Detecting Objectionable Videos[1)]

WANG Qian HU Wei-Ming TAN Tie-Niu

(*Institute of Automation, Chinese Academy of Sciences, Beijing* 100080)

(E-mail: qwang@nlpr.ia.ac.cn, wmhu@nlpr.ia.ac.cn, tnt@nlpr.ia.ac.cn)

Abstract This paper addresses the problem of detecting objectionable videos, which has never been carefully studied before. Our method can be efficiently used to filter objectionable videos on Internet. One tensor based key-frame selection algorithm, one cube based color model and one objectionable video estimation algorithm are presented. The key frame selection is based on motion analysis using the three-dimensional structure tensor. Then the cube based color model is employed to detect skin color in each key frame. Finally, the video estimation algorithm is applied to estimate objectionable degree in videos. Experimental results on a variety of real-world videos downloaded from Internet show that this method is promising.

Key words Tensor, key frame, skin segmentation, cube based color model, objectionable video estimation

1 Introduction

It is estimated that one in every ten respondents confessed to be addicted to sex on the Internet, and that they spent at least three hours a week on such pursuits. Thus, how to filter the objectionable information from the Internet has become a serious problem. For the past, many efforts have been made on network filtering based on key words searching, semantic understanding and IP filtering, such as the system of Smart Filter and Cyber Snoop. There were some attempts on image filtering[1,2]. Commercial products of image filtering also appeared, such as the system of LookThatUP by INRIA. Other efforts were made on video retrieval[3,4]. Howerer, none of them are not specific to objectionable video filtering. The fact that little work has been done on objectionable video filtering is mainly because of the difficulties in human motion analysis. Here we translate the problem of objectionable video filtering into the problem of detecting objectionable information on key frames of video.

In this paper we present a video filtering system. In the process of key frame extraction, a tensor based motion analysis is applied in our algorithm. A new color model (the cube based color model) is presented in segmentation of human skin regions in key frames. This color model has the advantage of quickness in training and detection. In addition, we also present an evaluation rule in video summarization.

The rest of this paper is organized as follows. We present our system framework in Section 2, along with the tensor based key frame extraction, the cube based color model and the objectionable estimation of the video. Results are shown in Section 3. In Section 4, we summarize our algorithms and discuss our idea about the future work.

2 System framework

The structure of the system is shown as a block diagram in Fig. 1. First, the original video is decoded into individual frames, and then we estimate local orientations in the spatio-temporal domain using the three-dimensional structure tensor. Based on the motion analysis of each frame, we search the local minimum of the sum of motion in each frame to detect key frames in a shot. On each key frame extracted, the human skin area is segmented. During the period of skin segmentation of each key frame, retrieval of skin database is executed, based on the cube based color model. The skin color model has 125000 records and each record corresponds to a small cube in the HSL color space. 600 manually edited skin images are inputted to construct the skin database. Our color model translates the skin images into skin database records. Finally we compute the objectionable score of each key frame and evaluate the whole objectionable degree of the video based on a predefined evaluation rule.

1) Supported by National Natural Science Foundation of P. R. China (60121302) and the National High Technology Research and Development Program of P. R. China (2002AA142100)

Received November 17, 2003; in revised form March 24, 2004

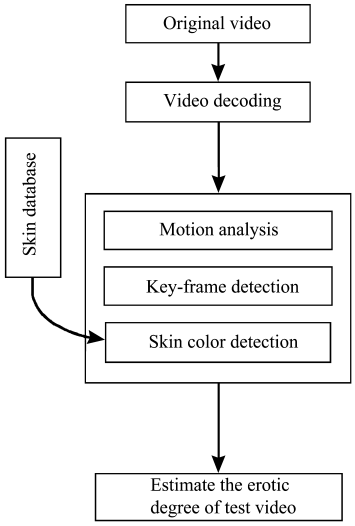

Fig. 1 Block diagram of the system

2.1 Tensor based key frame detection

Different methods for summarizing a long video have been proposed in the past; among these methods key frame selection is important. Key frames extracted directly from videos represent shots and sequences to inform users about content without requiring them to view the entire video. In our video filtering system, skin color detection is conducted on key frames rather than on the whole video, in order to reduce the computational complexity.

First classical algorithm of color histogram is used for shot segmentation. Then at each shot we extract key frames to summarize video.

There are many methods for selecting key frames. [5] and [6] selected the single key frame from each shot, based on the assumption that each shot has only one key frame. But many shots, especially shots of long duration, have more than one key frames. So this method cannot provide sufficient information on the video sequences. [7] selected key frames by video segmentation, which is conducted on edited videos. This method is not suitable for original videos. Given the complexity of video sources, Wolf has proposed an algorithm for selecting key frames by motion analysis[8], which uses optical flow to measure the motion in each shot and selects key frames at the local minima of the motion. Horn and Schunck′s algorithm[9] of optical flow is often used for motion analysis, but this method does not have enough precision in motion estimation, especially in noisy videos. Here we introduce a new key frame detection method based on tensor computing. The basic idea of our method is to analyze motion by estimating local orientations in the spatio-temporal domain using the thee-dimensional structure tensor. Motion analysis via the structure tensor has been described in [10]. Let n represent the orientation for constant gray values in image sequence, the product between n and the spatiotemporal gradient ∇I expresses the local deviation of the spatiotemporal gray value structure from an ideally oriented structure. If the gradient is perpendicular to n, the product is zero. It reaches a maximum when the gradient is either parallel or antiparallel to n. And we can determine moving and static parts on the image plane from the direction of minimal gray-value change in the spatiotemporal sequences. This minimization can be formulated as a total least square optimization problem

$$\int_{\Omega} (n^{\mathrm{T}} \nabla I(x, y, t))^2 \mathrm{d}x\mathrm{d}y\mathrm{d}t \tag{1}$$

Thus, we find the value of n to minimize (1), where $I(x, y, t)$ is a three-dimensional volume with two spatial (x, y) coordinates and one temporal (t) coordinate. (1) reaches a minimum if the vector n

is given by the eigenvector of the tensor J (2) of the minimum eigenvalue[11]

$$J = \int_{\Omega} \nabla I(x', y', t') \nabla I^{\mathrm{T}} \mathrm{d}x' \mathrm{d}y' \mathrm{d}t' \tag{2}$$

The eigenvalues $\lambda_1 \geqslant \lambda_2 \geqslant \lambda_3$ of the tensor can be used to classify the motion in local neighborhood. $\lambda_{1,2,3} \approx 0$ shows that no motion can be estimated. If $\lambda_1 > 0, \lambda_2 > 0$ and $\lambda_3 \approx 0$, real motion can be calculated by eigenvector $\boldsymbol{e}_3$ corresponding to the smallest eigenvalue

$$u = \left(\frac{e_{3,1}}{e_{3,3}}, \frac{e_{3,2}}{e_{3,3}} \right) \tag{3}$$

If $\lambda_{2,3} \approx 0$ and $\lambda_1 > 0$, there is an image structure with linear symmetry (spatial local orientation) which moves with a constant velocity. This is the well-known aperture problem and only the flow normal to the intensity structure can be calculated from e_1

$$u = \frac{-e_{1,3}}{e_{1,1}^2 + e_{1,2}^2} (e_{1,1}, e_{1,2}) \tag{4}$$

Finally, if $\lambda_{1,2,3} > 0$, the neighborhood shows no constant motion and we cannot estimate the optical flow.

Now we can calculate the sum of the magnitudes of the components of velocity at each pixel as a motion metric $M(t)$ for frame t

$$M(t) = \sum_x \sum_y ((u_x(x, y, t))^2 + (u_y(x, y, t))^2)^{\frac{1}{2}} \tag{5}$$

The $M(t)$ *vs* t curve is used to determine the local minimum of motion metric. We set $D(t) = M(t) - M(t-1)$. If the value of $D(t)$ is larger than a predefined threshold, frame t is regarded as an ascending sharp point and added to an array Sharp (t). Frame 1 and the last frame are also included in this array. The frame with minimum $M(t)$ between Sharp (t) and Sharp $(t+1)$ is selected as a key-frame.

Fig. 2 illustrates $|u(x, y)|$ on Frame 14 of the sequence of an advertising model. The results of motion representation by classical Horn and Schunck optical flow and tensor computing are also respectively shown. The tensor method can provide motion computing more precisely than the Horn and Schunck optical flow.

Fig. 3 illustrates $M(t)$ of the 44 frames of the advertising model sequences. Frame 13 and Frame 30 are the local minimum of curve $M(t)$ and thus be selected as key frames.

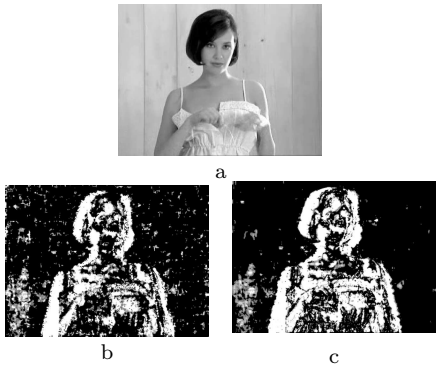

Fig. 2 The 14th frame of the advertising model and its motion repesentations
(a) The 14th frame of a video sequence
(b) Motion representation by Horn and Schunck optical flow
(c)Motion representation by tensor computing

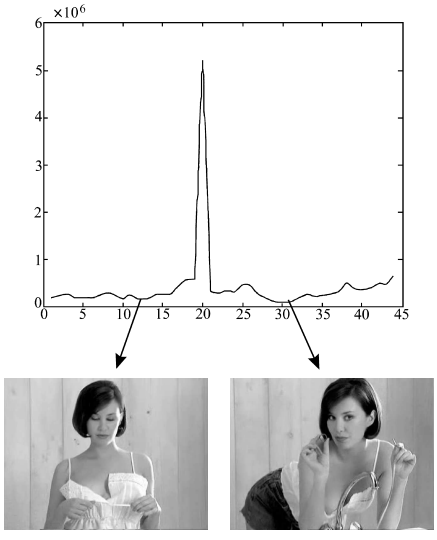

Fig. 3 The curve of $M(t)$ and selected key frames

2.2 Cube based color model

The skin color filtering is conducted on each key frame selected from video. The cube based color model is applied in the procedure of skin segmentation.

First we give a brief introduction of skin segmentation. Fig. 4 presents the framework of database training and segmentation of input image. During the training process, skin images are translated by cube based color representation and then stored in skin database. On the process of segmenting input images, this color model also translates the input key frames in order to search the skin database for matching.

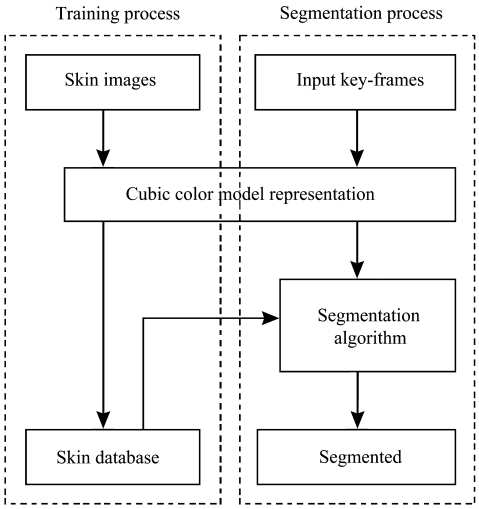

Fig. 4 Framework of the training and segmentation of image

There are many color models, among which the Gaussian skin color model is most widely used. Many models of this kind are based on the hypothesis that a Gaussian model $N(m, C)$ can represent the skin color distribution, where m is the mean and C is the covariance of color vectors[12,13]. Unfortunately, due to the complexity of skin color of different races, Gaussian models and Gaussian-Mixture models[14∼16] are not robust enough to describe the distribution of skin color. Unlike other skin color models used for skin filtering in a single image, we need a fast and reliable color model to segment human skin regions on the key-frames. We established a cubic based color model which judges whether the pixel is skin or not by database retrieval. In the following, we will present this skin color model and its results in skin segmentation.

Although many videos use RGB representation for colors, this representation is not suitable for characterizing skin color. In the RGB space, the triple component (r, g, b) represents not only color but also luminance. The HSV color space (hue, saturation, value) is often used by people who select colors (*e.g.*, of paints or inks) from a color palette, because the HSV color space corresponds to how people experience color better than the RGB color space does. Since the HSV works well for fluorescence images (monochromatic images) rather than for natural illumination and many real-world videos have natural illuminations, we replace the V (value) component with L (luminance). We use the HSL color space in our color model. A comparison of different color spaces (RGB, YUV, HSV, HSL) applied to our color model has been conducted; the results are presented in Section 3.

We use H, S, and L components to construct a large cube to contain all colors. The hue varies from 0° to 360°, saturation from 0 to 1, and luminance from 0 to 1. In order to load all the cubic elements in database, the discretization of the cubic is conducted. In convenience of the process of H (hue component), we map hue to the interval between 0 and 1. The large cube is split into small cubes, each cube is $0.02\times0.02\times0.02$, so that $50 * 50 * 50$ cubes have been formed, each of which corresponds to one record of the skin database. Fig. 5 shows the segmentation of HSL color space.

In order to accurately describe the manifold of the skin color distribution, constraints are added to small cubes. The constraints are based on RGB color space because the RGB components are isotropic and the constraints based on them have better convergence properties than those based on other color spaces.

Then the skin database training based on skin images is conducted. Skin regions in about 600 images are manually labeled. The RGB values of all skin pixels are converted to HSL. Each skin pixel is translated into one small cube and one record of the database. The constraints′ values were calculated to fill in related fields of corresponding record.

The human skin region can then be segmented based on skin database retrieval. Because the small cubes are arranged orderly in database, binary search algorithm is applied to locate a small cube, thus the searching is very fast. If all dataset in database are copied to a structure array in RAM (Random Access Memory), the searching procedure is even faster. Given one input image, the RGB values of each pixel are converted to HSL. We search the record of the small cube related to this pixel to see if there are enough pixels in this cube and if the constraints are between trained boundary values stored in skin database. If all these conditions are satisfied, the pixel is regarded as a skin pixel. The time of video filtering is depicted in Section 3. Fig. 6 illustrates the results of segmentation of two video frames extracted from sequences of an advertising model based on cubic color model.

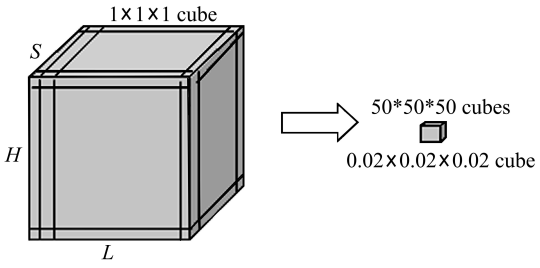

Fig. 5 The discretization of HSL cube

Fig. 6 The result of segmentation of 2 key frames in advertising model sequence

2.3 Estimation of objectionable degree in videos

The estimation of objectionable degree is based on segmentation of key frames. We give each key frame a score, which represents the proportion of skin color area to whole image area. If the value of $f(t)$ is more than a predefined threshold, then we mark this frame as an objectionable frame.

A single key frame is objectionable does not indicate that the whole video is objectionable. In fact, if a video is objectionable, most of its key frames are objectionable. So we need to create a new metric to estimate the video. Equation 6 gives a method of calculating the sum of $f(t)$ between $t1$ and $t1+\delta$ and then searching for the maximum of the sum. If this value is lager than a threshold, we decide that this video contains objectionable contents. The parameter δ is important in video estimation. We always set $\delta = 4$. Results of correct rate of the video detection under different values of δ are also shown in Section 3.

3 Results

In order to test our algorithm, we have established one objectionable video detection platform. The test video database includes 80 objectionable videos and 60 normal videos. All are real-world videos downloaded from Internet. The lengths of these videos are from 30 seconds to 5 minutes. The comparison of different color spaces used in objectionable video detector is illustrated in Table 1. We can clearly see that the algorithm based on HSL color space has the best performance.

To compare values of δ on objectionable video estimation, ROC curves on the test video database are shown in Fig. 7. The $\delta = 4$ curve has the best performance.

We test the video database on a Pentium IV 1GHZ PC. It takes 70 minutes to test 60 normal videos, with the average time of 1 minute and 10 seconds per video; it takes 30 minutes to test 80 objectionable videos, with the average time of 22.5 seconds per video. If we detect part of the video is

objectionable, we will not check the remainder of this video. This means that it always takes less time for objectionable videos to be detected than for normal videos.

Table 1 Comparison of erotic video detector using different color spaces on test data

		Correct detections (%)	False alarms (%)
RGB	(80 objectionable videos) (60 normal videos)	76.4%	18.6%
YUV	(80 objectionable videos) (60 normal videos)	80%	16.4%
HSV	(80 objectionable videos) (60 normal videos)	82.8%	15%
HSL	(80 objectionable videos) (60 normal videos)	87.1%	11.4%

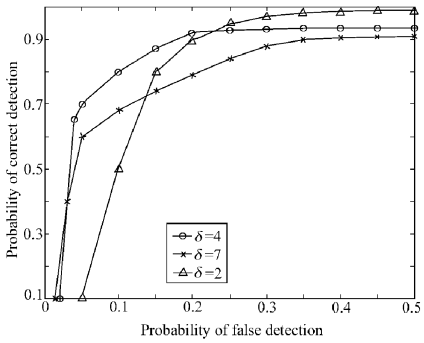

Fig. 7 ROC curves for the objectionable video detector on different values of parameter δ

4 Conclusions

Due to the complexity of natural videos, objectionable video detection is a challenging problem in video analysis. In this paper, we have presented an approach to automatically detect objectionable video from Internet. Three algorithms have been presented: the tensor based key frame extraction, the cube based color model, and the objectionable video estimation algorithm. We use three-dimensional structure tensor to represent video object′s motion. The process of key frame extraction is just to find the local minimums of summarized motion of all frames in video. Human skin segmentation is conducted on extracted key frames.

A score representing the objectionable degree marks each key frame. Finally, the objectionable video estimation algorithm evaluates average distribution of objectionable frames. The experiment results suggest that our algorithms work well on the real-world test videos.

In this paper we only give a coarse rule to evaluate the objectionable degree in a single key frame. Next we want to use the contour of the video object. In this way, the objectionable degree of each key frame will be represented by the proportion of skin color area of video object to the whole area of this object.

References

1 Forsyth D, Fleck M. Automatic Detection of Human Nudes, *International Journal of Computer Vision*, 1999, **32**(1): 63∼77

2 Fleck M, Forsyth D, Bregler C. Finding Naked People, In: Proceedings of European Conference on Computer Vision, **2**. Berlin, Germany: Springer-Veriag, 1996. 592∼602

3 Pickering M J, Rüger S M. Sinclair D. Video retrieval by feature learning in key frames, In: Proceedings of International Conference on Image and Video Retrieval. London, UK: Springer-Verlag, 2002. 309∼317

4 Hauptmann, Christel A, Papernick N M. Video retrieval with multiple image search strategies, In: Proceeding of The Second ACM/IEEE-CS Joint Conference on Digital Libraries, **11**(1). Portland, Oregon: USA: ACM Press, 2002. 56∼63

5 Tonomura Y, Akutsu A, Taniguchi Y, Suzuki G. Structured video computing, *IEEE Multimedia*, 1994, **1**(3): 34∼43
6 Mills M, Cohen J, Wong Y Y. A Magnifier Tool for Video Data, In: Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '92), Monterey, CA, USA: ACM Press, 1992. 93∼98
7 Smith M, Kanade T. Video Skimming Characterization through the Combination of Image and Language Understanding, In: Proceedings of the 1998 IEEE International Workshop on Content-Based Access of Image and Video Databases. Washington, DC, USA: IEEE Computer Society, 1998. 61∼70
8 Wayne Wolf. Key frame selection by motion analysis, In: Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing. Atlanta, GA, USA: IEEE Computer Society, 1996. 1228∼1231
9 Berthold, Horn K P, Schunck Brian G. Determining optical flow, *Artificial Intelligence*, 1981, **17**: 185∼203
10 HauBecker H, Motion H Spies, Jahne In B, HauBecker H, GeiBler P editors. Handbook of Computer Vision and Applications, volume 2, chapter 13. Academic Press, 1999
11 Bigun J, Granlund G H, Wiklund J. Multidimensional orientation estimation with applications to texture analysis and optical flow, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1991, **13**(8): 775∼790
12 Bergasa L M, Mazo M, Gardel A, Sotelo M A, Boquete L. Unsupervised and Adaptive Gaussian Skin-color Model, *Image and Vision Computing*, 2000, **18**(12): 987∼1003
13 Terrillon J C, David M, Akamatsu S. Automatic detection of human faces in natural scene images by use of a skin color model and of invariant moments,In: Proceedings IEEE International Conference on Automatic Face and Gesture Recognition. Nara, Japan: IEEE Computer Society, 1998. 112∼117
14 Caetano T S, Barone D A C. Skin segmentation via a Gaussian mixture model of chromatic features, In: Proceedings of AVBS 2001-International Workshop on Advanced Video Based Surveillance, London, UK, 2001
15 Caetano T S, Olabarriaga S D, Barone D A C. Performance Evaluation of Single and Multiple-Gaussian Models for Skin Color Modeling, *Brazilian Symposium on Computer Graphics and Image Processing*, XV Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI). Washington, DC, USA: IEEE Computer Society, 2002. 275∼282
16 Michael J. Jones, James M. Rehg, Statistical Color Models with Application to Skin Detection, *International Journal of Computer Vision*, 2002, **46**(1): 81∼96

WANG Qian Received his master degree (2001) from Southeast University, China. He is currently a Ph. D. candidate in the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include pattern recognition , computer vision, digital image processing and analysis, and video analysis.

HU Wei-Ming Received his Ph. D. degree from the Department of Computer Science and Engineering, Zhejiang University, China. From April 1998 to March 2000, he worked as a postdoctor at the Institute of Computer Science and Technology, Founder Research and Design Center, Peking University. From April 2000, he is an associate professor at the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His research interests include visual surveillance in dynamic scenes, neural network, image processing, and 3D computer graphics.

TAN Tie-Niu Received his bachelor degree (1984) in electronic engineering from Xi′an Jiaotong University, China, and M. Sc. (1986), DIC (1986) and Ph. D. degrees (1989) in electronic engineering from Imperial College of Science, Technology and Medicine, London, U. K. In October 1989, he joined the Department of Computer Science, The University of Reading, England, where he worked as Research Fellow, Senior Research Fellow and Lecturer. In January 1998, he returned to China to join the National Laboratory of Pattern Recognition, the Institute of Automation of the Chinese Academy of Sciences, Beijing, China. He is currently Professor and Director of the National Laboratory of Pattern Recognition and Director of the Institute of Automation. He is a Fellow of the IEEE and was an elected member of the Executive Committee of the British Machine Vision Association and Society for Pattern Recognition (1996∼1997). He serves as referee for many major national and international journals and conferences. He is an Associate Editor of the International Journal of Pattern Recognition, the Asia Editor of the International Journal of Image and Vision Computing and is a founding co-chair of the IEEE International Workshop on Visual Surveillance. His research interests include speech and image processing, machine and computer vision, and pattern recognition, and robotics.