Identification and annotation of erotic film based on content analysis

Donghui Wang*, Miaoliang Zhu, Xin Yuan, Hui Qian
College of Computer Science, Zhejiang University, Hangzhou, China, 310027

ABSTRACT

The paper brings forward a new method for identifying and annotating erotic films based on content analysis. First, the film is decomposed to video and audio stream. Then, the video stream is segmented into shots and key frames are extracted from each shot. We filter the shots that include potential erotic content by finding the nude human body in key frames. A Gaussian model in YCbCr color space for detecting skin region is presented. An external polygon that covered the skin regions is used for the approximation of the human body. Last, we give the degree of the nudity by calculating the ratio of skin area to whole body area with weighted parameters. The result of the experiment shows the effectiveness of our method.

Keywords: Content analysis, identification and annotation, erotic film, nude human body detection, skin detection, video segmentation

1. INTRODUCTION

In recent years interest in digital video applications has substantially increased due to advances in multimedia technologies and the publicity of the Internet. The availability of video as an accessible media leads to a new set of applications including video conferencing, video-telephony, VOD (video-on-demand), and distance learning. End-users can easily make digital videos on personal computer platform and upload to the network for share. Everyone that includes underage children can access and download all kinds of video steams from Internet. When we benefit from these innovations, serious security problem also appear on our sight. One of them is the spread of erotic films in Internet. For controlling the spread of erotic films, we are strongly interested in tools that can automatically identify and annotate the erotic films by content analysis.

The purpose of identification and annotation of erotic videos is to filter the relevant erotic shots from video streams. There are some existing solutions for providing automatic semantic annotation of video streams[1, 2, 3]. However, because of the complexity of the problem, most of the proposed solutions are domain-specific, being optimized to work only on specific video content such as news, sports, instruction, or specific categories of movies. And, all above methods need the assistant of peoples to fore-construct the map model that use the lower level visual and audio features to represent the high level semantic content. In the specific case of erotic videos, we must build a feature modal for automatic annotation at first. Usually, the visual features (continuous nude human body in many shots) are regarded as the significant clues for automatic semantic annotation[4, 5]. We can identify the continuous nude human body by detecting the human skin regions in key frames.

In this paper, we present a new method for identifying and annotating erotic films based on content analysis. First, the film is decomposed to video and audio stream. Then, the video stream is segmented into shots and key frames are extracted from each shot. This process achieves a partition of the entire video stream into collections of frames whose contents are regarded as a refined representation of whole video content. An image analysis process is introduced to detect the skin region and nude human body from collections of key frames. The results of nude human body detection can help to filter the shots that include potential erotic content. The whole process can be showed on Fig.1.

The remainder of the paper describes our technical solution to this problem. In Sec. 2 we introduce video segmentation required as a foundation for the annotation solution. In Sec. 3 a Gaussian model in YCbCr color space for detecting skin region is discussed. In Sec. 4 we discuss the degree of the nudity and an algorithm for nude detection is introduced. Sec. 5 describes the results of experiment and concludes the paper.

Electronic Imaging and Multimedia Technology IV, edited by Chung-Sheng Li,
Minerva M. Yeung, Proc. of SPIE Vol. 5637 (SPIE, Bellingham, WA, 2005)
0277-786X/05/$15 · doi: 10.1117/12.577235

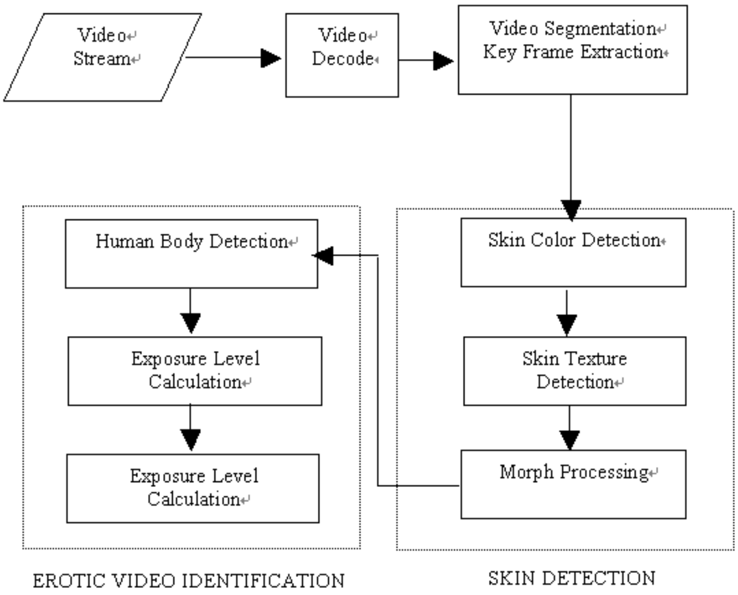
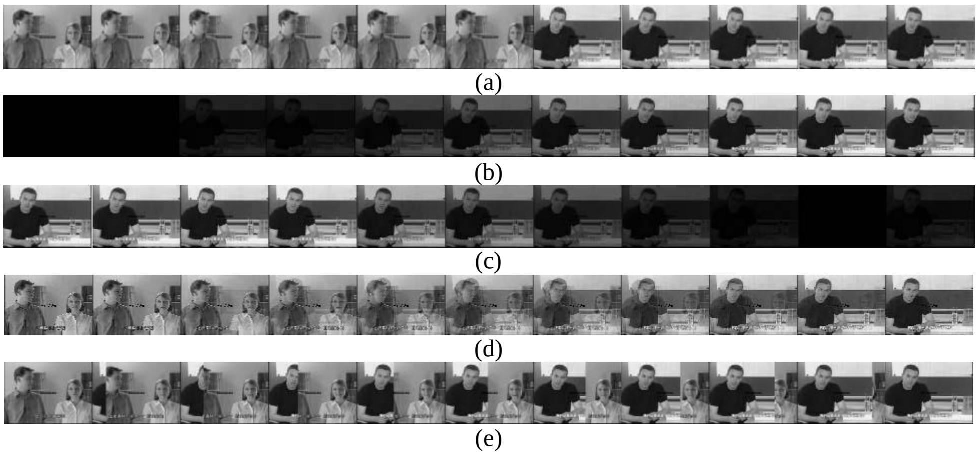

Video Stream
Video Decode
Video Segmentation Key Frame Extraction
Human Body Detection
Exposure Level Calculation
Exposure Level Calculation
Skin Color Detection
Skin Texture Detection
Morph Processing
EROTIC VIDEO IDENTIFICATION
SKIN DETECTION

Figure. 1 Erotic video identification process

2. VIDEO SEGMENTATION

Before starting the annotation process, video stream are partitioned into elemental units called shots. These shots are essentially sequences of contiguous frames formed by a continuous recording process. However, because this definition yields some difficulties for edited material (shot boundaries are vague), it can be modified to describe a contiguous sequence of frames whose content is common. The boundary between shots mainly has two types: cut and gradual transform. The cut type is a simple link of two shots. The gradual transform type includes fade-in, fade-out, dissolve and wipe. See Fig. 2.


(a)

(b)

(c)

(d)

(e)

Figure 2: (a) cut, (b) fade-in, (c) fade-out, (d) dissolve, (e) wipe.

For cut type, the visual content of frames at shot boundary change rapidly. But for gradual type, the transform is smooth. We can use the color histogram distance of two neighboring frames to measure the change of visual content and detect the boundary of shots. See Fig.3.

In our algorithm[10], considering the processing time, the change between frames is firstly influenced by color. When the color difference, after luminance influence is suppressed, is above the threshold, we think the corresponding image regions are different in content. Suppose two colors are represented in HSV color space as $C_1:(h_1,s_1,v_1)$ and $C_2:(h_2,s_2,v_2)$.

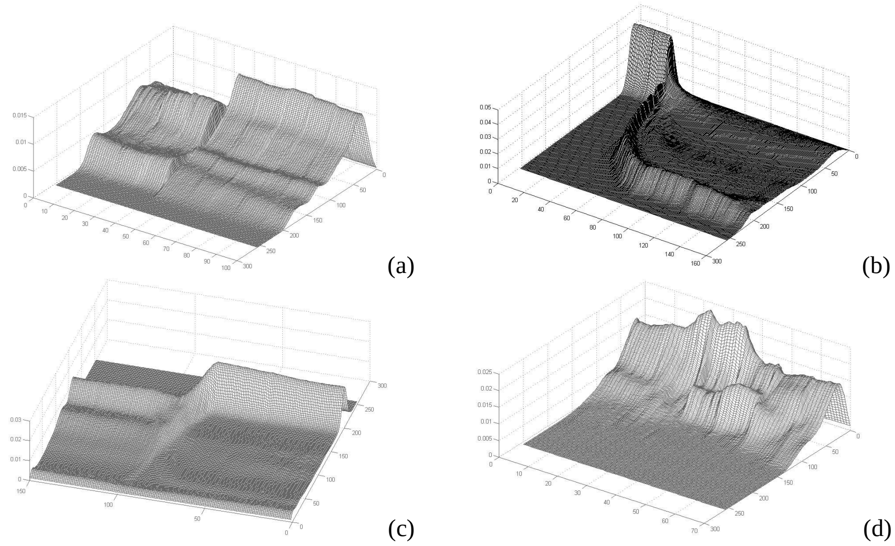


Figure 3: (a) cut, (b) fade-in, (c) dissolve, (d) wipe.

The hue difference is defined as:

$$D_t^{\ 2}=\left[s_1\cos(h_1)-s_2\cos(h_2)\right]^2+\left[s_1\sin(h_1)-s_2\sin(h_2)\right]^2 \tag{2.1}$$

The color difference is defined as:

$$D_c^{\ 2}=D_t^{\ 2}+(v_1-v_2)^2 \tag{2.2}$$

Suppose the sub-blocks of two images I_1, I_2 are R_1, R_2 ; the threshold of hue difference is ε_1 , the threshold of color difference is ε_2 . $D_t^{\ 2}(R_1,R_2)$ is defined as the average hue difference of two corresponding sub-blocks. $D_c^{\ 2}(R_1,R_2)$ is defined as the average color difference of two corresponding sub-blocks.

So we claim that: when both $D_t^{\ 2}(R_1,R_2)>\varepsilon_1$ and $D_c^{\ 2}(R_1,R_2)>\varepsilon_2$ are satisfied, we take R_1 and R_2 have content change. Suppose image I is segmented to *k* regions, each of which contains N_i sub-blocks, and R_i sub-blocks with content change， α_i is weighed coefficient.

The similarity between two images is calculated as:

$$Similarity(I_1,I_2)=\sum_{i=1}^{k}\frac{\alpha_i R_i}{N_i} \tag{2.3}$$

We compare the current frame I_k and the next I_{k+1} frame after the image similarity calculation known as $Similarity(I_k,I_{k+1})$. If the value of $Similarity(I_k,I_{k+1})$ is below the given threshold, I_k will be added to key frame sequence as a new one. The following skin detection and exposure evaluation step is done on key frames.

3. HUMAN SKIN DETECTION

The color of human skin results from a combination of blood (red) and melanin (yellow, brown). Human skin has a restricted range of hues and is not deeply saturated. Because more deeply colored skin is created by adding melanin. Finally, skin has little texture; extremely hairy subjects are rare. Ignoring regions with high-amplitude variation in intensity values allows the skin filter to eliminate more control images.

Detection of skin is complicated by the fact that skin's reflectance has a substantial non-Lambertian component[6, 7]. It often (perhaps typically) has bright areas or highlights which are desaturated. Furthermore, the illumination color varies slightly from image to image, so that some skin regions appear as bluish or greenish off-white. Skin detection model must adapt to the various luminance conditions and camera calibration problems.

Considering the convergent property of skin color distribution, we present the Gaussian model of skin color distribution in YCbCr color space and judge the skin pixels with Bayesian method. Because skin has smooth texture, we adopt a low pass filter and morphological process.

3.1 Color distribution of human skin

We choose YCbCr color space for extraction of color feature of human skin. There are two main reasons for this choice. First, the convergent property of skin color distribution has been proved to be effective[8, 9]. Second, some video standards such as MPEG choose YCbCr color space, which makes our method easier and faster to implement without further decoding video stream. We can also ignore the Y component to suppress luminance influence.

The diagram in Fig.4 is a typical human skin color distribution map in CbCr plane. The test samples are from our skin image database containing 232 nude pictures. We can see that the peak values are different for skin region and background region so that it is possible to distinguish the skin from background using statistically method.

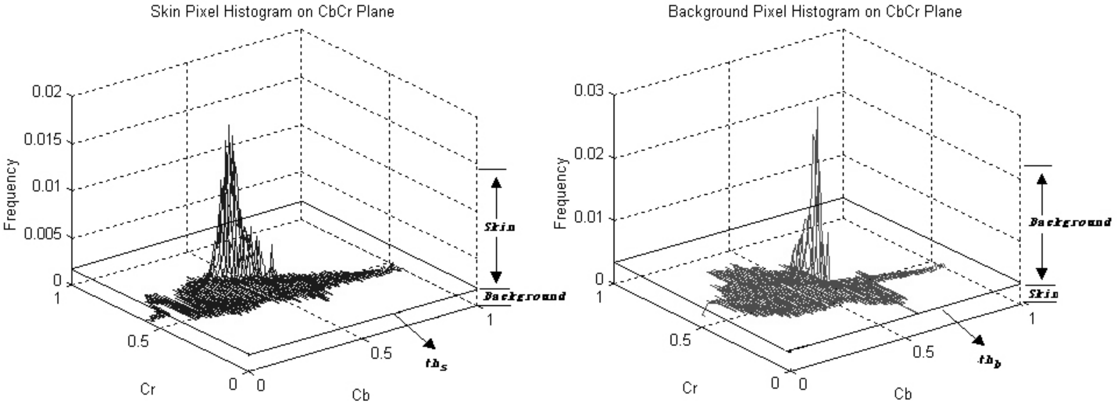


Figure 4: Skin Color Distribution Map in C_bC_r Plane

Suppose the color feature of pixel is defined as $C = \langle c_1, c_2 \rangle$. We build the Gaussian model to represent the probability of the skin color distribution in both Cb and Cr dimensions.

$$P(C|S) = \frac{1}{2\pi\sigma_{sc_1}\sigma_{sc_2}} e^{-(\frac{(c_1-\mu_{sc_1})^2}{2\sigma_{sc_1}{}^2}+\frac{(c_2-\mu_{sc_2})^2}{2\sigma_{sc_2}{}^2})} \tag{3.1}$$

$$P(C|B) = \frac{1}{2\pi\sigma_{bc_1}\sigma_{bc_2}} e^{-(\frac{(c_1-\mu_{bc_1})^2}{2\sigma_{bc_1}{}^2}+\frac{(c_2-\mu_{bc_2})^2}{2\sigma_{bc_2}{}^2})} \tag{3.2}$$

Here, S stands for skin while B stands for background. Let the average values of two color features of skin pixel c_1, c_2 to be μ_{sc_1} and μ_{sc_2}, mean square roots to be σ_{sc_1} and σ_{sc_2}. Let the average values of two color features of background pixel to be μ_{bc_1} and μ_{bc_2}, mean square roots to be σ_{bc_1} and σ_{bc_2}.

The formulas give the empirical probability of skin and background. According to Bayesian method, the threshold for acceptance of a pixel as skin is defined as:

$$\omega = \frac{P(S \mid C)}{P(B \mid C)} = \frac{P(C \mid S)P(S)}{P(C \mid B)P(B)} \qquad (3.3)$$

$P(S)$ and $P(B)$ is the probability of skin and background samples.

3.2 Texture feature of human skin

For image I_p, we do Daubechies wavelet transform to get the transformed image I_w, then we adjust the values of the three high frequent sub-bands HL, LH and HH to [0,255], and further segment to 8*8 size of small blocks, finally we can get the mean square root of the block to be the wanted texture feature. Due to the smoothness of skin texture, the corresponding mean square root should be relatively smaller. We can get rid of the high mean square root region, which might be rough texture, by presetting simply threshold.

Some fragments are likely to remain after color and texture filtering. The small fragments always seriously influence the detection of human body. Thus, usually we do morphological process to eliminate those small fragments.

3.3 Test results of skin detection

The test result of our skin detection algorithm is shown in Fig. 5. Here, the colors of the grassland, bathing dress and gold hair are similar to the human skin but their textures are different. We can use efficiently filtered out texture features by our layered algorithm while the skin area is largely remained.

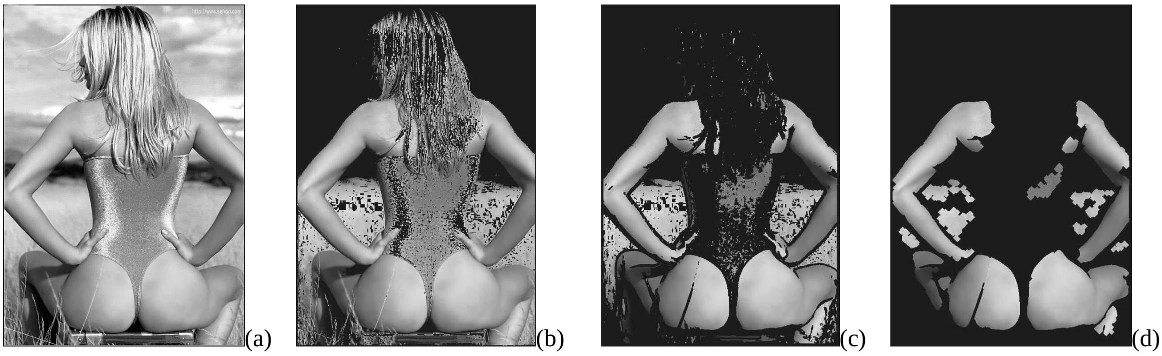


Figure 5: (a) original, (b) color filtering, (c) texture filtering, (d) morph filtering

4. EROTIC VIDEO IDENTIFICATION

It is difficult to semantically define eroticism and model. No efficient technique or theory supports the nude judgment from the skin detection results. But it is possible to evaluate the degree of the nudity (exposure level) by detecting human skin, which helps us to judge erotic content.

We define an external polygon that covered all the skin regions is used for the approximation of the human body. Then, the exposure level (EL) is defined as:

$$ExposureLevel = k_1 \times k_2 \times \frac{Area_{skin}}{Area_{body}} \qquad (4.1)$$

Here, $Area_{skin}$ is the area of skin regions and $Area_{body}$ is the area of body region (external polygon). Parameter k_1 is to reflect the influence of the ratio of skin area to the whole image. If the skin area is too small, k_1 is set to be 0. Parameter k_2 is to reflect the influence of shape of the skin region. If the shape is much like human face but not body, exposure result is also to be small.

The result of exposure level calculation using skin/body ratio is showed on Fig.6.

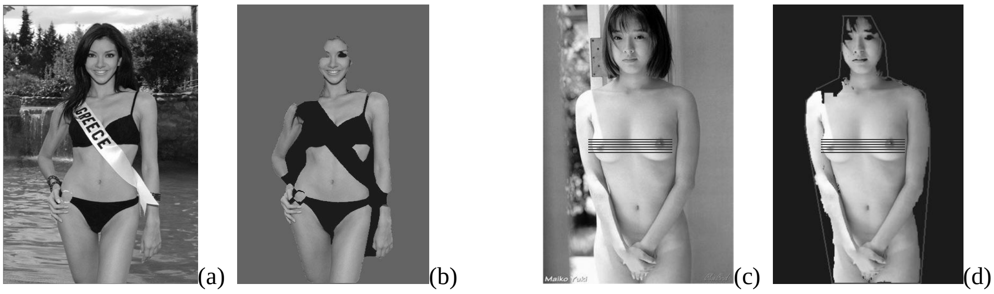
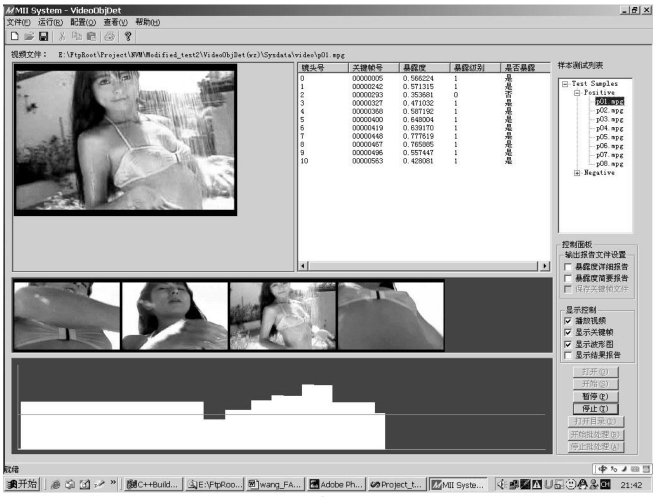


Figure 6: (a) Original image A, (b) For EL calculation of (a), (c) Original image B, (d) For EL calculation of (c).

Apparently, the EL value of Fig.6 (d) is higher than Fig.6 (b). If the key frames of video shots have high EL value, we can judge that the video includes potential erotic content.

5. EXPERIMENT AND CONCLUSION

We build a test system based on the descriptions above, which can automatically analyze the video stream and detect the video content to annotate whether it contain potential erotic content or not. Fig.7 shows the user interface of our test system.



视频文件： E:\FtpRoot\Project\NVM\Modified_test2\VideoObjDet(wz)\Sysdata\video\p01.mpg

镜头号	关键帧号	暴露度	暴露级别	是否暴露
0	00000005	0.566224	1	是
1	00000242	0.571315	1	是
2	00000293	0.353681	0	否
3	00000327	0.471032	1	是
4	00000368	0.587192	1	是
5	00000400	0.648004	1	是
6	00000419	0.639170	1	是
7	00000448	0.777619	1	是
8	00000467	0.765885	1	是
9	00000496	0.557447	1	是
10	00000563	0.428081	1	是

输出报告文件设置
□ 暴露度详细报告
□ 暴露度简要报告
□ 保存关键帧文件

显示控制
☑ 播放视频
☑ 显示关键帧
☑ 显示波形图
□ 显示结果报告

Figure 7: UI of our test system

We use Recall and Precision to be the testing norm. To test our system, we build a video library, which includes two types of videos: 1). Long shot: it contains many shots, some of which are erotic. 2). Short shot: it contains only one shot, either to be erotic or not. In this library there are 20 long shots and 112 short shots. The video content covers many aspects of human life as well as landscapes.

The experiment result is as follows:

Table 1 Experiment Result

	Detected erotic shots	Detected non-erotic shots	Erotic shots in lib	Non-erotic shots in lib	Precision	Recall
Long shot	198	24	214	131	89.2%	92.5%
Short shot	75	8	82	30	90.3%	91.5%

Experiment result shows the effectiveness of our method. The prototype system could be extended to be faster and more robust, which then can be used in security control in multimedia networking environment.

ACKNOWLEDGEMENTS

The research was supported by China Postdoctoral Science Foundation (No.32) and the Natural Science Foundation of Zhejiang Province (No.601055).

REFERENCES

1. Baoxin Li, M. Ibrahim Sezan, Event Detection and Summarization in Sports Video, IEEE Workshop on Content-based Access of Image and Video Libraries (CBAIVL'01), p. 132, December 14, 2001
2. Wensheng Zhou , Asha Vellaikal , C. C. Jay Kuo, Rule-based video classification system for basketball video indexing, Proceedings of the 2000 ACM workshops on Multimedia, p.213-216, October 30-November 03, 2000, Los Angeles, California, United States
3. Hisashi Miyamori , Shu-ichi Iisaku, Video Annotation for Content-based Retrieval using Human Behavior Analysis and Domain Knowledge, Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition 2000, p.320, March 26-30, 2000
4. M.M. Fleck, D.A. Forsyth and C. Bregler, ``Finding naked people,'' Proc. European Conf. on Computer Vision, Edited by: Buxton, B.; Cipolla, R. Berlin, Germany: Springer-Verlag, 1996. p. 593-602
5. Forsyth, D.A. and Fleck, M. M., ``Automatic Detection of Human Nudes,'' International Journal of Computer Vision , 32 , 1, 63-77, August, 1999
6. M. Soriano et.al. Skin color modeling under varying illumination conditions using the skin locus for selecting training pixels Real-time Image Sequence Analysis Workshop (RISA2000), August 31-September 1, Oulu, Finland, 43-49 (2000).
7. M. Störring Skin color detection under changing lighting conditions 7th Symposium on Intelligent Robotics Systems, pages 187-195, 20-23 July 1999, Coimbra, Portugal
8. M.J. Jones, J.M. Rehg, Statistical Color Models with Application to Skin Detection, Cambridge Research Laboratory Technical Reports, CRL98/11 Dec. 1998
9. J. Brand and J.Mason, A Comparative Assessment of Three Approaches to Pixel-level Human Skin-Detection, In International Conference on Pattern Recognition (ICPR'00)-VI, pp. 5056, Barcelona, September, 2000
10. Zhu miaoliang, Wang Donghui, Video stream segmentation method based on video page, Journal of Computer-Aided Design and Computer Graphics, Beijing, v12, n8, pp.585-589, August, 2000

* mii@zju.edu.cn; phone 86 571 8795-1916; fax 86 571 8795-1670; College of computer science, Zhejiang University.