methods such as Nearest Neighbor (NN) and Nearest Center (NC).

1. INTRODUCTION

Retrieval in video database is a challenging task because of the huge data volume and rich content of video material. Currently, the widely accepted query modes are keyword-based and example-based. The example-based approach is a good alternative to the text-based one, because the users could not always say clearly what they want by using text description. The example queries could be a video clip, a frame, an object or some low-leveled features such as color, texture and motion [1][2][3][4][5][6]. The system should search in the video database and cluster the relevant video segments (such as video shots) according to the content of the query example. The most important issues for such example-based approaches lie in two aspects: First, how to select features to represent the content of a video segment and to be used as indexes. Second, how to define a distance metric in the feature space which could be used for matching. This paper focuses on the second issue.

It is obvious inefficient to match the query frame by frame with the video shots in the database. Most of the presented methods work on the key-frames of the shots. Dimitrova et.al [1] regarded the average distance of corresponding frames between two videos as the similarity measure. Zhang et al. [3] defined another similarity according to the sum of the most similar pairs of key-frames. These methods could all be classified as the Nearest Center (NC) or Nearest Neighbor (NN) matching methods according to some feature space. The drawback for these kinds of methods lies in that they all leave out the temporal variations and correlation between key-frames within an individual shot.

In this paper, a new distance metric and an efficient classification algorithm named as Improved Nearest Feature Line (NFL) is proposed. It could be used in both query-by-frame and query-by-shot schemes.

Stan Z. Li et al. [7][8] first presented the NFL method and used it in face recognition and audio classification/retrieval. The NFL works in the following way: Each pair of prototypes (face pose) belonging to the same class are interpolated or extrapolated by a linear model. They are generalized by the feature line which is the line passing through the two points. The feature line provides information about linear variants of the two prototypes. Using the minimum distance between the feature points of the query and the feature lines does the classification and retrieval.

This method is inherently compatible with the shot-based retrieval issue, in which the key-frames of a shot could be looked as prototypes.

In our scenario, each frame of a shot is considered as a point in the feature space. When the content of video frames continuously changes from one to another in some way, the change will result in a trajectory linking the corresponding feature points in the feature space locally. The key-frames are just like some sampling points on the trajectory, so we could use the lines between feature points of key-frames as an approximation of the whole trajectory of a shot. In order to make it more appropriate for video classification/retrieval, we revise the original NFL method by imposing constrains on the feature lines according to the activity measurement of a shot.

The organization of this paper is as following. In section 2, we focus on the description of improved Nearest Feature

[*] This work is done while the author visiting Microsoft Research China.

Line (NFL) method. In section 3, the effectiveness of our proposed approaches is compared with some other classification methods by experiments over test video data. Concluding remarks are provided in section 4.

2. VIDEO SHOT RETRIEVAL USING THE IMPROVED NEATEST FEATURE LINE METHOD

NFL method assumes that there are at least two sample points (feature point) in each class and through these known samples linear extrapolation or interpolation could be made to generate the feature line.

We regard the key-frames in a shot as the known sample points in the feature space. Since in this paper we focus on the classification issue, we simply select the color histogram space as the feature space that we discussed.

In section 2.1 we describe how to use the NFL method in shot retrieval. In section 2.2 we introduce how to improve the original NFL method.

2.1 Video shot retrieval using NFL method

In a shot we think of that the distance in color histogram space between two adjacent key-frames mainly caused by the object motion or camera manipulation. So we use the line passing through the feature points of two key-frames to approximate the trajectory of the continuously frames between the two end key-frames.

We consider two frames F_i and F_j in video space mapping to the points f_i and f_j in the feature space. Let

$$f_k = \{f_{1k}, f_{2k}, \cdots f_{Mk}\} \qquad (1)$$

Where M is the dimension of the feature space. The difference between frames F_i and F_j can be measured as Euclidean distance $\Delta f = \left\| f_i - f_j \right\|$ in their feature space. The straight line passing through f_i and f_j of the same class, denoted by $\overline{f_i f_j}$, is called a feature line (FL) of that class (See figure 1).

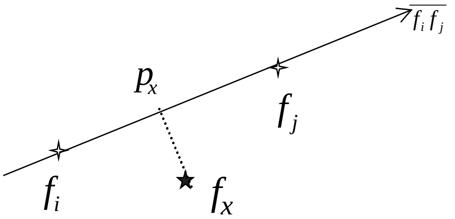


Figure 1: Feature points f_i and f_j, feature line $\overline{f_i f_j}$, and query feature point f_x.

Let $F^c = \{f_i^c \mid 0 < i \le N_c\}$ be the set of the N_c prototypical feature points belonging to class *c*. A number of $K_c = \dfrac{N_c(N_c - 1)}{2}$ lines can be constructed to represent the class. The FL space for class *c* is composed of the K_c feature lines:

$$S^c = \{\overline{f_i^c f_j^c} \mid 0 < i, j \le N_c, i \ne j\} \qquad (2)$$

This is a subset of the entire feature space. When there are ***M*** classes in the database, ***M*** such FL spaces can be constructed, with a total number of N_{total} FL's where $N_{total} = \sum_{c=1}^{M} K_c$.

Letting f_x be the query feature point, supposing the query-by-frame situation. We define the distance between the query point f_x and the feature line $\overline{f_i f_j}$ as $Dist(f_x, \overline{f_i f_j})$. Let p_x denote the projection of f_x on the feature line $\overline{f_i f_j}$ (see figure 1) :

$$p_x = f_i + \mu(f_j - f_i) \qquad (3)$$

Where

$$\mu = \frac{(f_x - f_i) \bullet (f_j - f_i)}{(f_j - f_i) \bullet (f_j - f_i)} \qquad (4)$$

We define corresponding NFL distance is:

$$Dist(f_x, \overline{f_i f_j}) = \left\| f_x - p_x \right\| \qquad (5)$$

Here, $\|\bullet\|$ denotes Euclidean distance.

The distances between the query point f_x and each line in the whole feature line space could be calculated, and these distances are then sorted in ascending order. The best matching shot in the video database should be the one containing the key-frames that form the feature line of the smallest distance with the query frame. The sorted list of shots gives the retrieval result in order.

2.2 Improved NFL method

There is a potential problem when using the original NFL method directly in video shot retrieval. Because the original feature line could extend finitely for beyond the two feature points. That is, the extension part of one feature line may cross the other feature line. This may cause problem when the feature line crosses the other feature spaces. So we need to limit the extension range of feature line, and could not let it be infinite in the feature space.

Here we adopt the concept of Shot Activity [9] and use it as a measurement added to constrain the feature line. When watching a video sequence, people may perceive it as being a slow sequence, fast paced sequence, or action sequence etc. The activity captures this intuitive notion of 'intensity of action' or 'pace of action' in a video segment. Therefore we use a descriptor that enables us to accurately express the activity of a given video sequence/shot. To evaluate this idea, we define:

$$Act_c = \underset{i}{MAX}(\|f_i - f_{center}\|) \tag{6}$$

Where $1 \le i \le N_c$, and $f_{center} = \frac{1}{N_c}\sum_{i=1}^{N_c} f_i$.

We define a bound on acceptable feature line distance:

$$SearchRange = C \cdot Act_c \tag{7}$$

Here C is a constant, which is set to *0.65* in our experiment.

Only when the condition

$$\|f_x - f_{center}\| \le SearchRange \tag{8}$$

is satisfied, the result is accepted.

So the idea here is that shot activity determines the distribution of feature points in a shot. If one feature point is far from the main distribution of the class, even it is very close to some feature line of the class, we still do not accept that the point is belongs to this class.

3. EXPERIMENTAL RESULT

In our experiment we use color histogram as the feature. The color feature is defined according to the 1976 CIE u'v' perceptually uniform color space. The generative model is a histogram: space u'v' is divided into 256 square bins (16 bins on a side). The model is that a bin is chosen with probability proportional to the stored histogram count and the color of a pixel is the color of the center of the chosen bin. There are 16 bins in u' that span from 0.16 to 0.2883. There are 16 bins in v' that span from 0.4361 to 0.5361. Colors outside of these spans are clipped to the nearest bin.

To evaluate the performance of the proposed improved NFL method, we build a video database of 160 shots. Among them, 16 shots can be unambiguous classed and are regard as test shots. These shots are taken from forty minutes Sports News of CCTV, including track and field, swimming, soccer, gig , advertisements etc.

We use precision and weight score to measure the performance. The following measure [8] will be used in performance evaluation:

$$\eta(q,m) = \sum_{k=1}^{m} w_k Match(q, r_k) \tag{9}$$

$$Match(q, r_k) = \begin{cases} = 1 & \text{If } r_k \text{ and } \boldsymbol{q} \text{ is correlative} \\ = 0 & \text{If } r_k \text{ and } \boldsymbol{q} \text{ is not correlative} \end{cases}$$

Where, ***q*** represents the query shot and $r_1, r_2, \cdots, r_m$ are the m top ranked matches for the query shot ***q*.** The judgments for the value of $Match(q, r_k)$ are given by subjective evaluation in our experiments.

$w_k = W \cdot \frac{1}{k}$ is a decreasing sequence of weights $(k = 1,2,\cdots)$ where $W = \frac{1}{\sum_{k=1}^{N_q} \frac{1}{k}}$, here N_q is the number of queried shots that match the query shot ***q***. Because the weights w_k are decreasing with the rank position ***k***, a higher ranked correct match contributes more to $\eta(q,m)$. The weights are normalized by the factor ***W*** in the following sense: When the top N_q matches are all correct, $\eta(q,m)$ reaches the highest possible value of 1.

Figure 2 gives the result of our experiment. We evaluate the conventional NFL method, the improved NFL method and other classification methods, including the Nearest Center (NC) and the Nearest Neighbor (NN).

The results show that the NFL based methods produce better performance than NC and NN, and the improved NFL method fits shot retrieval and classification better compared with the conventional NFL.

Figure 3 gives an example of query result. In this example, when using the conventional NFL method, the problem of one feature line passing through other feature space may occur (See Figure 3c). When using the improved NFL method, this problem could be solved

4. CONCLUSIONS

In this paper, we have presented a new approach to content-based representation of video shots and the application in example-based shot retrieval. In this approach, we use a new classification method (NFL) for video shot classification and retrieval. In accordance to the characteristics of video data, we have revised the original NFL method by adding constrains on the feature lines according to shot activity. The experimental results have shown that the proposed method achieves high performance not only better than the traditional methods for classification such as NN and NC, but also the original NFL method. The main advantage comes from the fact that both correlation between consecutive frames in a shot and activity constrains in a shot are considered in our approach.

In the experiments, we have only used color histogram as the feature for classification. In the future, we will add more features (texture, shape, etc) for classification; and even better performance will be achieved.

5. ACKNOWLEDGMENTS

This work is supported by Microsoft-Tsinghua Joint Lab.

6. REFERENCES

[1] Nevenka Dimitrova and Mohamed Abdel-Mottaled, “Content-based video retrieval by example video clip”, SPIE Vol. 3022, 1998.

[2] M.M. Yeung and B.Liu, “Efficient matching and clustering of video shots”, IEEE International Conference on Image Processing 1995 Vol.1 pp.338-341

[3] H.J.Zhang, D.Zhong and S.W.Smoliar, “An Integrated System for Content-Based Video Retrieval and Browsing,” Pattern Recognition, Vol.30, No.4, pp.643-658, 1997.

[4] D.Zhong, S.F.Chang, “Spatio-Temporal Video Search using the Object-Based Video Representation”, IEEE International Conference on Image Processing 1997 Vol 1, pp.21-24

[5] Man-Kwan Shan, Sun-Yin Lee, “Content-based Video Retrieval based on Similarity of Frame Sequence”, Proc. IEEE Conf. on Multimedia Computing and Systems, pp.90-97, 1998

[6] Y.Deng and B.S.Manjunath, “Content based search of video using color, texture and motion”, IEEE International Conference on Image Processing 1997 Vol 2, pp.13-16

[7] S. Z. Li and J. Lu, “Face recognition based on nearest linear combinations”, *IEEE Transactions on Neural Networks*, vol. 10, no. 2, pp.439-443, March 1999.

[8] S. Z. Li, “Content-based Classification and Retrieval of Audio Using the Nearest Feature Line Method”, *IEEE Transaction on Speech and Audio Processing*, vol. 8, no.5, pp.619-625, 2000.

[9] P. R. Hsu, H. Harashima, “Detecting Scene Changes and Activities in Video Databases”, IEEE Int. Conf. On Acoustics, Speech, and Signal Processing, 1994.

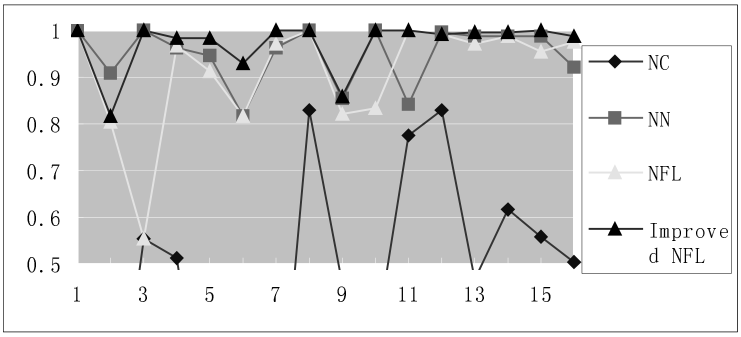
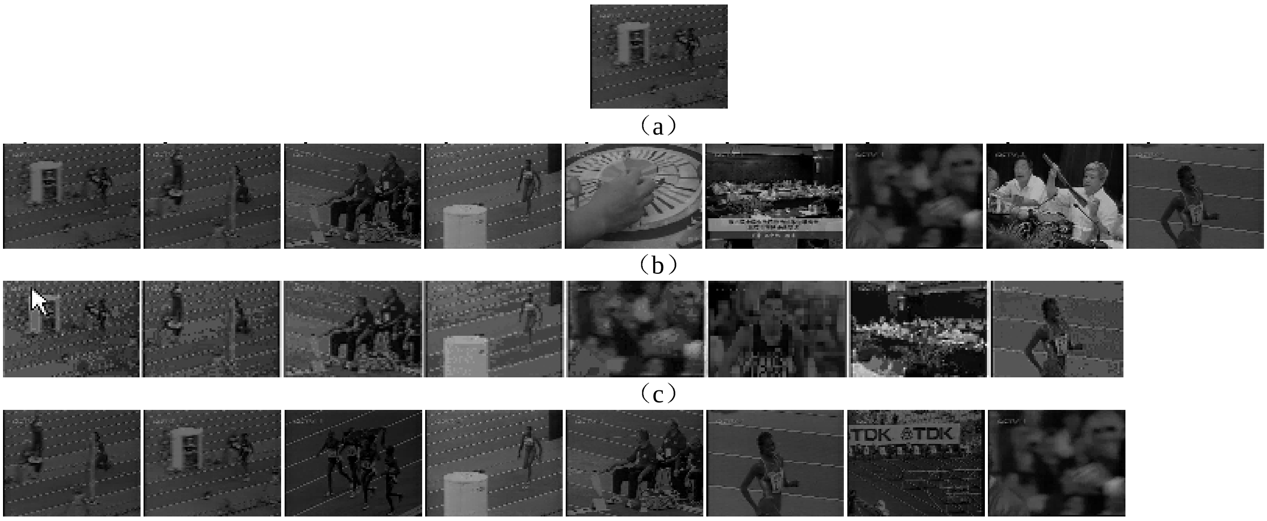


Figure 2: The score of different methods.


（a）

（b）

（c）

（d）

Figure 3: An example of retrieval result.
(a) Query Image. (b) The result of NN method. (c) The result of NFL method. (d) The result of Improved NFL method.