INFORMATION THEORETICAL BASED FEATURE SELECTION APPROACH FOR HUMAN SKIN DETECTION

K. S. Chenaoua

Computer Engineering Department
KFUPM,
Dhahran 31261
Saudi Arabia

A. Bouridane

School of Computer Science
Queens University of Belfast
Belfast BT7 1NN
UK

ABSTRACT

Detection of human skin in colored images has always been performed in known standard color spaces. In this paper a new color space coordinate is proposed based on popular existing color spaces but taking into account the most representative ones. Selection of the best color components is based on the use of the mutual information and maximum relevance minimum redundancy technique. A Gaussian model based classifier is used to test the performance of the proposed color space transformation.

Index Terms— Skin detection, feature selection, mutual information, Gaussian model

1. INTRODUCTION

Detection of human skin in color images has been tackled in many different works in the last decade [1]. Skin segmentation is commonly used in algorithms for face detection, hand gesture analysis [2], and objectionable image filtering [1]. In these applications, the search space for objects of interest, such as faces or hands, is reduced through the detection of skin regions.

Different existing color spaces, combination of color spaces, or sometimes authors’ own proposed color space transformations were proposed [1, 3]. Choice of a given color space is driven by its ability to cluster skin pixels and separate between skin and non-skin pixels. Even though comparative studies have been carried out to determine a suitable color space for skin detection [4].

Color is the main attribute used to detect skin regions in colored images. Assuming a color component as a feature, a good number of combinations of such features would be possible and hence to use all or part of the features for skin detection would be prohibitive in terms of computation requirements.

In this paper, an information theoretical based approach for the choice of an adequate color basis is proposed. The idea is to maximize the mutual information between the candidate features and the target classes. The mutual information between two input candidate features should be minimum [5]. For this reason, the Maximum-Relevance Minimum-Redundancy (MRMR) algorithm is used to select the best feature subset. The approach is divided in three steps: The first one involves extracting the characteristic attributes from the image data base of skin and non-skin pixels. The second step involves selecting the best feature subset. Finally, a Gaussian model is used to model the distribution of skin and non skin pixels, and hence the Bayes classifier is used to classify the input image pixels into skin and non skin.

In the first part of the paper, a description of the theoretical background is given, then the choice of the candidate features and the selection of the best color basis is explained. In the third part the Gaussian model is exposed, and finally some results are given.

2. THEORETICAL BACKGROUND

In this section the theory required to establish a basis for the used method is described. First the concept of mutual information is explained, then the maximum relevance minimum redundancy is briefly explained.

2.1. Definition of the Mutual Information

A classifying system maps input features onto output classes. However, only part of the input features are relevant and contain important information regarding output. In feature selection problems, techniques tend to find inputs that contain as much information about the output as possible. This information needs adequate tools to be measured. Shannons information theory provides a way to measure the information of random variables with entropy and mutual information [6, 7]. The entropy is a measure of uncertainty of a random variable. If a discrete random variable $X$ has $\chi$ alphabets and the probability density function (pdf) is $p(x) = Pr(x = X), x \in \chi$ the entropy of $X$ is defined as

$$H(X) = - \sum_{x \in \chi} p(x) \log p(x)$$

(1)

Here the base of is 2 and the unit of entropy is the bit. For two discrete random variables and with their joint pdf, the joint entropy of and is defined as:

$$H(X) = - \sum_{x \in \chi} \sum_{y \in Y} p(x, y) \log p(x, y)$$

(2)

When certain variables are known and others are not, the remaining uncertainty is measured by the conditional entropy:

$$H(Y/X) = - \sum_{x \in X} p(x) H(Y/X = x)$$

(3)

$$= - \sum_{x \in X} p(x) \sum_{y \in Y} p(y/x) \log p(y/x)$$

(4)
\[ H(X, Y) = H(X) + H(Y/X) \]  
\[ = H(Y) + H(X/Y) \]

This rule, known as the chain-rule, implies that the total entropy of two random variables \( X \) and \( Y \), is the entropy of \( X \) plus the remaining entropy of \( Y \) for a given \( X \).

The information found commonly in two random variables is known as the mutual information between two variables, and is defined in the discrete case by:

\[
I(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}
\]  

3. MUTUAL INFORMATION BASED FEATURE SELECTION

The process of selecting a subset of features from an original set of available features is called feature selection. This technique is a data reduction technique, which reduces the computational power when used for any pattern recognition problem. Some features can be more representative than others, and features can also be similar to each other, and thus are not providing significantly more information than any of them individually.

After the feature selection process, the selected features should have high interdependence between feature values and classes. However, the interdependence among the selected features should be minimized, so that redundancy is minimized.

3.1. Maximum Dependency

The Maximum Dependency (MD) algorithm consists in finding a subset \( S_m \) with \( m \) features \( x_1, x_2, \ldots, x_m \) which jointly has the Maximal Dependency on the target class \( C \). The measurement of this dependency is based on the estimation of the mutual information between \( S_m \) and the target class \( C \):

\[
I(X, C) = H(C) - H(C/X)
\]

The conditional entropy \( H(C/X) \) (9) measures the degree of uncertainty entailed by the set of classes \( C \) given the set of feature values \( X \). In this paper, only two classes are considered: \( C = \{ \text{skin}, \neg \text{skin} \} \).

\[
H(C/X) = \sum_{c \in C} p(c) \sum_{x \in X} p(x/c) \log \frac{p(x/c)p(c)}{p(x)}
\]

The first term in (8) is the entropy and depends only on the classes and does not depend of the design or selection of features. It provides an upper bound of the mutual information; \( I(X, C) \) since \( 0 < H(C/X) < H(C) \). The second term may be interpreted as the decrease in the uncertainty.

That is, with a higher interdependence between the feature values \( X \) and the classes \( C \), one has a higher certainty (i.e. lower uncertainty) in classifying a pixel given its feature value. The mutual information \( I(X, C) \) is at maximum when \( X \) and \( C \) are totally dependent on each other. Conversely, it is at minimum when there is no relationship between \( X \) and \( C \). Thus, in order to maximize the class separability, one objective is to maximize the mutual information \( I(X, C) \) between the feature values and the classes.

3.2. Minimum Redundancy

To avoid selecting two or more features that are similar to each other, the interdependence between the selected features should be minimized. The mutual information measures such interdependence. Using the PDFs of the different features, the pair-wise mutual information between two features is estimated by:

\[
I(X_i, X_j) = H(X_i) - H(X_i/X_j)
\]

3.3. Bayes Classifier

Both skin and non skin data is modeled by a Gaussian distribution. The color space is quantized into a number of bins, each corresponding to a particular range of color component values. After training histogram counts are normalized, and the values converted to a discrete probability distribution (11):

\[
p(c/skin) = \frac{1}{2\pi|\Sigma_s|^{1/2}} \cdot e^{-\frac{1}{2}(c-\mu_s)^T\Sigma_s^{-1}(c-\mu_s)}
\]

where \( c \) is the pixel value and the model parameters: mean \( \mu_s \) and covariance matrix \( \Sigma_s \), are computed by (12,13):

\[
\mu_s = \frac{1}{n} \sum_{i=1}^{n} c_i
\]

\[
\Sigma_s = \frac{1}{n} \sum_{i=1}^{n} (c_i - \mu_s)(c_i - \mu_s)^T
\]

The probability of observing skin given a color value \( c \) is given by \( P(skin/c) \). To compute this probability, the Bayes rule [8] is used:

\[
P(skin/c) = \frac{P(c/skin)P(skin)}{P(c/skin)P(skin) + P(c/\neg skin)P(\neg skin)}
\]

\[
P(c/skin) \text{ and } P(c/\neg skin)
\text{ are directly computed from skin and non-skin histograms. The prior probabilities } P(\text{skin}) \text{ and } P(\text{\neg skin}) \text{ can also be estimated from the overall number of skin and non-skin samples in the training set [9, 10, 11].}

The following inequality:

\[
\frac{P(c/skin)}{P(c/\neg skin)} \geq \theta
\]

where \( \theta \) is a threshold value, is used as a skin detection rule [9]. The receiver operating characteristics (ROC) curve [12] shows the relationship between correct detections and false detections for a classification rule as a function of the detection threshold.

4. PROPOSED APPROACH

This paper uses the Mutual Information Based Feature Selection (MIFS) algorithm to select \( M \) features from the original set of \( M_0 \) features. The MIFS algorithm takes into account both of the interdependencies described in Section 3.1 and Section 3.2. Outlines of the MIFS algorithm are described in the following sections.

The proposed technique gives an ordered list of features. Ordering is based on the MRRR principle. A problem arises as to what is the number of features among the 27 ordered ones should be taken into consideration.
Algorithm 1: Feature Selection

1: From the feature set, find the first feature that maximizes the mutual information between feature values and classes. (8).
2: Select the next feature based on the highest:
\[
I(X, C) - \beta \cdot \sum_{s \in \text{Selected}} (X, X_s)
\] (16)

The first term in (16) is the mutual information between feature values and classes. Like step 1, it is to be maximized. The second term is the summation of the mutual information between the candidate feature X and each of the previously selected features \(X_s\). The second term is to be minimized. The parameter \(\beta > 0\) controls the amount of interdependence among selected features. After a number of trials this parameter has been set to 0.5.
3: Repeat until M number of features are selected.

5. RESULTS

5.1. Feature Ranking

The following 27 features were initially considered (17). These represent most of the popular color spaces. Some of them are originally discrete ones, whereas the others were converted to integer numerical values in the range [0 . . 255] to keep the same interval as the previously known color components.

\[
X = \begin{bmatrix}
[\bar{R}_{\text{RGB}}, \bar{G}_{\text{RGB}}, \bar{B}_{\text{RGB}}, \bar{r}_{\text{rgb}}, \bar{g}_{\text{rgb}}, \bar{b}_{\text{rgb}}] \\
\bar{T}_{\text{TSL}}, \bar{S}_{\text{TSL}}, \bar{L}_{\text{TSL}}, \bar{L}_{\text{LUV}}, \bar{U}_{\text{LUV}}, \bar{V}_{\text{LUV}} \\
\bar{H}_{\text{HSL}}, \bar{S}_{\text{HSL}}, \bar{L}_{\text{HSL}}, \bar{L}_{\text{LAB}}, \bar{A}_{\text{LAB}}, \bar{B}_{\text{LAB}} \\
\bar{L}_{\text{LUX}}, \bar{U}_{\text{LUX}}, \bar{X}_{\text{LUX}}, \bar{Y}_{\text{YIQ}}, \bar{I}_{\text{YIQ}}, \bar{Q}_{\text{YIQ}} \\
\bar{Y}_{\text{cbcr}}, \bar{C}_{\text{cbcr}}, \bar{C}_{\text{cbcr}} \\
\end{bmatrix}
\] (17)

The feature selection approach was applied and the following ranking was obtained for the features, in order of highest conditional entropy between input features and output classes, and the interdependence between features:

\[
X = \begin{bmatrix}
[C_{\bar{r}_{\text{cbcr}}}, \bar{A}_{\text{LAB}}, \bar{R}_{\text{RGB}}, \bar{S}_{\text{TSL}}, \bar{T}_{\text{TSL}}, \bar{B}_{\text{LAB}}] \\
\bar{Q}_{\text{YIQ}}, \bar{L}_{\text{LAB}}, \bar{C}_{\text{cbcr}}, \bar{B}_{\text{RGB}}, \bar{g}_{\text{rgb}}, \bar{r}_{\text{rgb}} \\
\bar{X}_{\text{LUX}}, \bar{H}_{\text{HSL}}, \bar{I}_{\text{YIQ}}, \bar{G}_{\text{RGB}}, \bar{U}_{\text{LUX}}, \bar{b}_{\text{rgb}} \\
\bar{Y}_{\text{cbcr}}, \bar{S}_{\text{HSL}}, \bar{H}_{\text{HSL}}, \bar{L}_{\text{LUX}}, \bar{V}_{\text{LUV}}, \bar{U}_{\text{LUV}} \\
\bar{L}_{\text{TSL}}, \bar{L}_{\text{LUV}}, \bar{Y}_{\text{YIQ}} \\
\end{bmatrix}
\] (18)

Figures 1 and 2 show respectively the mutual information between input features and class output, and the mutual information between classes. The order in which features are indexed is similar to the one given in the initial feature set (17). The \(C_{\bar{r}_{\text{cbcr}}}\) component though is last in the initial set it has the highest mutual information with the skin class 1. Hence it is the first feature to be selected. The interdependence between color features can be seen from figure 2, which shows high mutual information between some features, and hence redundancies.

From the initial set of features (17) and the final set of features ranked in descending order of relevance (18), the following observations can be made:

- Most features appearing in the first ranks are those related to chrominance. Most features related to luminance and brightness are classified last.
- Once a feature is selected, e.g. the saturation from the TSL space, the same feature from the HSL space is classified in the last ranking.
- The first row contains both the R component from the RGB space and the Cr component from the YCbCr space, which both are related to the red component. This in fact can be explained by the dominance of the red component in skin based images.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{fig1}
\caption{Mutual Information between Input Features and Class}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{fig2}
\caption{Mutual Information between Input Features}
\end{figure}

5.2. Feature Selection

Once the features are ordered, remains a final stage to chose the best representative features. This process has been achieved through the use of the area under the ROC curve (AUC)[12] to find out which set of features gives the best classification results. The features are added one at a time and the AUC curve is computed. A total number of 50 images have been considered for performance evaluation. From the table (5.2) that represents the AUC versus the number of components, it can be noticed the existence of a high value at 5 features, meaning that with 5 selected features the performance is excellent. The first 5 to 6 features are the best from the selected ones (19). Notice also that even some features are good representative ones, they cannot be used on their own. The first selected feature for instance, the \(C_{\bar{r}_{\text{cbcr}}}\), has given a very low value for AUC when used on its own. However, when the second selected feature is added, the AUC increased to 0.789. This process continues until a maximum AUC value is found with the first 5 components. After this peak
value, addition of further components does not enhance much the AUC. Hence, the choice has been set to the first 5 color features to evaluate our skin classifier.

\[
X = [CT_{chcr}, A_{LAB}, R_{RGB}, S_{TSL}, T_{TSL}]
\]

(19)

5.3. Color Based Skin Detection

In order to evaluate the performance of the proposed approach, the best selected feature sets were used, with the Bayes classifier to derive the ROC and hence the AUC. Sample images are given in (figure 5.3 5.3).

<table>
<thead>
<tr>
<th>Component</th>
<th>(C_{T_{chcr}})</th>
<th>(A_{LAB})</th>
<th>(R_{RGB})</th>
<th>(S_{TSL})</th>
<th>(T_{TSL})</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC</td>
<td>0.601</td>
<td>0.789</td>
<td>0.905</td>
<td>0.963</td>
<td>0.969</td>
</tr>
</tbody>
</table>

Table 1. AUC for different Number of Input Features

The proposed technique has been compared to existing work. For instance, the model based on the RGB color space with a 16 component GMM model [9], when used with our database [11] gave an AUC of 0.909. Though our model is based on a single Gaussian model, detection results are excellent. It is understandable that to get good comparison results, similar conditions should be considered, nevertheless, accuracy in our case is based on many factors, such as the number of Gaussian components and the number of skin and non skin samples to consider.

6. CONCLUSION

A new attempt to select the best color set from existing color components for the purpose of skin detection has been proposed. The approach is based on the use of the mutual information between input features and output classes, and also between input features themselves to select the least redundant ones. The aim is to get the best and smallest feature set that could yield better results in term of skin data classification. As can be inferred from experimental results, the approach used with a Bayes classifier, performed very well. The approach is thus verified and could be further enhanced by adding more color features and using similar information theoretical based approaches. A larger human skin database and other classifiers could also be considered.

7. REFERENCES