FINDING FACES IN NEWS VIDEOS

Demir Gökalp and Selim Aksoy

Bilkent University
Department of Computer Engineering
Bilkent, 06800, Ankara, Turkey
Email: {dgokalp,saksoy}@cs.bilkent.edu.tr

ABSTRACT

People and their activities are important features in the content description of a multimedia document. The first step in people detection is the detection of their faces. We describe an algorithm for face detection using color and shape information. The proposed approach starts with skin segmentation based on color. Distributions of color features are modeled using mixtures of Gaussians and pixels are labeled as skin or non-skin in a Bayesian framework. After the pixels are labeled, sequential opening and closing operations are applied for noise cleaning. Resulting shapes are modeled according to their first and second order moments. Template matching is applied to the regions with elliptical shapes where the scale and orientation of the templates are estimated using the properties of fitted ellipses. Final decisions are made based on correlations of regions with the templates. We illustrate this approach on a large news video archive that proved to be a challenging data set because of its uncontrolled nature and complexity with respect to illumination, pose, location, scale, and occlusion. In comparative experiments with a popular face detector, our approach resulted in higher computational efficiency and comparable detection rates with fewer false alarms.

1. INTRODUCTION

Face detection [18] and recognition are important computer vision problems that have received significant attention in many applications such as multimedia indexing, biometric identification, human computer interaction, journalism, surveillance, people tracking, and film and video archiving. Example face detection techniques include knowledge-based methods that describe the relationships between facial features; feature-based methods that use edge, color and texture information; template matching methods that search for matches to a standard face pattern; and appearance-based methods that apply machine learning techniques such as eigenfaces, distribution-based models, neural networks, and support vector machines to find relevant characteristics of face and non-face images.

Recent success in face recognition in benchmark data sets such as the FERET database, or in face detection in data sets such as the one created by Rowley *et al.* [11], motivated algorithms for using face information for automatic multimedia indexing because people and their activities are important features in the content description of the visual modality of a multimedia document. A natural initial step in people detection is face detection. Most approaches even simplify the problem of people detection to detection of a human face [15]. Once a face is detected, face recognition techniques such as eigenfaces and Fisherfaces can be used to assign names to the detected face. Face detection can be combined with additional models for different body parts to verify the existence of a person. Speaker recognition techniques that identify people based on their speech utterance, and natural language processing techniques that label words as names in transcripts can also be used to improve person detection and recognition [15, 13].

In this paper, we concentrate on the problem of detecting faces in news videos. With the availability of public archives on the Internet and proprietary archives at news stations, news videos became an important source of information due to their rich content and high social impact. Designing systems for content-based indexing, analysis and retrieval of these archives have become a challenging research area. Therefore, the U.S. National Institute of Standards and Technology (NIST) selected news videos as the data set that has been used in the TRECVID workshop series for video retrieval evaluation [8].

Even though these videos include shots of anchorpersons in studio environments, a significant portion of the content contains recordings taken in both outdoor and indoor scenes that are completely uncontrolled with respect to factors such as illumination, pose, location, scale, occlusion, sound quality, etc. These environments introduce noise and unlimited variability, and provide a challenge to many systems designed for unimodal processing in specific data sets

This work was supported by the TUBITAK Grant 104E077 and the European Commission COST 292 Action.

with specific assumptions.

The constraints that we have considered in tackling this problem are as follows. First of all, the algorithm has to be fast because the size of the news archive is constantly increasing. The algorithm also cannot depend on uniform segmentation of image regions [5] or detailed templates and models based on symmetry assumptions [12] due to the uncontrolled nature of news videos. However, we can still make shape assumptions because faces appear elliptical even in different poses.

The proposed approach for face detection starts with skin segmentation. We evaluate the performances of different color models in labeling pixels as skin or non-skin. Distributions of color features are modeled using mixtures of Gaussians and the final decision about a pixel being part of a skin or a non-skin region is made in the Bayesian framework. After the pixels are labeled, sequential opening and closing operations are applied for noise cleaning. Resulting connected components are modeled according to their first and second order moments. Template matching is applied to the regions with elliptical shapes where the scale and orientation of the templates are estimated using the properties of the fitted ellipses. Finally, regions that have high correlations with the templates are labeled as being faces.

The rest of the paper is organized as follows. Segmentation of images into regions containing skin is described in Section 2. Locating face candidates based on color and shape information is discussed in Section 3. Template matching-based face detection is described in Section 4. Experiments are presented in Section 5 and suggestions for future work are given in Section 6.

2. SKIN SEGMENTATION

2.1. Selecting features

Our motivation for using skin color for initial modeling of faces comes from the fact that color of human skin is distinct from colors of other objects, and skin color does not depend on the pose of a face. Establishing a color model is a very important part of the color-based face detection to separate face regions from background. Choosing a suitable color space depends on the reliability on different lighting and environmental conditions. Many different color spaces have been used to extract skin regions in color images [18, 9]. Examples include:

- *RGB*: This is usually the original color space where colors are specified in terms of the three primary colors red (R), green (G) and blue (B) [4].

- *Normalized chromaticity*: RGB values are normalized by intensity so that the new values sum to one [1, 5]. Only two of these values are used for modeling skin because the third one is known when the other two are known.

- *HSV*: Colors are specified in terms of hue (H), saturation (S) and intensity value (V). Usually H and S are used for modeling skin [16].

- *CIE-LUV*: RGB values can be converted to the LUV space using a nonlinear transformation and the luminance (L) is usually discarded for modeling skin [17].

- *YES*: RGB values can be converted to the YES space using a linear transformation where Y represents the luminance and E and S denote the chrominance components [12]. Luminance is discarded and chrominance values are used for modeling skin.

- *Log-opponent*: RGB values can be converted to log-opponent values I, R_g, B_y using logarithmic transformations [2].

The RGB color space is generally considered as not being a suitable model at different lighting conditions. Color spaces where intensity information can be ignored are preferable because differences in the skin color of people from different ethnic groups lie largely between the intensity values rather than the chrominance components [18].

We investigated the effects of different color space transformations on skin detection in the TRECVID news video archive. Color constancy is the basic problem for skin detection. Brightness, shadows and makeup can affect appearance of skin partly or completely in these videos. Since the videos in our data set are very noisy and have low quality, three different color spaces were examined to reduce lighting effects. Reliability of color-based face detection strongly depends on the overlap between skin and non-skin color information so an initial evaluation was done according to the overlap of skin and non-skin values in histograms.

Given the original RGB data, we used the normalized chromaticity values that are computed as

$$C_r = \frac{R}{R+G+B} \tag{1}$$

$$C_b = \frac{B}{R+G+B}, \tag{2}$$

the E and S components of the YES color space that are computed as

$$\begin{bmatrix} Y \\ E \\ S \end{bmatrix} = \begin{bmatrix} 0.250 & 0.684 & 0.063 \\ 0.500 & -0.500 & 0.000 \\ 0.250 & 0.250 & -0.500 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}, \tag{3}$$

and the R_g and B_y components after transformation to the log-opponent values that are computed as

$$R_g = L(R) - L(G) \tag{4}$$

$$B_y = L(B) - (L(G) + L(R))/2 \tag{5}$$

where $L(x) = 105 \log_{10}(x + 1 + n)$ and n is a uniform noise value generated from a distribution uniform over the range $[0, 1)$. The histograms of skin and non-skin pixels with respect to individual color components are shown in Fig. 2 and the details are discussed in Section 5.2.

2.2. Detecting skin

Both non-parametric and parametric methods are used for modeling skin color in the literature. Non-parametric techniques include piecewise linear decision boundaries such as thresholds on color features, and histogram intersection techniques that compare a control histogram computed from a skin patch given by the user and a test histogram computed from a new region in an image. Parametric techniques include modeling of skin color distributions using unimodal Gaussians [5, 12] or mixtures of Gaussians [17]. Nonlinear models such as multilayer perceptrons were also used [9]. Several studies compared effectiveness of different models such as piecewise linear decision boundaries, Bayesian classifiers that use histogram-based density estimation or estimation using Gaussian models, and multilayer perceptrons, and concluded that histogram-based estimation performs better than Gaussian models [9, 4]. However, non-parametric techniques such as histogram-based models or nonlinear techniques such as multilayer perceptrons usually require a large amount of training data and computations can be quite demanding when the data size increases, so they may not be practical in many cases (e.g., the histogram models in [9] that performed better than Gaussian models were constructed with 256^3 bins using 680.7 million training samples).

As can be seen from Fig. 2, the distributions of neither skin nor non-skin pixels are unimodal. To model these distributions, we use Gaussian mixture models because closed form estimates of their parameters can be computed using the Expectation-Maximization (EM) algorithm. The probability density function of a mixture model with k components for the feature vector $\mathbf{x}$ is defined as

$$p(\mathbf{x}) = \sum_{j=1}^{k} \alpha_j p(\mathbf{x}|j) \tag{6}$$

where α_j is the mixture weight and $p(\mathbf{x}|j)$ is the Gaussian density model for the j'th component

$$p(\mathbf{x}|j) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma_j}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu_j})^T \boldsymbol{\Sigma_j}^{-1}(\mathbf{x}-\boldsymbol{\mu_j})} \tag{7}$$

where $\boldsymbol{\mu_j}$ is the mean vector and $\boldsymbol{\Sigma_j}$ is the covariance matrix for the j'th component, respectively, and d is the dimension of the feature space, $\mathbf{x} \in \mathbb{R}^d$.

We use the k-means algorithm to determine the initial configuration. Then, the EM algorithm is run on the training data using a stopping criterion that checks if the change in log-likelihood between two iterations is less than a threshold. Given the training data $\mathcal{D} = \{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$, the parameters of the mixture can be estimated as

$$p(j|\mathbf{x_i}) = \frac{\alpha_j p(\mathbf{x_i}|j)}{\sum_{t=1}^{k} \alpha_t p(\mathbf{x_i}|t)} \tag{8}$$

$$\hat{\alpha}_j = \frac{\sum_{i=1}^{n} p(j|\mathbf{x_i})}{n} \tag{9}$$

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{j}} = \frac{\sum_{i=1}^{n} p(j|\mathbf{x_i})\mathbf{x_i}}{\sum_{i=1}^{n} p(j|\mathbf{x_i})} \tag{10}$$

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{j}} = \frac{\sum_{i=1}^{n} p(j|\mathbf{x_i})(\mathbf{x_i} - \hat{\boldsymbol{\mu}}_{\boldsymbol{j}})(\mathbf{x_i} - \hat{\boldsymbol{\mu}}_{\boldsymbol{j}})^T}{\sum_{i=1}^{n} p(j|\mathbf{x_i})}. \tag{11}$$

The number of components in the mixture can be either supplied by the user or chosen using some optimization criteria. We use the Minimum Description Length (MDL) Principle [10] that tries to find a compromise between the model complexity (still having a good data approximation) and the complexity of the data approximation (while using a simple model). Under MDL, the best model M is the one that minimizes the sum of the model's complexity ($\frac{\kappa_M}{2} \log n$, where κ_M is the number of free parameters in model M) and the efficiency of the description of the training data with respect to that model ($-\log p(\mathcal{D}|M)$). For a Gaussian mixture model with k components, the number of free parameters becomes $\kappa_M = (k-1) + kd + k\frac{d(d+1)}{2}$ and the best k can be found as

$$k^* = \arg\min_k \left[\frac{\kappa_M}{2} \log n - \sum_{i=1}^{n} \log \left(\sum_{j=1}^{k} \alpha_j p(\mathbf{x_i}|j) \right) \right]. \tag{12}$$

After the densities for both skin pixels and non-skin pixels are estimated as in (6) using their corresponding training data, an unknown pixel with feature vector $\mathbf{x}$ is labeled as skin using the Bayesian classifier

$$\text{label } \mathbf{x} \text{ as skin if } p(\mathbf{x}|\text{skin}) > p(\mathbf{x}|\text{non-skin}) \tag{13}$$

under the minimum-error with equal priors assumption.

3. LOCATING FACE CANDIDATES

3.1. Post-processing skin detector output

After labeling of pixels as skin or non-skin, the next step is to locate the face candidates. However, because of the enormous variations in the lighting conditions such as shining or shadow effects, there are some holes in the face regions. Similarly, because of the complexity of the backgrounds in the scenes in news videos, there are often a lot of noise pixels that are mistakenly labeled as belonging to skin. These errors must be corrected to improve the accuracy of face

detection. Eliminating noise pixels also improves the efficiency of the rest of the algorithm.

We adopted mathematical morphology operators [3] for post-processing the skin detector output. We apply sequential opening and closing operations using 5×5 disk structuring elements. Opening eliminates small bumps that are connected to skin regions due to clothing or lighting effects, and closing recovers the holes within skin regions.

Finally, the connected components labeling algorithm [3] is applied to the resulting binary images to label candidate face regions. Very small connected components are also eliminated using an area threshold.

3.2. Estimating face scale and orientation

Our goal is to be able to detect all faces at all scales and orientations. We assume that faces have elliptical shapes [16, 12] and model the face regions using their first and second order moments.

Given the list of pixels in a region connected component as $\{(r_1, c_1), \ldots, (r_n, c_n)\}$ where r and c represent row and column coordinates, respectively, the first step in fitting an ellipse to this region is to find its centroid (μ_r, μ_c) as

$$\mu_r = \frac{1}{n}\sum_{i=1}^{n} r_i, \tag{14}$$

$$\mu_c = \frac{1}{n}\sum_{i=1}^{n} c_i. \tag{15}$$

Then, the pixel coordinates can be normalized with respect to translation around the centroid, and ellipse properties such as the major axis length a_{major} and minor axis length a_{minor} can be computed using the central moments as [3]

$$a_{\text{major}} = 2\sqrt{2}\sqrt{u_{cc} + u_{rr} + \sqrt{(u_{cc} - u_{rr})^2 + 4u_{rc}^2}} \tag{16}$$

$$a_{\text{minor}} = 2\sqrt{2}\sqrt{u_{cc} + u_{rr} - \sqrt{(u_{cc} - u_{rr})^2 + 4u_{rc}^2}} \tag{17}$$

where $u_{rr} = \frac{1}{n}\sum_{i=1}^{n}(r_i - \mu_r)^2$, $u_{cc} = \frac{1}{n}\sum_{i=1}^{n}(c_i - \mu_c)^2$, and $u_{rc} = \frac{1}{n}\sum_{i=1}^{n}(r_i - \mu_r)(c_i - \mu_c)$.

Similarly, orientation θ of the ellipse can be computed as

$$\theta = \begin{cases} \arctan\left(\frac{u_{rr} - u_{cc} + \sqrt{(u_{rr} - u_{cc})^2 + 4u_{rc}^2}}{-2u_{rc}}\right) & \text{if } u_{rr} > u_{cc} \\ \arctan\left(\frac{-2u_{rc}}{u_{cc} - u_{rr} + \sqrt{(u_{cc} - u_{rr})^2 + 4u_{rc}^2}}\right) & \text{if } u_{rr} \leq u_{cc} \end{cases} \tag{18}$$

and its eccentricity e can be computed as

$$e = \sqrt{1 - \left(\frac{a_{\text{minor}}}{a_{\text{major}}}\right)^2}. \tag{19}$$

Before computing these properties, we iteratively fill the holes inside connected components to improve localization.

Fig. 1. Face templates are the centroids of the frontal face clusters found using the k-means algorithm with $k = 5$. The rows show the RGB and E color components, respectively.

To further eliminate noise regions, we select only the ones that have an eccentricity less than 0.89. This value is chosen so that the regions that are fitted ellipses with a major and minor axis ratio greater than approximately 2.2 (i.e., too long to be a face) are eliminated. Regions that pass this final test are kept as face candidates for template matching-based face detection as described in the next section.

4. FACE DETECTION

We use template matching for making the final decision about a candidate region being a face or not. The templates are generated by clustering public domain frontal face examples (http://pics.psych.stir.ac.uk/) using the k-means algorithm. We set $k = 5$ and use the cluster centroids shown in Fig. 1 as the templates.

However, instead of resizing the whole image (or template) at different scales, rotating each scaled image (or template) to different orientations, and searching for matches to the template at all pixel locations, we use the fitted ellipse to estimate the scale and orientation of a face candidate. Then, we apply an affine transformation to each template and verify its existence only at the candidate locations. This significantly reduces the computational load caused by other approaches that test for all possible scales, orientations and translations. The final labels are assigned by comparing the correlation of the templates with the candidate regions to a threshold. The largest correlation among the values for all templates is used as the final measure.

Face detection is done very fast due to the sequential elimination of regions in multiple levels. However, the approach is biased towards frontal faces because of the templates used. Nevertheless, we have observed that correlations between faces at different poses and the templates scaled and rotated according to the shape properties of the candidate regions are still higher than correlations with non-face regions that happen to pass all the checks before template matching.

5. EXPERIMENTS

5.1. Experimental setup

The training and testing data come from the development set for the TREC Video Retrieval Evaluation 2004 (TRECVID) organized by the U.S. National Institute of Standards and Technology. The development set includes approximately 120 hours (241 30-minute programs) of ABC World News Tonight and CNN Headline News recorded between January through June 1998. 62.2 hours (51.6GB) of this data were annotated by the Video Collaborative Annotation Forum [6] that was composed of researchers from 21 institutions. Annotations were assigned from a lexicon with 133 items.

We used the annotation data as the ground truth for evaluating the face detection algorithm. From the annotation lexicon, first, we selected the items that were related to a person: Bill Clinton, Face, Female Face, Female News Person, Female News Subject, Human, Madeleine Albright, Male Face, Male News Person, Male News Subject, Newt Gringrich, People, People Event, Person, Person Action, Peter Jennings, Saddam Hussein, Running, Walking. Then, all video shots that contained at least one of these items in their corresponding annotations were marked as containing a face. There is a possibility that this may result in mislabeled truth data because some of the actions in the lexicon (such as running, walking, person action) may exist without a visible face, and the annotation is also known to contain some errors. However, we treat the resulting 20,897 shots out of the total 33,549 as the ground truth data for images containing faces supplied by TRECVID without further labeling to be as objective as possible with the evaluation.

To estimate the performance of the algorithm, we define the following detection problem:

- w_1 denotes the case when a shot is annotated as containing a face.
- w_0 denotes the case when a shot is not annotated as containing a face.
- α_1 is the action of deciding that there is a face in the shot.
- α_0 is the action of deciding that there is no face in the shot.

Therefore, w_1 and w_0 come from the TRECVID annotation, α_1 and α_0 are related to the face detector output. Then,

- $P(\alpha_1|w_1)$ is the correct detection rate.
- $P(\alpha_0|w_1)$ is the misdetection rate.
- $P(\alpha_1|w_0)$ is the false alarm rate.
- $P(\alpha_0|w_0)$ is the correct rejection rate.

Tab. 1. Training and testing data used in evaluating the skin detector.

Datasets	# images	# skin pixels	# non-skin pixels
Training	144	15,379	83,353
Testing	60	13,345	28,669

We manually check the output of the face detector for each shot, verify that the regions marked as faces are indeed faces, and compute these rates using the annotation labels.

5.2. Evaluation of skin detection

The first step in performance evaluation is skin detection. We manually collected image patches from skin and non-skin regions from TRECVID videos. The number of pixels in the training and testing examples are shown in Tab. 1. Test data includes skin and non-skin patches extracted independently from keyframes that are not in the training set.

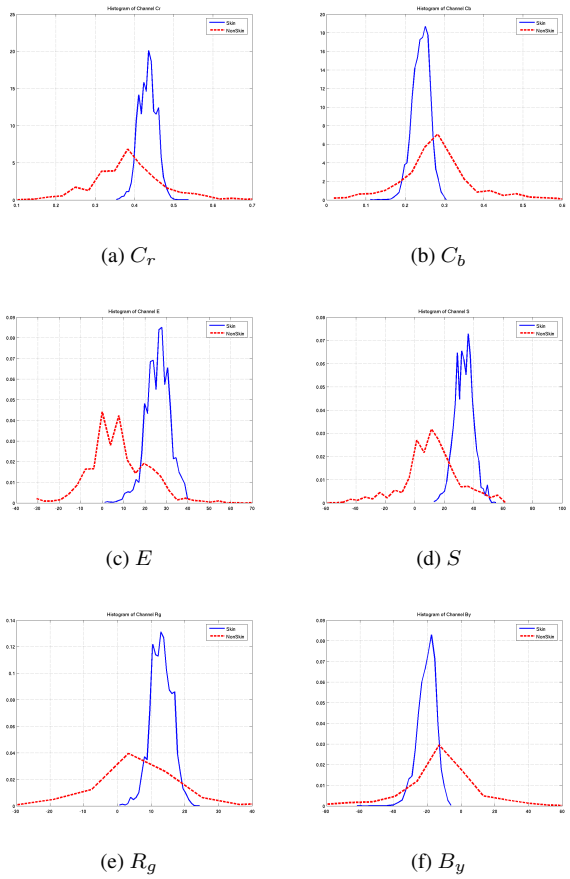
Histograms for both skin and non-skin pixels in the training set are shown in Fig. 2. Due to the variations in recording conditions, complexity of scene backgrounds, and the uncontrolled nature of the appearances of faces in the videos, most of the histograms for skin and non-skin have a significant amount of overlap. This overlap was found to be the smallest in the E and S components of the YES color space for our data.

In addition to comparing the histograms for selecting color components, we trained Bayesian classifiers using univariate Gaussians and their mixtures as described in Section 2.2. The number of components in mixture densities were estimated using the Minimum Description Length formulation. For example, the optimum number of mixture components for the E color component was found to be 4. Plots for fitted densities are shown in Fig. 3. Confusion matrices for Bayesian classification are given in Tab. 2. As a result, we selected to use only the E component of the YES color space in the rest of the experiments because of the small overlap in skin and non-skin histograms, and the smallest classification error obtained using the Bayesian classifier.

5.3. Evaluation of face detection

After selecting the color feature and estimating the densities for skin and non-skin classes, we applied the algorithm that involves Bayesian classification, post-processing using morphology, scale and orientation estimation via ellipse fitting, and template matching steps to all of the 33,549 annotated shots in the news video archive. We manually examined all results and computed the detection and error rates using the setting described in Section 5.1.

We performed three separate experiments for face detection that differ in the number of templates and the correlation thresholds used. In addition, the area threshold for

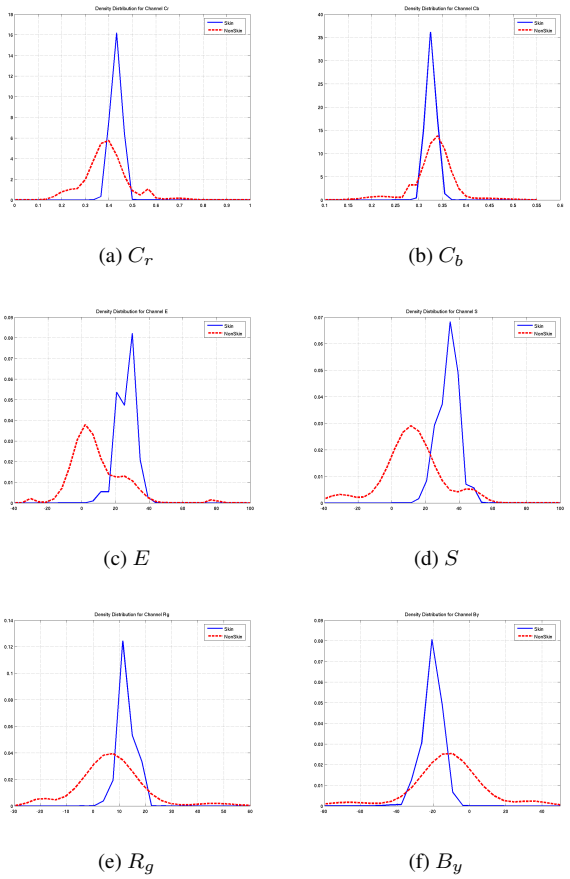


(a) C_r (b) C_b

(c) E (d) S

(e) R_g (f) B_y

Fig. 2. 1-D histograms of skin vs. non-skin pixels with respect to individual color components. Solid blue lines represent skin histograms and dashed red lines represent non-skin histograms.



(a) C_r (b) C_b

(c) E (d) S

(e) R_g (f) B_y

Fig. 3. Fitted distributions for skin vs. non-skin pixels using the models described in Section 2.2. Solid blue lines represent skin densities and dashed red lines represent non-skin densities. Only the univariate versions computed for individual color components are shown here.

eliminating connected components that were too small to be face candidates was set to 1000 pixels in both experiments. Confusion matrices for face detection results are given in Tab. 3. For a single template with a correlation threshold of 0.3, correct detection rate was 22.63% with a false alarm rate of 11.78%. For the same single template with a correlation threshold of 0.6, correct detection rate was 15.17% with a false alarm rate of 5.08%. For five templates with a correlation threshold of 0.6, correct detection rate was 23.19% with a false alarm rate of 9.90%.

The proposed approach produces conservative detections. The correct detection rates are not very high but the false alarm rates are relatively lower. This means that when the algorithm labels a region as being a face, it is usually correct. Correct detection rate can be increased by adjusting the prior probabilities in the Bayesian skin classifier and reducing the correlation threshold in template matching while still keeping the false alarm rate relatively low.

We also performed comparative experiments using a popular face detector, an implementation of the Schneiderman-Kanade's face detection algorithm [14] by Mikolajcyzk [7]. This detector was run on the same shots with annotations and the results were manually inspected to compute detection rates. This algorithm resulted in a correct detection rate of 32.02% but the false alarm rate was 44.17%. This detector gave a higher correct detection rate at the expense of significantly increased number of false alarms. Note that the correct detection rate of this detector on this data set is significantly lower than the rates achieved on faces in photographs as reported in the literature.

Processing speeds of both algorithms were also different. Our approach processed the whole data set within 12 hours using an unoptimized Matlab implementation. The other face detector took multiple days to process the same data set using an implementation in C.

Example shots and regions where faces were correctly detected or where non-face regions were mistakenly labeled

Tab. 3. Confusion matrices for face detection on 33,549 shots.

		Detected	
		face (α_1)	no face (α_0)
True	face (w_1)	4,729	16,168
	no face (w_0)	1,490	11,162

(a) Using a single template with a correlation threshold of 0.3. Correct detection rate is 22.63% with a false alarm rate of 11.78%.

		Detected	
		face (α_1)	no face (α_0)
True	face (w_1)	3,170	17,727
	no face (w_0)	643	12,009

(b) Using a single template with a correlation threshold of 0.6. Correct detection rate is 15.17% with a false alarm rate of 5.08%.

		Detected	
		face (α_1)	no face (α_0)
True	face (w_1)	4,847	16,050
	no face (w_0)	1,253	11,399

(c) Using five templates with a correlation threshold of 0.6. Correct detection rate is 23.19% with a false alarm rate of 9.90%.

Tab. 2. Confusion matrices for skin vs. non-skin classification. Components in different color spaces are used both individually and together to learn density models as described in Section 2.2. The E component of the YES color space resulted in the smallest classification error of 5.58%.

		Assigned with C_r		Assigned with C_b		Assigned with C_r and C_b	
		skin	non-skin	skin	non-skin	skin	non-skin
True	skin	11,966	3,937	10,694	5,988	11,230	3,702
	non-skin	1,758	24,383	3,030	22,332	2,494	24,618

(a) Classification using the normalized chromaticity color space.

		Assigned with E		Assigned with S		Assigned with E and S	
		skin	non-skin	skin	non-skin	skin	non-skin
True	skin	12,361	984	10,963	5,416	10,719	1,233
	non-skin	1,363	27,336	2,761	22,904	3,005	27,087

(b) Classification using the YES color space.

		Assigned with R_g		Assigned with B_y		Assigned with R_g and B_y	
		skin	non-skin	skin	non-skin	skin	non-skin
True	skin	11,873	4,265	10,355	6,164	12,004	9,191
	non-skin	1,851	24,055	3,369	22,156	1,720	19,129

(c) Classification using the log-opponent color space.

as being faces are shown in Fig. 4. The performances of both approaches show that finding faces in these news videos is indeed a very hard problem.

6. DISCUSSION AND FUTURE WORK

We described a face detection algorithm that consists of color-based Bayesian skin segmentation and template matching using correlations of elliptical skin regions with face templates. Experiments performed on the news videos and manual annotation provided by TREC video retrieval evaluation showed that the approach is capable of detecting faces in uncontrolled scenes with complex backgrounds while still keeping the false alarm rate low. The intermediate steps of the algorithm were designed so that it produces conservative detections. In other words, when the algorithm labels a region as being a face, it is usually correct.

We are investigating several ways for improving the detection results. Skin segmentation will be extended by developing separate models for people from different ethnic groups. Template matching will be enhanced by introducing multiple templates with different poses. A novel extension with the highest potential for improvement is multimodal detection by combining visual, audio and textual information, and its application to content-based retrieval by combining visual information with text (e.g., speech transcripts) and audio (e.g., human voice, speaker ID).

7. REFERENCES

[1] T. S. Caetano and D. A. C. Barone. A probabilistic model for the human skin color. In *International Conference on Image Analysis and Processing*, pages 279–283, 2001.

[2] D. A. Forsyth and M. M. Fleck. Automatic detection of human nudes. *International Journal of Computer Vision*, 32(1):63–77, August 1999.

[3] R. M. Haralick and L. G. Shapiro. *Computer and Robot Vision*. Addison-Wesley, 1992.

[4] M. J. Jones and J. M. Rehg. Statistical color models with application to skin detection. *International Journal of Computer Vision*, 46(1):81–96, January 2002.

[5] S. H. Kim, N. K. Kim, S. C. Ahn, and H. G. Kim. Object oriented face detection using range and color information. In *International Conference on Automatic Face and Gesture Recognition*, pages 76–81, 1998.

[6] C.-Y. Lin, B. L. Tseng, and J. R. Smith. Video collaborative annotation forum: Establishing ground-truth labels on large multimedia datasets. In *NIST TREC-2003 Video Retrieval Evaluation Conference*, Gaithersburg, MD, November 2003.

[7] K. Mikolajczyk. Face detector. Technical report, INRIA Rhone-Alpes.

[8] U.S. National Institute of Standards and Technology. TREC video retrieval evaluation (TRECVID), 2004. http://www-nlpir.nist.gov/projects/trecvid/.

[9] S. L. Phung, A. Bouzerdoum, and D. Chai. Skin segmentation using color pixel classification: Analysis and comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(1):148–154, January 2005.

[10] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

[11] H. A. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):23–38, January 1998.

(a) Examples of faces correctly detected by our algorithm.

(b) Example regions incorrectly labeled as containing faces by our algorithm.

(c) Example regions incorrectly labeled as containing faces by the face detector implementation by Mikolajcyzk.

Fig. 4. Face detection examples.

[12] E. Saber and A. M. Tekalp. Frontal-view face detection and facial feature extraction using color, shape and symmetry based cost functions. *Pattern Recognition Letters*, 17(1):669–680, 1998.

[13] S. Satoh, Y. Nakamura, and T. Kanade. Name-it: Naming and detecting faces in news videos. *IEEE Multimedia*, 6(1):22–35, January-March 1999.

[14] H. Schneiderman and T. Kanade. Object detection using the statistics of parts. *International Journal of Computer Vision*, 56(3):151–177, February 2004.

[15] C. G. M. Snoek and M. Worring. Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications*, 25(1):5–35, January 2005.

[16] K. Sobottka and I. Pitas. Face localization and feature extraction based on shape and color information. In *IEEE International Conference on Image Processing*, pages 483–486, 1996.

[17] M.-H. Yang and N. Ahuja. Gaussian mixture model for human skin color and its application in image and video databases. In *SPIE Storage and Retrieval of Image and Video Databases VII*, pages 458–466, 1999.

[18] M.-H. Yang, D. J. Kriegman, and N. Ahuja. Detecting faces in images: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(1):24–58, January 2002.