VIDEO BACKGROUND RETRIEVAL USING MOSAIC IMAGES

Xue Mei, Mahesh Ramachandran

Center for Automation Research
and Department of ECE
University of Maryland, College Park, MD 20742
{xuemei,maheshr}@cfar.umd.edu

Shaohua Kevin Zhou

Integrated Data Systems Department
Siemens Corporate Research
Princeton, NJ 08540
kzhou@scr.siemens.com

ABSTRACT

Content-based video retrieval is one of the most active and exciting research areas in the field of multimedia technology. In this paper, we present an approach for video background retrieval using mosaic images and a Support Vector Machine(SVM). The video is captured by a moving camera and a portion of the scene is visible at any time. The Kanade-Lucas-Tomasi(KLT) feature tracker is used to get the correspondences between consecutive images and the homography is calculated using these correspondences. We use the homography to construct the mosaic background image and a Mixture of Gaussian(MoG) background subtraction algorithm to remove the moving objects in the scene. An SVM is then applied to classify the mosaic background image. The experimental results show the efficiency and effectiveness of the proposed approach.

1. INTRODUCTION

Content-based retrieval has become an active research area since the early 1990's and a large number of retrieval systems has been developed [1-3]. It is still a challenging task to search and retrieve a query video from large numbers of videos. One of the interesting subproblems in video retrieval is indexing and retrieval of background regions from a video database. For example, the task may be ”Please retrieve the videos that have the background of mountains?”. How to solve this particular problem is the focus of this paper.

For the video captured by a moving camera, each frame only captures a portion of the scene which is not suitable for retrieval. We should be able to process the whole video that captures a full view of the scene. Suppose, for example, a video of a tall building is taken. It is natural to move the camera vertically along the tall building to take a full view of it. In order to handle this situation, mosaic background image could be used for video retrieval.

Image mosaics have attracted a growing attention in recent years and have been used in many applications, such as video stabilization, background subtraction and virtual environment [4]. We can construct the mosaic image from a video by aligning and properly blending together partially overlapped images acquired by a moving camera.

Once we get the mosaic image, we input it to an SVM for classification. SVM [5, 6] is a very efficient binary classifier and can also be applied to multi-class classification and has attracted much attention in image retrieval literature recently. Based on SVM, a classifier can be learned from the training data which is marked by the users. Then the model can be used to find more relevant videos in the database.

The rest of the paper is organized as follows. In section 2, we introduce the mosaicing algorithm used in our experiment. In section 3, we introduce the background subtraction algorithm used for removing the moving objects in the background image. In section 4, we provide a brief overview of SVM and its training algorithm. The proposed method is then tested and experimental results are given in section 5. Finally, concluding remarks and future work are discussed in section 6.

2. MOSAICING

An image mosaic is a panoramic image obtained by collating a sequence of frames after alinging all the images onto a common reference frame.

Two views of the same scene can be related by a nonsingular linear transformation of the projective plane called homography or collineation. This happens in two cases: i) The camera performs a pure rotation or ii)The scene can be well approximated by a single plane(that is, the depth range of the scene is small compared to the distance from the camera). In these two cases, there will be no parallax and images can be composited to form a mosaic image.

2.1. Homography Computation

The homography between two consecutive frames is computed using the KLT tracker [7, 8]. Features are tracked

0-7803-8874-7/05/$20.00 ©2005 IEEE

through the video and correspondences are obtained between each pair of consecutive frames. Using homogeneous coordinates, we relate the correspondences by a non-singular 3×3 matrix H:

$$\begin{bmatrix} x'_1 \\ x'_2 \\ x'_3 \end{bmatrix} = \begin{bmatrix} H_{1,1} & H_{1,2} & H_{1,3} \\ H_{2,1} & H_{2,2} & H_{2,3} \\ H_{3,1} & H_{3,2} & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \quad (1)$$

2D points in the image plane are denoted as $m = (u, v) = (x_1/x_3, x_2/x_3)$, $m' = (u', v') = (x'_1/x'_3, x'_2/x'_3)$. Each point correspondence in the plane provides two equations in the unknown entries of H:

$$u'(H_{3,1}u + H_{3,2}v + 1) = H_{1,1}u + H_{1,2}u + H_{1,3} \quad (2)$$

$$v'(H_{3,1}u + H_{3,1}v + 1) = H_{2,1}u + H_{2,2}u + H_{2,3} \quad (3)$$

We need four correspondences to define the the transformation matrix uniquely. If $n > 4$ points, the equation can be reduced to the form $Ax = b$. The solution of this equation is $A^{+}b$, where $A^{+} = (A^T A)^{-1} A^T$ is the pseudo-inverse of A.

The Least Median of Squares(LMeS) algorithm is used to remove the outliers due to tracking error or features attached to the moving object.

2.2. Mosaic Construction

In frame to frame approach to mosaic construction, the first frame is chosen as the reference frame, while all the others are registered to it, i.e.,

$$H_i = H_{i-1} H_{i-1,i} \quad (4)$$

where H_i is the homography between the ith frame and the first frame, $H_{i,i-1}$ is the homography between the ith frame and the $(i-1)$th frame.

3. BACKGROUND SUBTRACTION

Before we input the mosaic image to an SVM, we should identify the moving object and remove it from the background image. Pixels in the current frame that deviate from the background are considered to be moving objects. We will use Mixture of Gaussians(MoG) for the background model [9].

The MoG method tracks multiple Gaussian distributions simultaneously and maintains a density function for each pixel. The pixel distribution is represented as

$$p(I_t) = \sum_{i=1}^{K} \omega_{i,t} \cdot \eta(I_t, \mu_{i,t}, \sigma_{i,t}) \quad (5)$$

where $\eta(I_t, \mu_{i,t}, \sigma_{i,t})$ is the ith Gaussian component with intensity mean $\mu_{i,t}$ and standard deviation $\sigma_{i,t}$. $\omega_{i,t}$ is the weight for the ith component.

The parameters of the component are updated [9] as follows:

$$\begin{cases} \omega_{i,t} = (1-\alpha)\omega_{i,t-1} + \alpha(M_{i,t}) \\ \mu_{i,t} = (1-\rho)\mu_{i,t-1} + \rho I_t \\ \sigma^2_{i,t} = (1-\rho)\sigma^2_{i,t-1} + \rho(I_t - \mu_{i,t})^T(I_t - \mu_{i,t}) \end{cases} \quad (6)$$

where α is the learning rate and $M_{i,t}$ is 1 for model which matched and 0 for the remaining models.

$$\rho = \alpha\eta(I_t|\mu_{i,t}, \sigma_{i,t}) \quad (7)$$

is the learning factor for adapting current distributions.

The Gaussians are ordered by the value of ω/σ. Higher rank components have low variances and high probabilities, which are typical characteristics of background. The first B distributions are chosen as the background model, where

$$B = argmin_b(\sum_{i=1}^{b} \omega_{i,t} > \gamma) \quad (8)$$

where γ is a measure of the minimum portion of the data that should be accounted for by the background.

4. SUPPORT VECTOR MACHINE

SVM is a very powerful classification algorithm. It has been effectively used in Content-Based Image Retrieval(CBIR) in [10, 11].

Given a linear separable training samples

$$(x_i, y_i)_{i=1}^{N} \quad and \quad y_i = \{+1, -1\} \quad (9)$$

where x_i is an n-dimensional vector and y_i is the class label that the vector belongs to. The general form of the linear classification function is

$$g(x) = w \cdot x + b \quad (10)$$

which corresponds to a separating hyperplane

$$w \cdot x + b = 0 \quad (11)$$

where x is an input vector, w is a weight vector, and b is a bias. The goal of SVM is to find the parameters w and b for the optimal hyperplane to maximize the geometric margin $2/\|w\|$ between the hyperplane, subject to

$$y_i(w_i^T x_i + b) \geq +1 \quad (12)$$

The solution can be found through a Lagrangian dual objective function

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i^T x_j \quad (13)$$

We maximize L_D subject to $\alpha_i \geq 0$ and $\sum_{i=1}^{N} \alpha_i y_i = 0$.

We can represent the optimization problem and its solution via the inner product. We do this directly for the transformed feature vectors $h(x_i)$.

$$x_i \cdot x_j \rightarrow h(x_i) \cdot h(x_j) = K(x_i, x_j) \tag{14}$$

where $K(\cdot)$ is a kernel function. We then get the kernel version of the dual function

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{15}$$

The SVM classification function is given by

$$f(x) = sign(\sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b) \tag{16}$$

SVM can be extended to multi-class problems, essentially by solving many two-class problems. A classifier is built for each pair of classes, and the final classifier is the one that dominates the most pair of classes. It has particular advantage when applied to problems with limited samples in high dimensional spaces.

5. EXPERIMENTS

To evaluate the performance of the proposed algorithm, we develop the following video background retrieval system. In figure 1, image features are extracted from different background images and used as training data for designing the SVM classifier. A query is sent by the user and video is taken from the video database. The video is processed to form a mosaic image of the background and sent to SVM for classification. The retrieval results are then sent to the user.

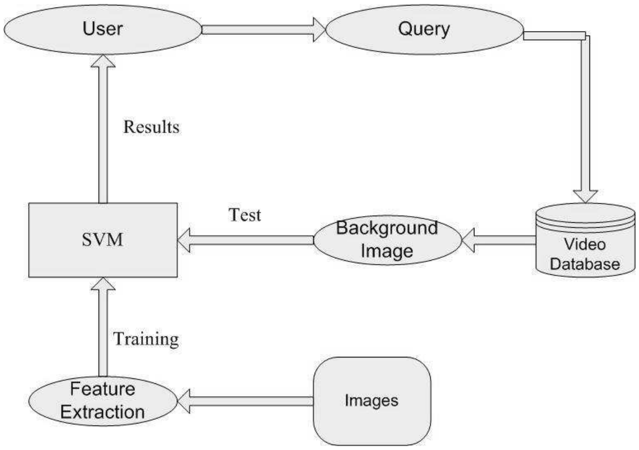
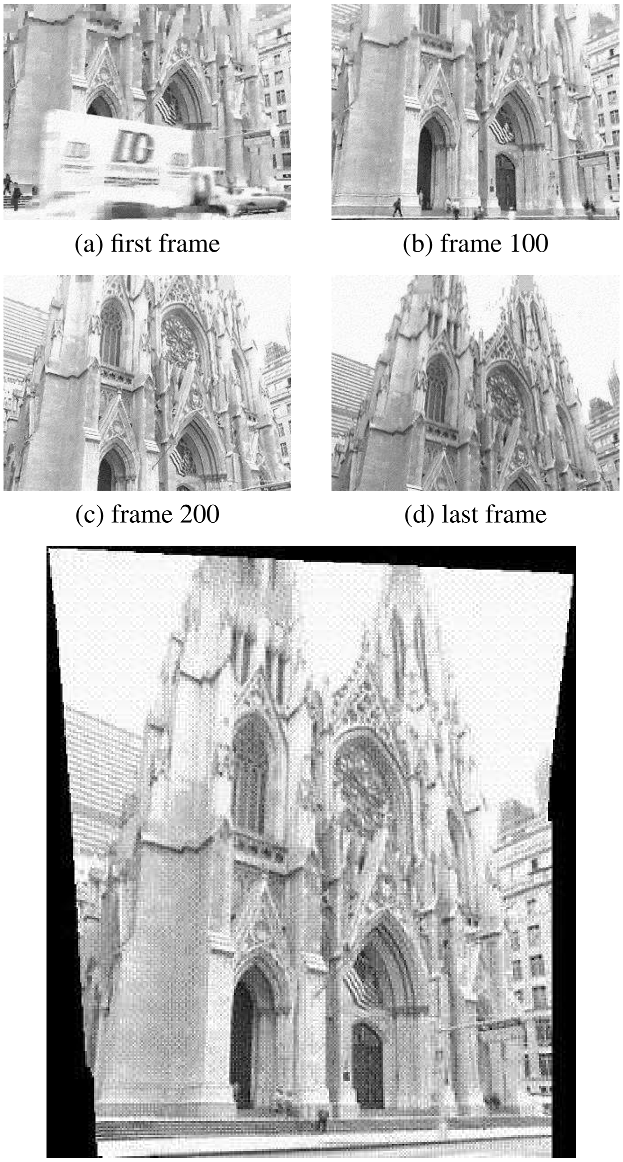


Fig. 1. Flowchart of the video background retrieval system

In our retrieval system, three main features, color, texture and shape are extracted to represent the image. For the color feature, we use the color moment [12] in HSV color space. The color histogram is quantized into 256 levels. Hue, Saturation and Value are quantized into 16, 4, 4 bins respectively. Texture is extracted using Gabor wavelet filter [13]. Four scales and six orientations are used and the feature length is 4×6. Shape features are extracted by edge histogram [14]. Edges are grouped into five categories, vertical, horizontal, 45° diagonal, 135° diagonal and isotropic (non-orientation specific). We normalize each feature and combine them to form a feature vector.



(a) first frame

(b) frame 100

(c) frame 200

(d) last frame

(e) An example of mosaic image

Fig. 2. A mosaic image of church background constructed from the video.

	beach	church	lake & rivers	clouds	mountain
Precision	90.90%	90.48%	90.00%	94.74%	94.44%
Recall	86.96%	86.36%	94.74%	90.00%	85.00%

Table 1. Experimental results for the video background retrieval.

We constructed a fully labeled training image database. It has six classes each with 100 images which are taken from the Corel Photo Gallery. These are: beach, church, lake and rivers, clouds and mountain. The features of the images are sent to the SVM for the background training.

In our experiments, 100 videos are taken and stored in a video database each of which is about 1000 frames. The frame size is 320 × 240 pixels. In advance of the retrieval, a mosaic background image is created for each video.

A mosaic of church background is shown in Figure 2. The first four images in the two rows are the first, 100, 200 and the last frames in the video. The last image is the mosaic image constructed using the above four images. The mosaic image gives a panoramic of the church and big moving objects that occlude the scene are removed.

For evaluation of the retrieval performance, precision-recall metrics are used. Recall is the ratio of the number of relevant images returned to the total number of relevant images. Precision is the ratio of the number of relevant images returned to the total number of images returned.

The results of the experiment are summarized in Table 1. The average precision and recall rate are 92.11% and 88.61% respectively. We achieve very good results using our proposed approach.

6. CONCLUSION AND FUTURE WORK

In this paper, we presented an approach for video background retrieval using a mosaic image and an SVM classifier. The experiment results prove our method's robustness and effectiveness.

Our future work will include applying bootstrap or probabilistic characterization to model pixel value of the mosaic image for further possible improvement.

7. ACKNOWLEDGEMENTS

We would like to thank Prof. Rama Chellappa for very helpful and encouraging suggestions.

8. REFERENCES

[1] A. Yoshitaka and T. Ichikawa. A Survey on Content-Based Retrieval for Multimedia Databases. *IEEE Trans. on Knowledge and Data Engineering*, Vol.11, No.1, pp.81-93, January/February 1999.

[2] S. C. Chen and R. L. Kashyap. A Spatio-Temporal Semantic Model for Multimedia Database Systems and Multimedia Information Systems. *IEEE Trans. on Knowledge and Data Engineering*, Vol.13, No.4, pp.607-622, July/August 2001.

[3] S.L. Feng, R. Manmatha and V. Lavrenko. Multiple Bernoulli Relevance Models for Image and Video Annotation. *Computer Vision and Pattern Recognition*, 2004.

[4] M. Irani, P. Anandan, J. Bergen, R. Kumar, and S. Hsu. Mosaic Representations of Video Sequences and Their Applications. *Signal Processing: Image Communication, special issue on Image and Video Semantics: Processing, Analysis, and Application*, Vol.8, No.4, May 1996.

[5] Vapnik, V.: *The Nature of Statistical Learning Theory*, Springer-Verlag, New York 1995.

[6] J. C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. *Int. J. on. DMKD* 2, pp.121-167, 1998.

[7] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. *In Proceedings of the International Joint Conference on Artificial Intelligence*, 1981.

[8] C. Tomasi and T. Kanade. Detection and tracking of point features. *Technical Report CMU-CS-91-132*, Carnegie Mellon University, Pittsburg, PA, April 1991.

[9] C. Stauffer and W. Grimson. Learning patterns of activity using real-time tracking. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22, pp. 747-57, Aug 2000.

[10] P. Hong, Q. Tian, and T. S. Huang. Incorporate Support Vector Machines to Content-based Image Retrieval with Relevant Feedback. *In Proc. IEEE ICIP*, 2000.

[11] Y. Chen, X. Zhou, and T. S. Huang. One-class SVM for learning in image retrieval. *In Proc. IEEE ICIP*, 2001.

[12] M. Stricker, M. Orengo. Similarity of color images. *Proc. SPIE on Storage and Retrieval for Image and Video Databases*, Vol. 2420, pp. 381-392, San Jose, USA, February, 1995.

[13] B. S. Manjunath and W. Y. Ma. Texture Features for Browsing and Retrieval of Image Data. *IEEE Trans. on PAMI*, vol.18 no. 8 pp. 837-42, Aug. 1996.

[14] B.S Manjunath, J. Ohm, V. Vasudevan, and A. Yamada. Color and texture descriptors. *IEEE Trans. on CSVT*, Vol. 11, pp. 703 -715, Jun 2001.