Color-Based and Context-Aware Skin Detection for Online Video Annotation

Christian Liensberger, Julian Stöttinger, Martin Kampel

Institute of Computer Aided Automation,
Pattern Recognition and Image Processing Group,
University of Technology Vienna, Austria
[liensberger|julian|kampel]@prip.tuwien.ac.at

***Abstract*—By analyzing the low level features of images only, skin detection in visual data cannot be solved. To compensate for this major drawback of many approaches, we combine a state of the art recognition algorithm with color model based skin detection. Detected faces in videos are the basis for adaptive skin-color models, which are propagated throughout the video, providing a more precise and accurate model in its recognition performance than pure color based approaches. The approach is able to run in real-time and does not need prior data-specific training. We received challenging online videos from an online service provider and use additional videos from public web platforms covering a grand variety of different skin-colors, illumination circumstances, image quality and difficulty levels. In an extensive evaluation we estimated the best performing parameters and decide on the best model propagation techniques. We show that adaptive model propagation outperforms static low level detection.**

I. Introduction

User generated content became very popular in the last decade. With the international success of several Web 2.0 websites (focusing on the interaction aspect of the internet) the amount of publicly available content from private sources is growing rapidly.

The state of the art approach to block content on the internet is based on contextual keyword pattern matching technology [1]. This approach has certain drawbacks for websites which allow uploading videos by the user, like e.g. YouTube[1]. The uploaded videos do not explicitly need to report textual information for the content. Since there is no reliable automated process available the classification is done voluntarily by the user community. Therefore, an automated method to categorize videos based on skin-color will help the service providers to gain additional information about the videos as soon as they have been uploaded to the platform. Skin-color detection is also used as a preliminary step in a broad range of computer vision tasks, including gesture analysis, tracking [2], or content-based image retrieval systems [3]. We evaluate a fast and straightforward adaptive skin detection method for videos. The decision rules of our method adapt upon a first step of face detection using the well-known approach from Viola et al. [4]. This leads to the proposed method which takes high advantage of the temporal relationship between frames in an image sequence and deals well with time dependent illumination changes.

There are other approaches that rely upon a successful face detection for skin classification, e.g. [5]. We extend their propagation model and add trainable parameters to the framework. That gives us a fast classification technique for low quality videos using multiple models at one time [6]. Additionally, we use different color spaces which are combined using voting. The method can be carried out in a real-time classification system and is therefore useful for an automated pre-selection and classification for large video databases.

We suggest a method for extracting meaningful key frames from the videos for the possible task of filtering adult content in the video. Based on the results of the skin coverage graph we are able to extract the meaningful frames for further manual classification.

In the following Section II we describe the state of the art in low level skin detection and its adaption towards time-varying color circumstances and video segmentation. Section III describes the multiple model approach for fast skin detection. The experiments and results are outlined in Section IV. A conclusion is given in Section V.

II. Related Work

We aim to build a decision rule that will discriminate between skin and non-skin pixels for skin-color detection and the classification of skin contents. The ways to classify skin by color in videos may be divided into three types: 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 [7]. The second approach, nonparametric methods, estimate the skin-color from the histogram that is generated by the training data used [8]. The third approach relies on thresholding of different color space coordinates and is used in many approaches, e.g. [9]. It explicitly defines the boundaries of the skin clusters in a given color space. The underlying hypothesis here is that skin pixels have similar color coordinates in the chosen color space, which means

[1]http://www.youtube.com

MMSP'09, October 5-7, 2009, Rio de Janeiro, Brazil.
978-1-4244-4464-9/09/$25.00

Fig. 1. Example frames from the video dataset used.

that skin pixel are grouped in a color space. The main drawback of this method is the comparably high number of false detections [10]. We are able to compensate for this issue in our approach by using the multiple adaptive model approach.

Color is a low level feature which is broadly used for real-time object characterization, detection and localization [1]. Following [10], the major difficulties in skin-color detection are caused by illumination circumstances, camera characteristics, ethnicity and individual characteristics of the displayed persons. Regarding user-generated video content, we face additional problems: Online portals restrict the resolution of their videos to minimize their server load and bandwidth. YouTube, for example, recommends a resolution of 640x360 for 16:9 or 480x360 for 4:3[2]. Moreover, capture devices with low aperture like mobile phones and web cams produce a higher amount of noise than professional devices. Additionally, many videos are compressed several times in the work flow of user-generated video publishing, including on the user side and on the platform, e.g. YouTube[1]. Dealing with the typical amount of data that has to be processed by a video platform (YouTube allows uploads of files up to 1 GB[1]) the runtime of the algorithm should be real-time or faster to be of use for this task. Additionally, no presumptions can be made about the appearance of skin or scene circumstances. Varying lighting conditions might also appear within the video itself.

A. Skin-color

To model and classify skin-color properly the choice of the appropriate color space is crucial. Clusters in normalized RGB are an appropriate model for skin-color [11]. Still, the normalized RGB color space suffers from instability with dark colors. The $HS*$ color spaces are known to be invariant to illumination change. This property is helpful in the process of skin detection and that is why they are often used to detect skin in images [12]. Orthogonal color spaces like YC_bC_r, YC_gC_r, YIQ, YUV, YES try to form as independent components as possible. YC_bC_r is one of the most successful approaches for skin-detection and used by e.g. [13], [14]. A single color space may limit the performance of the skin-color filter and that better performance can be achieved using two or more color spaces. Using the most distinct invariant color coordinates of different color spaces increases the performance. The combination of different color spaces also eliminates false positives since the combination stabilizes the area that is used for skin detection.

[2]http://www.youtube.com/handbook_popup_produce_upload?pcont=bestformats

B. Skin detection

In many approaches e.g. [13] pixel level skin detection is used as one of the first steps for a successful face detection, face recognition and gesture tracking. This is a valuable assumption since the human face is often not completely covered and at least some skin is visible.

Viola et al. [4] introduced a stable face detection algorithm based on their integral image, Haar-like features and a cascade structure that applies more specialized filters as the cascade is walked. The algorithm is applicable in real-time (see Section III). The performance and simplicity of the face detector inspired several authors for using this approach as an initial step for further skin-color estimation [5]. In contrary to our approach they use more sophisticated classifiers, which rely on different assumptions and just one model at a time.

Skin detection under varying illumination in image sequences is addressed in [15]. Some of these approaches try to map the illuminance of the image into a common range to assure that skin always exposes the same luminance and tone.

Neural networks [16], Bayesian Networks e.g. [11], Gaussian classifiers e.g. [8] or self organizing maps (SOM) [17] are high level classifiers that try to overcome issues of low level classifier and try increase the classification accuracy. These methods typically demand long training times and are too slow for real-time classification. Therefore they are not suitable for high speed classification as required in our scenarios.

III. Method

We address the problem of changing lighting conditions, different skin-colors and varying image quality in videos in adapting the skin-color model according to reliably detected faces. Prior to any detected face, the combination of the static YC_bC_r, normalized RGB and RGB skin model is applied for skin detection. These three color spaces are used in a combination: Two votes out of the three color spaces make the decision final. In an extensive evaluation (see Section IV) we show that voting to be more robust than only using one of the three color spaces.

Due to its real-time performance, Khan et al. [6] use the Viola et al. [4] face detector in their model propagation. Wimmer et al. [5] point out that the performance of this detection algorithm allows a precise and reliable estimation of

the skin-color. In our approach any detected face introduces a new skin-color model, which allows to detect skin of different color and under different lighting conditions. In case there is no face detected static color voting is used for the whole video.

A. Skin sample localization through face detection

Viola-Jones use a set of features which are similar to Haar basis functions but also extend them to complexer features that can not be modeled with the Haar basis functions only. The integral image, was introduced to model these features rapidly at many scales. Computing the integral image is done with only a few operations per pixel:

$$ii(x,y) = \sum_{x' \leq x, y' \leq y} i(x', y') \tag{1}$$

where $ii(x,y)$ is the integral image and $i(x',y')$ is the original image.By using the following pair of recursions:

$$s(x,y) = s(x,y-1) + i(x,y) \tag{2}$$

$$ii(x,y) = ii(x-1,y) + s(x,y) \tag{3}$$

(where $s(x,y)$ is the cumulative row sum, $s(x,-1) = 0$ and $ii(-1,y) = 0$), the integral image can be computed in one pass over the original image. Once the integral image is computed the Haar-like features can be evaluated in constant time.

They construct a classifier that selects only a small subset of important features by using Ada-Boost [18]. The Ada-Boost learner is modified so that each returned weak classifier can depend on only one single feature. This modification is based on the work of Tieu and Viola [19]. As a result each stage of the boosting process that selects a new weak classifier can be seen as a feature selection process. The benefit of Ada-Boost is that it provides an effective learning algorithm and is also very strong on generalization performance [20].

A low false positive detection rate is crucial for our approach. We try to overcome this problem by tracking detections in a simple and fast way over time. If a face cannot be tracked below a certain time threshold it is discarded as noise and not used for the skin detection at all. With that threshold all of the false detections got eliminated (see Eq. 5). Additionally the algorithm detects a bigger area as face as the one that is found in the ground truth. This behavior might also be caused by the compression and only pops out for a few frames. We overcome this problem by applying a simple geometric tracking approach in Section III-D

B. Color Space for Skin-Color Tracking

The transformation simplicity and explicit separation of luminance and chrominance components makes YC_bC_r attractive for skin-color modeling [14]. YC_bC_r is an encoded nonlinear RGB signal, commonly used by European television studios and for image compression work, such as JPEG and MPEG.

$$\begin{aligned} Y &= (0.299 * (R-G)) + G + (0.114 * (B-G)) \\ C_b &= (0.564 * (B-Y)) + 128 \\ C_r &= (0.713 * (R-Y)) + 128 \end{aligned} \tag{4}$$

It is constructed as a weighted sum of the RGB values, and two color difference values C_b and C_r that are formed by subtracting luma from RGB's red and blue components. For 24 bit color depth, it can be estimated in linear time (compare Eq. 4).

C. Introducing a new face as model

We rely on a computationally simple face tracker and confidence check for reasons of the runtime of the algorithm: For every given set of detected pixel A in the frame number n it is regarded as a new and trustful face detection if

$$(A_n \cap A_{n-1}) \wedge (A_n \cap A_{n+3}) \geqslant 0.5 \tag{5}$$

The parameters $n+3$ and 0.5 are chosen after the parameter training. In other words, we regard a face as reliable detection, if it is present in 3 subsequent frames and both its position and its size does not change by more than 50%. In the prototype, the operation $\cap$ is implemented as a logical AND between binary arrays of detected and not detected pixels and gives the overlapping area of two detections. This condition suppresses every background detection in the test set given in Section IV as they tend to "flicker" through the scene. No true face is disregarded.

Once the face is lost the model is still applied, unless a new face is found. The idea behind this approach is that in subsequent frames the face detector may fail to detect a face, because of occlusion, face rotation etc.

The main assumption of this approach is that a detected face contains a certain amount of skin-color and is the base for a new model. The Viola-Jones face detection system returns detected faces as a square that contains the face. The square is arranged in such a way that it covers hair and parts of the background to the left and right region of face. In case we add background information to the model generation, we end up with a wrong assumption about the skin color. We chose the straight forward rule of truncating the square by 30% on every border to make sure that only the skin area of the face is returned. With this smaller area as a basis, we start the adaptive skin color modeling described in the following Section III-D.

D. Adaptive skin-color modeling

At the starting frame we use the static range (multiple color spaces combined by voting) for skin detection (the explicit values are given in [21]). After a face has been detected its appearance is examined: The range for the C_b and C_r components are used to generate a newly adapted range model. The Y component is ignored since it encodes only the luminance. We use every detected face in each frame to adapt our model continuously as the lighting in the scene changes or a new face is introduced to the sequence.

We do not use the original C_b and C_r ranges that has been found in the face but rather generate the average values for each of them. We choose the mean color $M_{[Cb|Cr]}$ of the extracted face region A becoming the model center. The ranges of the models are estimated by using "clamping" values R_{Cb}

and R_{Cr} for the according the two channels. A successful skin detection is given if

$$[(R_{Cb} - M) < M_{Cb} < (R_{Cb} + M)] \wedge [(R_{Cr} - M) < M_{Cr} < (R_{Cr} + M)] \quad (6)$$

holds. It is applied since the detected facial area usually still contains certain parts that are not skin, such as possible open eyes, mouth, eye brows etc. The clamping values are percentage values of the static ranges for the C_b and C_r channels. We model the adapted skin-color borders as a range by starting from the detected average and expanding it in both directions by the percentages defined in the clamping.

We evaluate the skin detection accuracy under varying R and determine the best performing solution in Section IV-C. A separate model is used for each separate face and the skin is detected on base of these models. This approach solves the problem of multiple people with different skin tones.

IV. Experiments

Our experiments include evaluation of the key parameters of our approach, a set of color spaces and an online poll to understand how people classify skin. All the videos used in the experiments have been annotated by hand to provide a valid per pixel ground truth that we can evaluate the videos against. Throughout this Section, the accuracy of the detection approaches is detected per pixel wise location (compare annotation in Fig. 4). In the following Section IV-A, we show that humans are - as computer vision approaches - not able to detect skin reliably without knowledge of the context of the scene.

A. Skin classification by humans

To better understand the importance of context for humans when classifying skin we carried out an online poll where people were asked to rate fragments of images as whether they contain skin or not. We made sure a lot of people were reached by publishing the poll in several web portals where there is a broad distribution of visitors coming from different parts of the world. In total we got 403 people from six continents participating who rated 18338 fragments.

The poll consisted of a set of random frames from the videos in our dataset. To remove eventual context these frames were cut into small fragments (see Figure 2). Each person who participated got presented with a random set of fragments and asked whether each fragment contained skin and how much skin was visible.

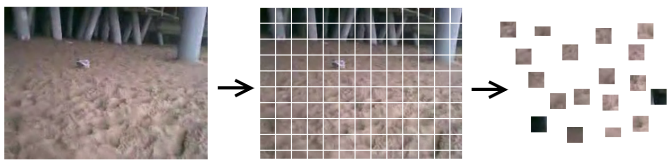

Fig. 2. Images where cut into fragments to remove the context.

Fig. 3. Example frame Video 25. Humans do not classify skin properly without context. The darker the color the more people misclassified it as skin.

Fig. 4. Example frames with the generated ground truth.

The results pointed out that humans are not able to detect skin without context. Near skin color materials like sand or wood is likely to be misinterpreted. They tend to fall back to only use color and in some scenarios with skin-color like material fail completely (see Figure 3).

B. Dataset composition and ground truth

The dataset used is a set of 25 videos. Half of it represents a random selection from YouTube[3]. The other half is a set of random videos that has been given by an external company to evaluate our approach for their future products. Most of the sequences also contain scenes with multiple people, multiple visible body parts and scenes shot 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. The videos vary in length from 100 frames to 500 frames. They are generally challenging as they contain skin-color like content as pink backgrounds, beaches, sand, cork boards which are easily detected as false positives (see Figure 1). For all of the videos ground truth has been

[3]www.youtube.com

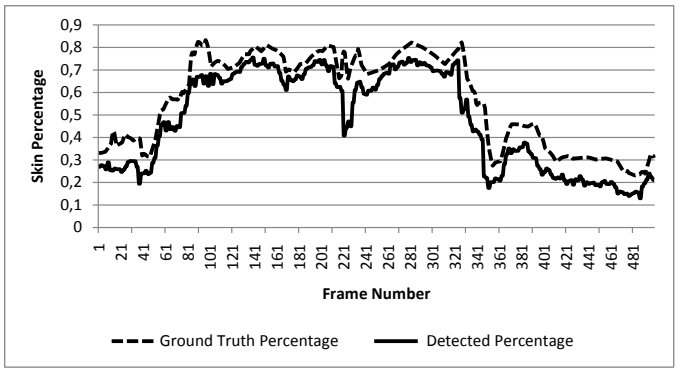

Fig. 5. Graph showing the detected skin distribution percentage compared with the ground truth data in one of the videos of our dataset. This video has been run with face detection. C_b range: 30%, C_r range 17.5%.

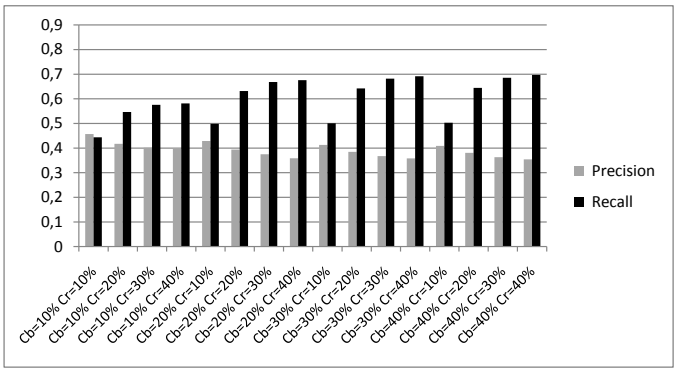

Fig. 6. The average dataset results by using various range parameter values for C_b and C_r.

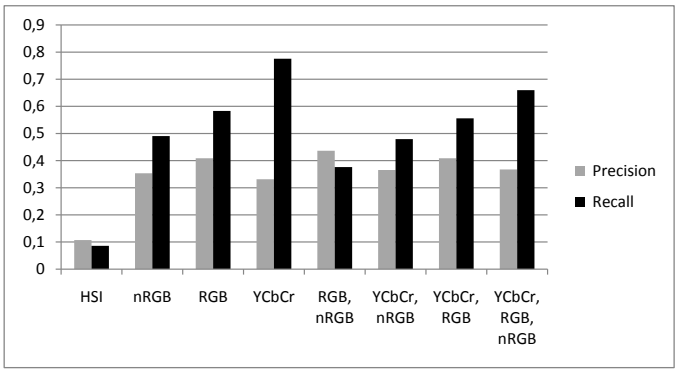

Fig. 7. The average dataset results by using different color spaces and combination of color spaces.

generated. We used Adobe Flash[4] since it allowed us to output a binary ground truth video with a per pixel precision, which was easier to process than using the XML that Viper GT[5] produced. Examples of the annotated ground truth can be seen in Fig. 4. White pixels indicated annotated skin. Non-skin facial features like eyes, eyebrows or similar are left out in case they are visible. Because of the poor quality of the videos this is not always possible.

Figure 5 gives a graph that compares the skin percentage that has been found by the algorithm in one video with the ground truth that has been generated for that video. During the generation of the ground truth it was made sure that eye brows, open mouth and eyes were excluded if these were distinguishable in the video. Sometimes, due to the low resolution and bad visual quality of the video it was not possible to exactly mark some of these non-skin elements (see Figure 4).

C. Model range parameter evaluation

We tested the algorithm against our dataset with various relative model parameters (clamping parameters) for C_b and C_r and got the best results by using 30% for C_b and 17.5% for C_r of the static skin color model. The full set of results is found in Figure 6.

D. Evaluation of color spaces

To get a broad overview of the static approach by using various fixed ranges for color spaces we processed all the videos in the dataset also by using the HSI, $NRGB$, RGB and YC_bC_r color spaces. The static ranges for these color spaces are available in various papers, such as [21], [22]. The results are found in Figure 7 and it shows that the usage of a combination of color spaces results in a more robust detection than only using a single color space.

[4]www.adobe.com

[5]http://viper-toolkit.sourceforge.net/

E. Performance evaluation

After having tested several values for the C_b and C_r components and having done the tests with the static color based approach we wanted to investigate the performance of our approach on a per video basis. We wanted to understand how the algorithm performs with the best performing parameters for C_b (30%) and C_r (17.5%) by looking at the results for each frame of the videos in the dataset. The average classification results over all of the 25 videos compared with the ground truth is displayed in Figure 8.

We compared the skin color detection results that are generated by the algorithm with the one that have been returned by the humans who were classifying skin without context. Figure 9 shows the results of the comparison between the algorithm and humans. The figure also contains the percentage of ground truth that is found in the image. The figure shows that the images where faces were detected during the classification the results are nearer to the ground truth compared with the classification by humans (e.g. in Video 25 (compare Fig. 3), there are neither skin nor faces present). Further it shows that the skin-color restriction usually results in the detection of less than the present skin. For some images, e.g. “Image 6” and “Image 8”, the algorithm is closer to the ground truth than classification by humans despite faces being present in

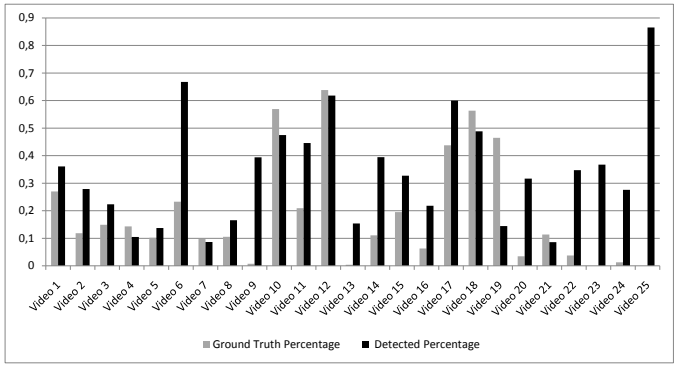

Fig. 8. The average probability for each video in the dataset. The gray bar is the average of skin percentage specified by the ground truth. The black bar represents skin percentage that was detected by the algorithm.

the images. The static ranges restrict the skin-color to fit the ground truth closer than the classification done by humans. The algorithm outperformed the humans in 7 of the 27 images.

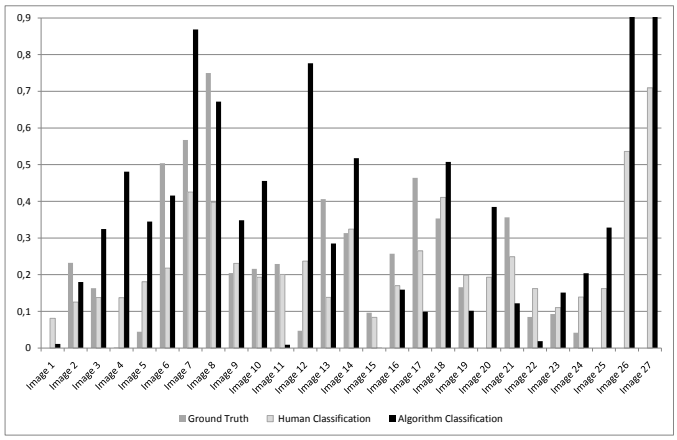

Fig. 9. The average probability for each frame. The light gray bar is the average of skin percentage identified by the humans. The gray bar represents the ground truth average skin percentage for the image. The black bar represents the percentage of skin identified by the algorithm.

V. Conclusion

The results of the online poll show that even humans are not able to reliably classify skin-color when there is no context involved. People tend to randomly guess upon the color whether or not something is skin. We argue that this also holds for low level detection approaches, which are not able to overcome this issue and that we have to have contextual information for reliably detecting skin in visual data. We try to model this contextual information with adapting the skin color classification based on detected and tracked faces in the scene. By using a combination of color spaces and an adaptive multiple model approach to dynamically adapt skin-color decision rules we are able to significantly reduce the number of false positive detection and the classification results become more reliable. The runtime of the algorithm is below real-time and can be carried out in parallel because the frames are independent of each other. No prior training is needed, and no parameters have to be adapted for unknown data. Therefore our approach is applicable for large scale deployment in online web portals to reduce the number of inappropriate user generated videos.

Acknowledgment

This work was partly supported by the Austrian Research Promotion Agency (FFG), project OMOR 815994, and the CogVis[6] Ltd. However, this paper reflects 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] J.-S. Lee, Y.-M. Kuo, P.-C. Chung, and E.-L. Chen, "Naked image detection based on adaptive and extensible skin color model," *PR*, vol. 40, no. 8, pp. 2261–2270, 2007.

[2] C. Garcia and G. Tziritas, "Face detection using quantized skin color regions merging and wavelet packet analysis," *TM*, vol. 1, no. 3, pp. 264–277, 1999.

[3] J. Stöttinger, J. Banova, T. Pönitz, N. Sebe, and A. Hanbury, "Translating journalists' requirements into features for image search," in *VSMM*, 2009, p. to appear.

[4] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *CVPR*, vol. 1, pp. 511–518, 2001.

[5] M. Wimmer, B. Radig, and M. Beetz, "A person and context specific approach for skin color classification," in *ICPR*, 2006, pp. 39–42.

[6] R. Khan, J. Stöttinger, and M. Kampel, "An adaptive multiple model approach for fast content-based skin detection in on-line videos," in *MM AREA*, 2008.

[7] M. hsuan Yang and N. Ahuja, "Gaussian mixture model for human skin color and its application in image and video databases," in *SPIE*, 1999, pp. 458–466.

[8] M. J. Jones and J. M. Rehg, "Statistical color models with application to skin detection." *IJCV*, vol. 46, no. 1, pp. 81–96, 2002.

[9] S. L. Phung, A. Bouzerdoum, and D. Chai, "Skin segmentation using color pixel classification: Analysis and comparison," *PAMI*, vol. 27, no. 1, pp. 148–154, 2005.

[10] P. Kakumanu, S. Makrogiannis, and N. Bourbakis, "A survey of skin-color modeling and detection methods," *PR*, vol. 40, no. 3, pp. 1106–1122, 2007.

[11] N. Sebe, I. Cohen, T. S. Huang, and T. Gevers, "Skin detection: A bayesian network approach," in *ICPR*, 2004, pp. 903–906.

[12] Z. Fu, J. Yang, W. Hu, and T. Tan, "Mixture clustering using multidimensional histograms for skin detection," in *ICPR*, 2004, pp. 549–552.

[13] A. Senior, R.-L. Hsu, M. A. Mottaleb, and A. K. Jain, "Face detection in color images," *PAMI*, vol. 24, no. 5, pp. 696–706, 2002.

[14] V. Vezhnevets, V. Sazonov, and A. Andreev, "A survey on pixel-based skin color detection techniques," in *CCGV*, 2003, pp. 85–92.

[15] L. Sigal, S. Sclaroff, and V. Athitsos, "Skin color-based video segmentation under time-varying illumination," *PAMI*, vol. 26, no. 7, pp. 862–877, 2004.

[16] J.-Y. Lee and Y. Suk-in, "An elliptical boundary model for skin color detection," in *ICIS*, 2002, pp. 579–584.

[17] D. Brown, I. Craw, and J. Lewthwaite, "A som based approach to skin detection with application in real time systems," in *BMVC*, 2001, pp. 491–500.

[18] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," in *COLT*, 1995, pp. 23–37.

[19] K. Tieu and P. Viola, "Boosting image retrieval," in *IJCV*, 2000, pp. 228–235.

[20] E. Osuna, R. Freund, and F. Girosit, "Training support vector machines: an application to face detection," *CVPR*, pp. 130–136, 1997.

[21] D. Chai and K. Ngan, "Locating facial region of a head-and-shoulders color image," *AFGR*, pp. 124–129, 1998.

[22] F. Gasparini and R. Schettini, "Skin segmentation using multiple thresholding," in *SPIE*, vol. 6061, no. 1. SPIE, 2006.

[6] http://www.cogvis.at/