Skin Paths for Contextual Flagging Adult Videos

Julian Stöttinger[1,2], Allan Hanbury[3],
Christian Liensberger[4], and Rehanullah Khan[1]

[1] PRIP, Vienna University of Technology, Austria
[2] CogVis Ltd., Vienna, Austria
[3] Information Retrieval Facilities, Vienna, Austria
[4] Microsoft, Redmond, Washington, USA

Abstract. User generated video content has become increasingly popular, with a large number of internet video sharing portals appearing. Many portals wish to rapidly find and remove objectionable material from the uploaded videos. This paper considers the flagging of uploaded videos as potentially objectionable due to sexual content of an adult nature. Such videos are often characterized by the presence of a large amount of skin, although other scenes, such as close-ups of faces, also satisfy this criterion. The main contribution of this paper is to introduce to this task two uses of contextual information in the form of detected faces. The first is to use a combination of different face detectors to adjust the parameters of the skin detection model. The second is through the summarization of a video in the form of a path in a skin-face plot. This plot allows potentially objectionable segments of videos to be found, while ignoring segments containing close-ups of faces. The proposed approach runs in real-time. Experiments are done on per pixel annotated and challenging on-line videos from an on-line service provider to prove our approach. Large scale experiments are carried out on 200 popular public video clips from web platforms. These are chosen from the community (top-rated) and cover a large variety of different skin-colors, illuminations, image quality and difficulty levels. We find a compact and reliable representation for videos to flag suspicious content efficiently.

1 Introduction

User generated content has become very popular in the last decade and has significantly changed the way we consume media [2]. With the international success of several *Web 2.0* websites (platforms that concentrate on the interaction aspect of the internet) the amount of publicly available content from private sources is vast and still growing rapidly.

The amount of video material being uploaded every day is too large to allow the operating companies to manually classify the content of every submitted video as appropriate or objectionable. The predominant methods to overcome this problem are to block contents based on keyword matching that categorizes user generated tags or comments. Additionally, connected URLs can be used to check the context of origin to trap these websites [8]. This does not hold true for

G. Bebis et al. (Eds.): ISVC 2009, Part II, LNCS 5876, pp. 303–314, 2009.

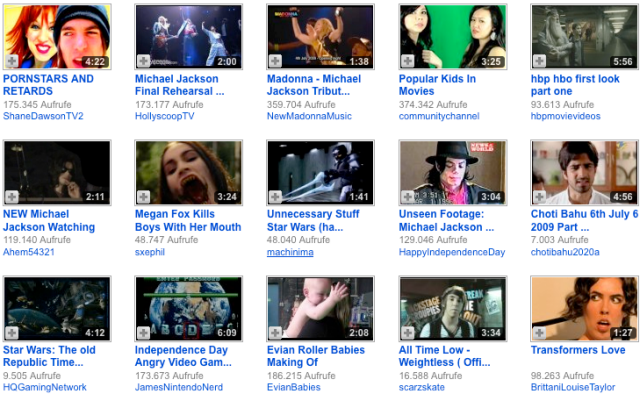


Fig. 1. Most popular videos from youtube.com on July 4th, 2009

websites like YouTube that allow uploading of videos. The uploaded videos are not always labeled by (valid) keywords for the content they contain (compare Fig. 1). As no reliable automated process exists, the platforms rely on their user community: Users flag videos and depending on this, the administrators may remove the videos flagged as objectionable. This method is rather slow and does not guarantee that inappropriate videos are immediately withdrawn from circulation. A possible solution for rapid detection of objectionable content is a system that detects such content as soon as it is uploaded. As a completely automated system is not feasible at present, a system that flags potentially objectionable content for subsequent judgement by a human is a good compromise. Such a system has two important parameters: the number of harmless videos flagged as potentially objectionable (false positive rate), and the number of objectionable videos not flagged (false negative rate). In the context of precision and recall of a classification application, these two parameters present a trade-off. For a very low false negative rate, a larger amount of human effort will be needed to examine the larger number of false positives. These parameters should be adjustable by the end-users depending on the local laws (some regions have stricter restrictions on objectionable content) and the amount of human effort available. A further enhancement to reduce the amount of time required by the human judges is to flag only the segments of videos containing the potentially objectionable material, removing the need to watch the whole video, or search the video manually. One reason why videos may be considered objectionable is due to explicit sexual content. Such videos are often characterized by a large amount of skin being visible in the frame, so a commonly used component for their detection is a skin detector [8,15]. However, this characteristic is also satisfied by frames not considered as objectionable, most importantly close-ups of faces.

This paper considers the flagging of user-uploaded videos as potentially objectionable. The main contribution of this paper is to introduce two uses of contextual information in the form of detected faces. The first is to use tracked faces

to adjust the parameters of the skin detection model. As it is shown in Fig. 1, user generated content contains many faces. We develop classification rules based upon a prior face detection using the well known approach from Viola et al. [13]. This work builds on [7] where it is shown that more precise adaptive color models outperform more general static models especially for reducing the high number of false positive detections. In [9] it is shown that humans need contextual information to interpret skin color correctly. We extend their approach by using a combination of face detectors: We combine frontal face detection and profile face detection in a combined tracking approach for more contextual information in the skin color representation.

The second use of face information is through the summarization of a video in the form of a path in a skin-face plot. This plot allows potentially objectionable segments of videos to be extracted, while ignoring segments containing close-ups of faces. We show that the properties of the skin paths give a reliable representation of the nature of videos. The proposed approach was kept algorithmically simple, and currently runs at over 30 frames per second. A high level of performance is required in such an application to cope with the large number of uploaded videos.

In Section 2 we summarize some related work, while Section 3 describes our multiple model approach for fast skin detection with face information, as well as our summarization of the videos on the skin-face plot. The experiments and results are presented in Section 4. Section 5 concludes.

2 Related Work

In computer vision, skin detection is often used as a first step in face detection, e.g. [11], and for localization in the first stages of gesture tracking systems, e.g. [1]. It has also been used in the detection of naked people [4,8]. The latter application has in most cases been developed for still images.

The approaches to classify skin in images or videos can be grouped into three types of skin modeling: parametric, nonparametric and explicit skin cluster definition methods. The parametric models use a Gaussian color distribution since they assume that skin can be modeled by a Gaussian probability density function [14]. Non-parametric methods estimate the skin-color from the histogram that is generated by the used training data [5].

An efficient and widely used method is the definition of classifiers that build upon the approach of skin clustering. This thresholding of different color space coordinates is used in many approaches, e.g. [10] and explicitly defines the boundaries of the skin clusters in a given color space. The underlying hypothesis is that skin pixels have similar color coordinates in the chosen color space, which means that skin pixels are found within a given set of boundaries in a color space. Although this approach is extremely rapid, its main drawback is a comparably high number of false detections [6]. We are able to compensate for this issue in our approach by using a multiple adaptive model approach and contextual information in the form of faces.

3 Method

In this section we describe the adaptive skin-color modeling in detail. We address the problem of changing light conditions, different skin colors and varying image quality in videos in adapting the skin color model according to reliably detected faces. Figure 2 gives an overview of the main steps. We can do the face detection and tracking (see Section 3.1) and the color conversion (Section 3.2) in parallel on the input frame. With this data, we can build our skin model and propagate it to adjust to present skin-color variations and illumination changes in Section 3.3 for a robust skin color classification.

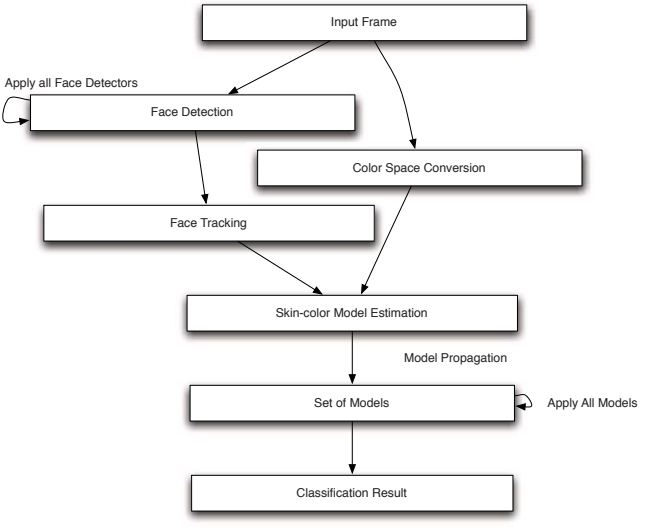


Fig. 2. Overview of the proposed method

3.1 Face Detection and Tracking

Due to its real-time performance, we use the face detector proposed by Viola et al. [13], as done by Khan et al. [7]. Opposed to their approach, we run profile face detectors and frontal face detectors in parallel. We track faces in the videos to adjust the model propagation strategy. Additionally, false positive detections are likely to "pop out" of the background for a short time, which can be easily suppressed with a tracking algorithm. Color or feature based tracking techniques may fail when there are large changes in the illumination, the facial expression or in the viewpoint. We rely on a geometrical approach that removes every false positive in the annotated data set and lets us track faces from one detector to the other. For every given detection of detector D^i where $i = 1..n$ is the detector

identifier and n the number of detectors, we merge every dependable detection for any frame m by

$$\bigvee_{i=1}^{n} (D_m^i \cap D_{m-1}^i) \wedge (D_m^i \cap D_{m+3}^i) \geqslant 0.5 \tag{1}$$

We merge cases where multiple detectors give the same faces. When a head is turned, profile face detections are merged with frontal faces over time.

3.2 Skin-Color Modeling

Choosing a color space that is relatively invariant to minor illuminant changes is crucial to any skin color tracking system. The transformation simplicity and explicit separation of luminance and chrominance components makes $YCbCr$ attractive for skin color modeling [12]. For 24 bit color depth, the following values apply:

$$\begin{aligned} Y &= (0.299 * (R - G)) + G + (0.114 * (B - G)) \\ Cb &= (0.564 * (B - Y)) + 128 \\ Cr &= (0.713 * (R - Y)) + 128 \end{aligned}$$

The favorable property of this color space for skin color detection is the stable separation of luminance, chrominance, and its fast conversion from RGB. These points make it suitable for our real-time skin detection. The static values used when initially no face is detected are [3]: $Cb_{max} = 127, Cb_{min} = 77, Cr_{max} = 173, Cr_{min} = 133$. If we do not detect any face in a video, these values are the general static skin-color model. In that case, we do not gain any advantage from this approach. These values apply to a very broad range of illumination circumstances and a range of skin color, thus having a large number of false positives. However, it overlaps significantly with the idea humans have about skin-color [9]. Our approach for more specific skin-color models is explained in the next Section.

3.3 Skin Color Model Instance Initialization and Destruction

Any detected and tracked face introduces a new skin color model instance, which allows skin of different color and under different light conditions to be detected. After a face has been detected its color is examined: The range for the Cb and Cr components (of the $YCbCr$ color space) are used to generate a newly adapted range model. The Y component is ignored since it encodes only the luminance.

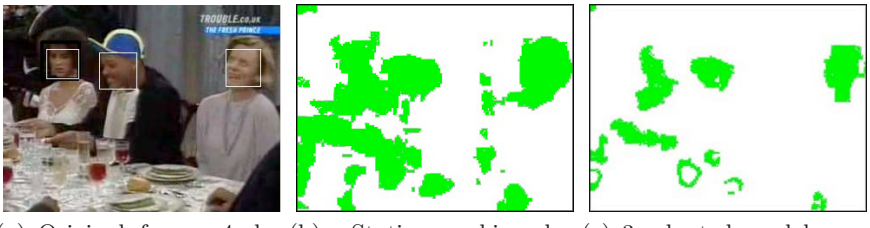
In a first step, we have to estimate how much skin is present in the detected face. Detections usually contain certain parts that are not skin, such as hair, open eyes, mouth, eye brows etc. Therefore we statically cluster skin-color with the borders defined in Section 3.2. Having the possible skin region extracted,

(a) Original frame, 1 detected faces.

(b) Static skin-color model, green indicates detected skin.

(c) 1 adapted model applied, green indicates detected skin.

Fig. 3. Video 2 example frame and its classification result with a near skin-color background



(a) Original frame, 4 detected faces. The left one is detected twice.

(b) Static skin-color model, green indicates detected skin.

(c) 3 adapted models applied, green indicates detected skin.

Fig. 4. Video 5 example frame and its classification result. There are 4 different skin-colors present in the scene, 4 detected faces from two different detectors leading to 3 connected color models.

we adjust the model: The average skin pixel color gives the median value for the new skin-color model. As evaluated in [9], the Cb channel gives best performance using a range of 30% of the static color model, the Cr channel is more stable having a range $17,5\%$. With this adapted cluster definitions, we are able to classify every pixel by 4 simple threshold operations per model. This contextual information makes the approach more precise reducing the number of false positives (compare Fig. 3).

Additionally, in the same manner we track multiple faces, we are able to track multiple skin-colors per scene robustly, as it can be seen in Fig. 4. A tracked face adapts its skin-color model dynamically per frame. When the face is lost by the tracker, a possible re-detection of the face gives a second very similar skin-color model. We do not forget about any model we create in the course of the video. This is an assumption which works well for rather short online video clips with a limited number of persons, but does not hold for long movies with a grand variation of illumination circumstances and actors.

3.4 Skin Paths for Video Classification

In detecting adult video material, we are interested in the amount of skin visible. It is possible to visualize this information in the form of skin graphs, showing the number of skin pixels detected per frame [7]. However, a major problem of such skin color based classification systems is that e.g. portrait shots like interviews and news do have a large amount of skin present in the scene, which makes a decision based on skin pixel count difficult.

After a successful face detection, we overcome this problem by estimating the relation between skin inside the face region and the whole frame. This measure gives an idea about the scale of the people in that shot. Plotting this measure against the overall skin detection, we are able to describe the property of the given frame meaningfully. Videos differ heavily from scene to scene and from shot to shot. To get an idea of a video, we have to provide a compact representation for our detection method.

We introduce skin paths, which average the described measure over a fixed number of frames and give an intuitive idea of the character of a given video. In Fig. 5, the skin path for video number 9 and the corresponding frames are given. On the x-axis, the mean quotient of the skin color area inside a tracked face and the skin coverage of a fixed number of frames is given. The y-axis represents the mean total skin color coverage in these frames. The path starts at the very left, as in the beginning there is no face detected and much sand is detected as skin. Following the path, more faces are detected and the skin color model adjusts towards the right color model, giving a better idea about the amount of skin present in the scene. We show in the following Section 4, that there are certain areas in the skin graph correlated with properties and content of the videos. From the information of the skin paths, we can categorize the nature of videos reliably by the position of the data points of the graphs. Additionally, there is a trend that data points with $x = 0$ provide more unreliable results as there are none or few faces detected. This gives a confidence measurement for the skin detection itself.

4 Experiments

In this Section, we develop a robust classification rule for flagging adult videos and prove our approach on a large data set containing the most popular online videos. In the following Section 4.1, we describe the videos used in more detail. Section 4.2 evaluates the approach on a per pixel basis and proposes the classification rules that are evaluated in a large scale experiment in Section 4.3.

4.1 Data Sets

15 videos have been provided by an internet service provider that requires a skin detection application for their on-line platform. Their aim was to choose challenging videos with near skin-color backgrounds. Pink and brown backgrounds

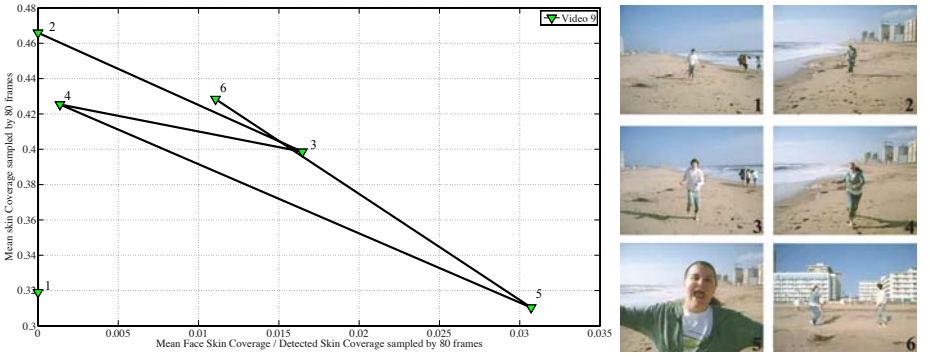


Fig. 5. Skin path for the classification of Video 9 and the corresponding key frames. The detection and incorporation of facial skin makes the results more reliable.

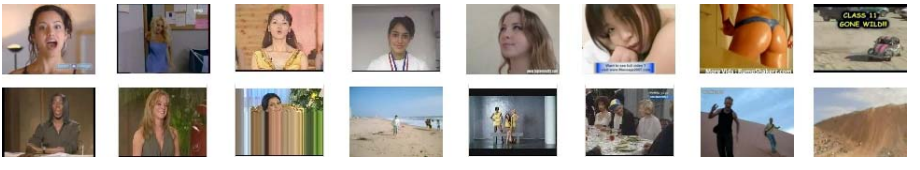


Fig. 6. Example frames from the annotated video data-set used

such as beaches, sand, cork boards or similar are detected as false positives easily (see Fig. 3 and compare Fig. 6). We added 10 videos to encounter additional challenges as a larger variety of skin-colors, especially different skin-colors in one frame. Most of the sequences also contain scenes with multiple people and/or multiple visible body parts and scene shots both indoors and outdoors, with steady or moving camera. The lighting varies from natural light to directional stage lighting. Sequences contain shadows and minor occlusions. Collected sequences vary in length from 100 frames to 500 frames. They also contain data errors and are generally of poor quality, varying size and frame rate. Ground truth has been generated for all of the 25 videos on a per pixel basis annotating 10764 frames manually.

The second data set consists of 200 publicly available videos. To provide an objective collection of videos, we chose the 100 most popular videos from youtube[1] on July 7th, 2009. As there are not more videos available in this category, we additionally gathered 50 videos "being watched right now" which are not in the previous category. For the adult material, we chose the 50 most popular videos from youporn[2] as this platform provides explicit adult material only and is publicly available. We want to make sure that our classification is not biased by the

[1] `http://www.youtube.com`
[2] `http://www.youporn.com`

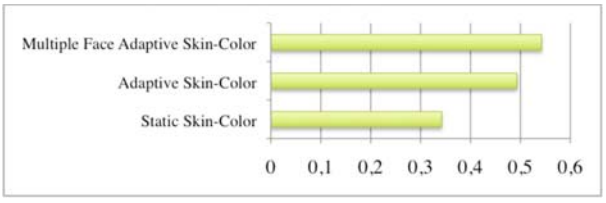


Fig. 7. The Fscore of the classification results. The proposed approach using multiple face detectors outperforms the adaptive skin-color modeling [9] and the static skin-color classification [3].

two different data sources. There is a probability that the two classes of video material differ e.g. in frame rate, size, video quality or noise level just because of the two platforms they are downloaded from. Such criteria would nullify any classification success. Therefore we chose 10 videos of the adult material where we encounter rather extended non-adult scenes and deleted the adult scenes out of it. Finally, the second data set consists of 160 videos with non-adult material (100 most popular, 50 being watched, 10 edited adult material) and 40 videos with explicit content. Example frames are shown in Fig. 1.

4.2 Adult Video Classification and Detection

Fig. 7 shows that the use of multiple face detectors provides a significant increase of classification performance of almost 10% compared to single face detection and a combined color space voting [9]. The weighted harmonic mean indicates the per pixel evaluated classification results. In Fig. 8 the skin paths for the whole data-set are given. As it is shown by the red paths, adult material tends to have few and small faces compared to a large amount of skin present. This intuitive criteria is well suited for classification of videos in the skin path diagram: We make the assertion that the skin path of suspicious video material enters the area defined by x ¡ $x < 0.08$ and $y > 0.55$. We classify our whole data-set correctly with two false positive detections of videos. These two contain desert shots without faces present.With this classification technique, we can detect adult material reliably with the tendency to get false positive detections but very few false negative ones. Additionally, the distance to the upper left corner gives an idea of the character of the scene: Video 8 and 21 sporting girls in bikinis show to have a path beginning below $x = 0.5$, $y = 0$ (and therefore near adult material) and smoothly adapt themselves towards the lower right (towards the unsuspicious space). Videos without much non-facial skin visible (e.g. Interviews) have skin paths significantly towards the bottom.. The main areas are summarized in Table 1: Adult is the zone where we encounter adult material to be flagged, suspicious videos tend to show persons in full with lots of visible skin as they appear e.g. in sports clips. 5 videos are classified as suspicious material. They are intuitively correct as they are beach, dance and massage scenes. We separate videos with

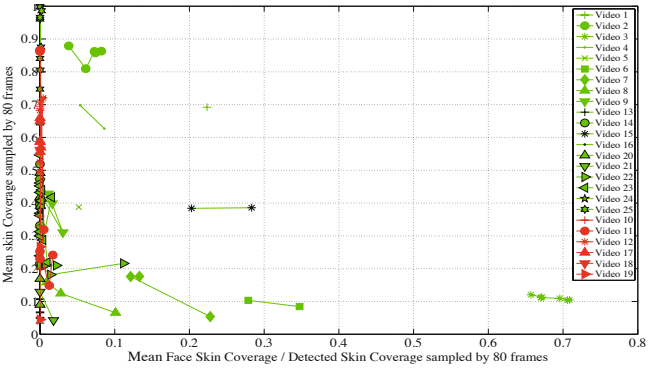


Fig. 8. Skin paths of the relation between facial skin pixels and other skin pixel drawn onto the overall skin coverage. Red indicates adult material, green unsuspicious video material.

Table 1. The 3 main characters of videos that can be extracted from the skin paths reliably and their classification performance on the annotated data set

character of video	classification rule	classification accuracy	absolute false negatives
adult material	$x < 0.08, y > 0.55$	0.85	0
suspicious	$x < 0.08, y > 0.43$	-	-
portrait	$x \gg 0.08, y < 0.25$	1	0

portrait shots as there are in news, interviews and most of the "webcam" video messages robustly into the portrait area with an accuracy of 1. We show in the next section that these values hold for arbitrary online videos as well.

4.3 Flagging Adult Online Videos

We repeat the experiment from the previous Section on the 200 online videos described in Section 4.1. As it can be seen in Fig. 9, adult material has again a strong trend towards the upper left corner. We apply the classification rules defined for adult material and reach an accuracy of 0.91. The two false negatives are both classified as suspicious character, which can be seen in the second line of Tbl. 2. The reason for this wrong classification is in the nature of the two videos: The first one is of explicit adult material, but almost no skin and no face is visible as the actors are dressed fully. The other false positive contains a couple that apparently takes the video with their webcam on their own. In contrast to all other adult videos, the actors appear rather small in the image. Although the skin color is detected precisely, the amount of skin is not enough for our adult material classification rule. For portrait videos, 54 videos are classified as portrait videos. They all contain portrait shots. We do not encounter false positive detections of tracked faces, although we do not know precisely how many we missed.

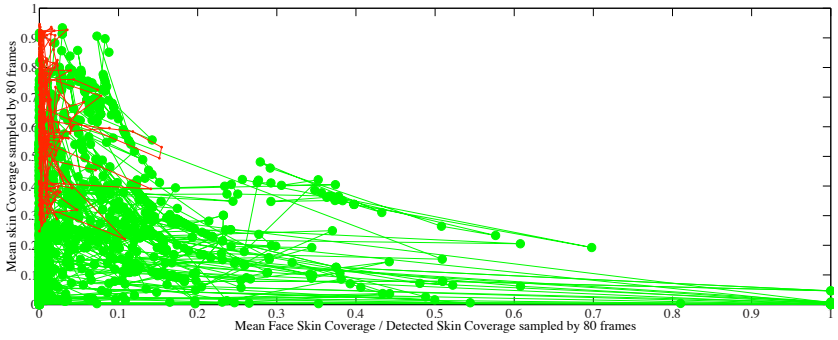


Fig. 9. Skin paths of the relation between facial skin pixels and other skin pixel drawn onto the overall skin coverage. Red indicates adult material, green unsuspicious video material.

Table 2. Classification accuracy and absolute number of false negatives and false positives. 1st line shows flagging performance of adult material, second line for both adult and suspicious material.

classification rule	classification accuracy	nr. of false negatives	nr. of false positives
$x < 0.08, y > 0.55$	0.91	2	15
$x < 0.08, y > 0.43$	0.82	0	37

5 Conclusion

We present a practical approach to detecting skin in on-line videos in real-time. Instead of using solely color information, we include contextual information in the scene through multiple face detection and combined face tracking. By using a combination of face detectors and an adaptive multiple model approach to dynamically adapt skin color decision rules we are able to significantly reduce the number of false positive detections and the classification results become more reliable compared to static color threshold based approaches or approaches using multiple color spaces. The runtime of the algorithm is still real-time and can be carried out in parallel.

We give the skin path as a compact and powerful representation of videos. We are able to extract reliable features from facial and non facial skin and classify on-line videos successfully. The number of false negatives is very low, providing a reliable flagging of adult material. The approach is computationally inexpensive and can be carried out in real-time.

Acknowledgment

This work was partly supported by the Austrian Research Promotion Agency (FFG), project OMOR 815994, and the CogVis[3] Ltd. However, this paper reflects

[3] `http://www.cogvis.at/`

only the authors' views; the FFG or CogVis Ltd. are not liable for any use that may be made of the information contained herein.

References

1. Argyros, A.A., Lourakis, M.I.: Real-time tracking of multiple skin-colored objects with a possibly moving camera. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3023, pp. 368–379. Springer, Heidelberg (2004)
2. Cha, M., Kwak, H., Rodriguez, P., Ahn, Y.-Y., Moon, S.: I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In: Int. Conf. Internet Measurement, pp. 1–14 (2007)
3. Chai, D., Ngan, K.N.: Locating facial region of a head-and-shoulders color image. In: Int. Conf. Automatic Face and Gesture Recognition, pp. 124–129 (1998)
4. Fleck, M.M., Forsyth, D.A., Bregler, C.: Finding naked people. In: Buxton, B.F., Cipolla, R. (eds.) ECCV 1996. LNCS, vol. 1064, pp. 593–602. Springer, Heidelberg (1996)
5. Jones, M.J., Rehg, J.M.: Statistical color models with application to skin detection. IJCV 46(1), 81–96 (2002)
6. Kakumanu, P., Makrogiannis, S., Bourbakis, N.: A survey of skin-color modeling and detection methods. PR 40(3), 1106–1122 (2007)
7. Khan, R., Stöttinger, J., Kampel, M.: An adaptive multiple model approach for fast content-based skin detection in on-line videos. In: Int. Workshop Analysis and Retrieval of Events/Actions and Workflows in Video Streams (2008)
8. Lee, J.-S., Kuo, Y.-M., Chung, P.-C., Chen, E.-L.: Naked image detection based on adaptive and extensible skin color model. PR 40(8), 2261–2270 (2007)
9. Liensberger, C., Stöttinger, J., Kampel, M.: Color-based and context-aware skin detection for online video annotation. In: MMSP (to appear, 2009)
10. Phung, M.-S.L., Bouzerdoum, S. M.-A., Chai, S. M.-D.: Skin segmentation using color pixel classification: Analysis and comparison. PAMI 27(1), 148–154 (2005)
11. Senior, A., Hsu, R.-L., Mottaleb, M.A., Jain, A.K.: Face detection in color images. PAMI 24(5), 696–706 (2002)
12. Vezhnevets, V., Sazonov, V., Andreev, A.: A survey on pixel-based skin color detection techniques. In: ICCGV, pp. 85–92 (2003)
13. Viola, P., Jones, M.J.: Robust real-time face detection. IJCV 57(2), 137–154 (2004)
14. Yang, M., Ahuja, N.: Gaussian mixture model for human skin color and its application in image and video databases. In: SPIE, pp. 458–466 (1999)
15. Zheng, H., Daoudi, M., Jedynak, B.: Blocking adult images based on statistical skin detection. ELCVIA 4(2), 1–14 (2004)