International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:1, No:3, 2007

A New Face Detection Technique using 2D DCT and Self Organizing Feature Map

Abdallah S. Abdallah, A. Lynn Abbott, and Mohamad Abou El-Nasr

***Abstract*—**This paper presents a new technique for detection of human faces within color images. The approach relies on image segmentation based on skin color, features extracted from the two-dimensional discrete cosine transform (DCT), and self-organizing maps (SOM). After candidate skin regions are extracted, feature vectors are constructed using DCT coefficients computed from those regions. A supervised SOM training session is used to cluster feature vectors into groups, and to assign "face" or "non-face" labels to those clusters. Evaluation was performed using a new image database of 286 images, containing 1027 faces. After training, our detection technique achieved a detection rate of 77.94% during subsequent tests, with a false positive rate of 5.14%. To our knowledge, the proposed technique is the first to combine DCT-based feature extraction with a SOM for detecting human faces within color images. It is also one of a few attempts to combine a feature-invariant approach, such as color-based skin segmentation, together with appearance-based face detection. The main advantage of the new technique is its low computational requirements, in terms of both processing speed and memory utilization.

***Keywords*—**Face detection, skin color segmentation, self-organizing map.

I. Introduction

FACE detection is the problem is identifying locations within images at which human faces appear. This is an important task that can facilitate higher-level applications, such as face recognition, face tracking, surveillance, and human-computer interaction (HCI). Face detection has been an area of active research for more than a decade, and is complicated by variations in scale, orientation, illumination, occlusion, and skin color. Comprehensive surveys of the area are given in [1] and [2].

This paper presents a novel approach for face detection that derives from an idea suggested by Hjelmås and Low[1]. In their survey, they describe a preprocessing step that attempts to identify pixels associated with skin independently of face-related features. This represents a dramatic reduction in computational requirements over previous methods, where many systems generated multiresolution image pyramids before preparing lists of candidate face regions, but these pyramids are no longer needed after segmentation based on skin color.

The extraction of skin regions can be either pixel-based [3] or region based [4], [5], [6]. Even skin color varies by individual, studies have demonstrated that intensity rather than chrominance is the main distinguishing characteristic [2]. The choice of color space is of low importance, because the next stage typically uses an intensity (grayscale) representation of the segmented image for further processing. This grayscale version, sometimes called a "skin map," contains intensity values for skin pixels and the background is represented as black. An example of a color input image and its resulting skin map is shown in Fig. 1. For simplicity, previously prepared skin maps from the VT-AAST Database [14] have been used throughout this research.

A block diagram for our proposed technique is presented in Fig. 2. Using a segmented image as input, the first stage applies a morphological opening operation to refine region shapes slightly, and to subdivide regions that consist of large areas joined only by narrow connections. Region labeling is performed so that each region can be distinguished from all others. In the second stage, the two-dimensional (2D) discrete cosine transform (DCT) for each region is computed, and feature vectors are formed from the DCT coefficients. The last stage uses a self-organizing map (SOM) to classify each feature vector as either "face" or "not face." The output is therefore a set of locations in the original image at which the system has identified faces.

For benchmarking and evaluation of this new technique, a new image database has been compiled for training and testing purposes. The database consists of 286 images, containing a total of 1027 faces. Each image is available in four formats. In addition to original color versions, images segmented by skin color are provided as previously described. These GIF images are of size 300 × 225 pixels. More details concerning the database are given in [14].

The rest of this paper is organized as follows: Section II discusses region analysis and labeling steps for the new algorithm. Section III describes the process of feature extraction. Section IV is describes the design and architecture of the SOM neural network that is used in the system. Section V shows experimental results, and discusses possible modifications and improvements to the system. Section VI presents concluding remarks.

A. S. Abdallah is with Bradley Dept. of Electrical and Computer Engineering, Virginia Tech, Blacksburg, Virginia, USA.

A. L. Abbott, Associate Professor is with Bradley Dept. of Electrical and Computer Engineering, Virginia Tech, Blacksburg, Virginia, USA.

M. A. El-Nasr, Assistant Professor is with Computer Engineering Dept., Arab Academy for Science, Technology, and Marine Transport, Alexandria, Egypt.

(a)

(b)

Fig. 1 Example of image segmentation for face detection. (a) Color input image. (b) Segmented image, with skin pixels represented using intensity values

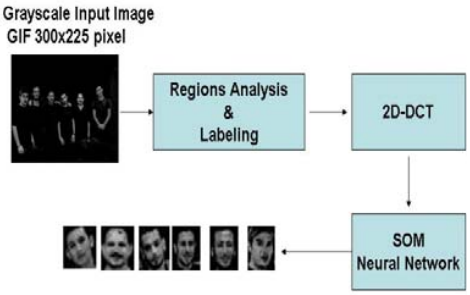


Fig. 2 Block diagram of the face-detection system. The input is a segmented "skin map." After a region analysis step, features are extracted using DCT coefficients for skin regions. Faces, if present, are then detected using a self-organizing map

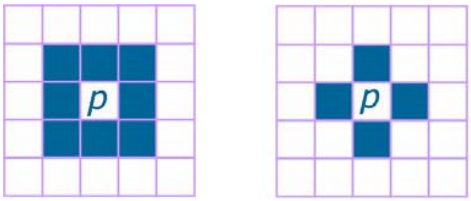
II. Region Analysis and Labeling

The input to the region analysis stage is a segmented image in which pixel values of detected skin regions are represented using intensity values. For simplicity of implementation, the value zero (black) is assigned to background pixels.

In this stage a morphological opening operation is applied to the segmented image, and this is followed by connected-components labeling [7]. The opening step modifies the boundary of a region slightly, typically reducing its curvature. More importantly for this application, morphological opening has the effect of subdividing large regions that are connected by a narrow connecting path, sometimes called a "neck." This has the effect of separating candidate face regions from other body parts, thereby improving the chances of correct detection.

Connected components labeling is a procedure that assigns a unique label to each region in the image. In effect, each skin region will receive a unique index that aids in further analysis. Region labeling in this system is done using eight-neighbor connectivity. A common alternative would be to use four-neighbor connectivity instead (Fig. 3), but this was not tested in our system. The last step for this stage is to use a nearest-neighbor interpolation scheme to resize each labeled region into a block of size 32 × 32 pixels, for later processing by the DCT.

Finally, the output of the region analysis and labeling stage is a set of arrays. Each array is of size 32 × 32 pixels, and represents a single candidate skin region. These steps are represented in Fig. 4.



(a) (b)

Fig. 3 The labeling connectivity schemes. (a) Eight neighbors. (b) Four neighbors

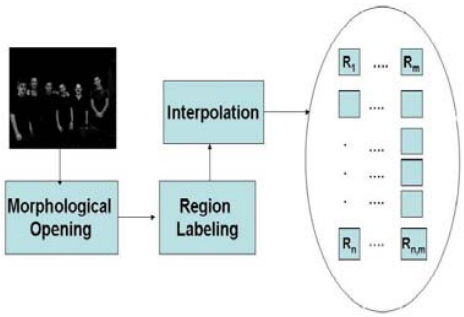


Fig. 4 Steps in the region analysis and labeling stage. Morphological opening is followed by region labeling to generate a set of candidate blocks of the same size. Blocks are ordered based on their upper-left corner indices

III. Feature Extraction

The discrete cosine transform is widely used for different applications. The most popular use of the DCT is for data compression, and it forms the basis for the well-known JPEG image compression format [8]. The extracted DCT coefficients can also be used as a type of signature that is useful for recognition tasks, such as facial expression recognition [9], [10]. Conceptually, each DCT coefficient can be viewed as representing a different feature dimension. This is the approach that has been developed for the face-detection presented here.

The proposed technique calculates the 2D-DCT for each cropped skin block coming out of the previous stage. This

results in a matrix of 32 × 32 coefficients. A subset of these values is taken to construct the feature vector. Empirically, the upper left corner of the 2D-DCT matrix contains the most important values, because they correspond to low-frequency components within the processed image block. For the initial experiments presented below, we use a set of 16 × 16 coefficients.

IV. Self Organizing Map

A. Overview

The Self Organizing Map, also called a Kohonen map, is a well known artificial neural network that can be trained using unsupervised or supervised learning approaches [11]. In the system described here, a SOM is employed to classify each DCT-based feature vector as either "facial skin" or "not facial skin".

The SOM used here contains N nodes ordered in a two-dimensional lattice structure. Both rectangular and hexagonal lattices were tested. In these cases, each node has 4 or 6 neighboring nodes, respectively. Typically, the SOM has a life cycle of three phases: the initialization phase, the learning phase, and the testing phase. The codebook vector for each node can be initialized randomly or linearly. During training, a learning function is responsible for updating the codebook vectors of the neurons that are located in predefined neighborhoods of the winning neuron. The most widely used training function is a parameterized Gaussian function,

$$h_{ci}(t) = \lambda(t) \exp\left(-\frac{\| x_c - x_i \|}{2\sigma^2(t)} \right) \quad (1)$$

where x_c and x_i represent a codebook vector and an input vector, respectively, $\lambda(t)$ is the learning rate, and $\sigma(t)$ represents the radius of the neighborhood set of nodes. The winning metric is usually based on Manhattan or Euclidean distance between the input vector and each entry in the node's codebook:

$$\arg \min_{1 \le i \le N} \left\{ \left\| x - v_i \right\| \right\} \quad (2)$$

As the training progresses, both $\lambda(t)$ and $\sigma(t)$ decrease. Finally, testing is performed by comparing the error computed for each input feature vector against a specified threshold.

Training and testing for our system were performed using the SOM Toolbox [12]. Three phases of the SOM implementation are now described.

B. Supervised Training

During the training phase, labeled feature vectors are presented to the SOM one at a time. For each node, the number of "wins" is recorded along with the label of the input sample for each win. The codebooks for the nodes are updated as described above.

By the end of this stage, each node of the SOM has two recorded values: the total number of winning times for facial input samples, and the total number of winning times for non-facial input samples.

C. Voting

A simple voting approach is next used to assign a label to each SOM node, such that the label represents the class that is associated with the largest number of winning times. In case of a tie, the processed node receives a blank label. This is also true for nodes having zero winning times.

D. Testing Phase

During the testing phase, each input vector is compared with all nodes of the SOM, and the best match is found based on minimum distance, as given in (2). Euclidean distance was used in our tests. The final output of the system is the label associated with the winning node.

V. Experimental Results

The VT-AAST image database [14] was created for the purpose of benchmarking face-detection techniques. It is divided into two subsets, for separate training and testing of the system. During SOM training, 129 images were used, containing 439 facial skin regions and 770 non-facial skin regions. The testing phase used 157 images with 588 facial and 1369 non-facial cases.

The face detection method presented in this paper was developed, trained, and tested using MATLAB™ 7.1. The computer was a Windows XP machine with a single 2.00 GHz Centrino processor and 512 MB of RAM.

This section presents results for 4 experiments in which different system parameters were altered. In the first experiment, the effect of the size of the SOM was studied in order to find the optimum grid size. Table I shows the detection results for this experiment, including detection rates, and false positive rates for different sizes of the SOM network. Each of these networks was trained using a sample array of size 1209 × 256, and each was tested using a sample array of size 1957 × 256. The best detection rate obtained here was 77.71% for an SOM of 100 nodes arranged in a 10 × 10 hexagonal lattice. This SOM size was used for all subsequent experiments. Training was performed using the batch algorithm, and a total of 3000 epochs.

The second experiment studied the effect of DCT block size on accuracy of detection, with each DCT coefficient used in the feature vector. Table II shows that slightly better performance was obtained for the case of 16 × 16 block sizes, and this block size was used in all subsequent experiments. This is also shown graphically in Fig. 5.

The third experiment compared the detection results for a hexagonal lattice topology with the case of a rectangular lattice. Table III presents the results for both cases. Better results were obtained for the hexagonal lattice, presumably because each updating step affects a larger number of neighbors. This is also the reason that hexagonal lattices require shorter training times than rectangular lattices, as the update rule for more nodes leads to faster convergence.

The previous experiments were not particularly concerned with computational requirements. Most of the computational

load comes from the large sizes of the feature vectors being used, whereas the memory needs derive primarily from the size of the SOM. Then the aim of fourth experiment was to determine whether a smaller feature vector could be constructed from a given set of DCT coefficients, without significant degradation in system performance.

TABLE I
THE EFFECT OF NETWORK SIZE FOR THE SELF-ORGANIZING MAP (SOM)
The SOM size versus detection results for three different sizes

	SOM Structure		
	Size = 50 (5 × 10)	Size = 100 (10 × 10)	Size = 150 (15 × 10)
Detection Rate (%)	74.78	77.71	73.29
False Positive Rate (%)	5.97	5.14	9.46

TABLE II
DETECTION RATES VS. DCT BLOCK SIZE
Maximum detection rate is obtained at (16 x 16) block

	DCT Block size				
	8 x 8	12 x 12	16 x 16	20 x 20	24 x 24
Detection Rate (%)	76.84	76.56	77.71	76.45	76.39

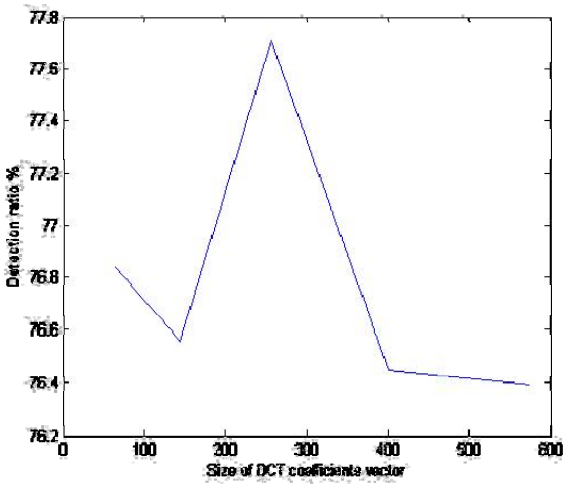


Fig. 5 Detection rate versus size of feature vector. The best results were obtained for a feature space of dimensionality 256

TABLE III
HEXAGONAL LATTICES VS. RECTANGULAR LATTICES

	Size =100 (10 × 10) Hexagonal	Size = 100 (10 × 10) Rectangular
Detection Rate (%)	77.71 %	76.89 %
Total training time (s)	446.844	406.156

A statistical analysis was conducted on a set of 3166 facial and non-facial skin samples. For the chosen DCT block size of 16 × 16, a total of 256 DCT coefficients were computed for each sample. Each of these coefficients can be viewed as representing a separate dimension in a 256-dimensional feature space. By assessing the variance in each dimension of this space, it is possible to determine which of the coefficients contribute most to the final decision of the classifier. Variances were computed using

$$\mathrm{var}\left(x_j\right) = \sum_{i=1}^{k} (X_i - \bar{X})^2 \quad (3)$$

where variable j is the DCT coefficient index, i is the sample index, and k is equal to the available number of samples.

Fig. 6 shows the statistical variances of the 256 features for this set of samples. The graph indicates that there are seven prominent peaks where high variance values occur. This suggests that these particular features perform prominent roles during classification. To exploit this, we defined a reduced-size feature space based on these high-variance DCT coefficients alone, whereas all other coefficients were excluded. The new feature vectors consist of only 27 DCT coefficients. Note that this reduction of dimensionality is similar to a Karhunen-Loeve-type analysis.

Table IV compares the performance of the system for full-size and reduced-size feature spaces. In spite of the dramatic reduction from 256 features to only 27, the detection rates are essentially the same. In addition to detection rates, the table also shows training times and memory usage for the both batch training and sequential training. The training length is kept constant, for fair comparison regarding the training time. The minimum value of the needed memory size M in case of batch training is roughly estimated by

$$M = 8(5(m+n)d + 3m^2) \quad (4)$$

as described by [12], where m is the number of SOM neurons, n represents the number of samples, and d is the dimensionality of the input feature vector. The memory requirement is therefore proportional to the size of the feature vector.

This last experiment has demonstrated that good face detection performance is possible, even with feature vectors that are dramatically reduced in size relative to the usual case for DCT-based analysis. This makes the proposed method much more attractive for low-cost, real-time implementation of a face-detection system. Commercial implementations for the SOM already exist [13]; thus it is conceivable that practical SOM-based face detection may be possible in the future.

VI. CONCLUSIONS AND FUTURE WORK

This paper has presented a novel face detection technique that uses features derived from DCT coefficients, along with a SOM-based classifier. The system has been tested using a sizable database containing 1027 faces in many orientations, sizes, and skin colors, and it achieved a detection rate of 77.94%. A reduced feature space, described for experiment 4 above, dramatically reduces the computational requirements

of the method as compared with standard DCT feature-extraction methods. This makes our system well suited for low-cost, real-time hardware implementation.

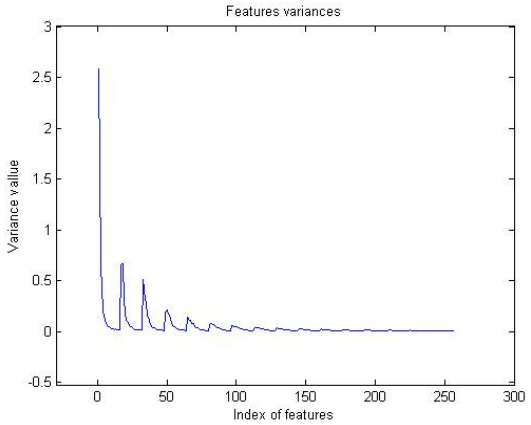


Fig. 6 Statistical variances of the feature component values indicate seven significant cases of high-variance features

TABLE IV
THE EFFECT OF REDUCING FEATURE VECTOR SIZE

DCT size	Batch time (s)	Batch detection rate	Memory consumption (Byte)	Sequential time ($s \times 10^3$)	Sequential detection rate
27	82.516	77.94 %	1068240	1.0009	76.28 %
256	400.328	77.71 %	12712320	1.9070	77.60 %

REFERENCES

[1] E. Hjelmås, and B. K. Low, "Face detection: a survey", *Computer Vision and Image Understanding*, Vol. 83, No. 3, Sept. 2001, pp. 236-274.
[2] M. H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: a survey", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 1, Jan. 2002, pp. 34 - 58.
[3] V. Vezhnevets, V. Sazonov, and A. Andreeva, "A survey on pixel-based skin color detection techniques", in *Proc. Graphicon-2003*.
[4] H. Kruppa, M. A. Bauer, and B. Schiele, "Skin patch detection in real-world images", in *Proc. of the DAGM-Symposium*, 2002, pp. 109-116.
[5] M.-H. Yang and N. Ahuja, "Detecting human faces in color images," in *Proc. of the International Conference on Image Processing*, vol. 1, pp. 127-130, 1998.
[6] B. Jedynak, H. Zheng, M. Daoudi, and D. Barret, "Maximum entropy models for skin detection," IRMA Technical Report, Vol. 57, No. XIII, Universite des Sciences et Technologies de Lille, France, 2002.
[7] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*. Prentice Hall, 2002.
[8] E. Ifeachor and B. Jervis, *Digital Signal Processing: A Practical Approach.* Prentice Hall, 2001.
[9] L. Ma, Y. Xiao, K. Khorasani, and R. K. Ward, "A new facial expression recognition technique using 2D DCT and k-means algorithm", in *Proc. International Conference on Image Processing*, Oct. 2004, pp. 1269-1272.
[10] L. Ma and K. Khorasani, "Facial expression recognition using constructive feedforward neural networks", *IEEE Transactions on Systems, Man and Cybernetics*, Part B, Vol. 34, No. 3, June 2004, pp. 1588 – 1595.
[11] T. Kohonen, *Self-Organizing Maps.* Springer, Berlin, Heidelberg, 1995.
[12] J. Vesanto, J. Himberg, E. Alhoniemi, and J. Parhankangas," Self-organizing map in Matlab: the SOM Toolbox", in *Proc. of Matlab DSP Conference*, Espoo, Finland, November 1999, pp. 30-40.
[13] D. Brown, I. Craw, and J. Lewthwaite, "A SOM based approach to skin detection with application in real time systems", in *Proc. of the British Machine Vision Conference*, 2001.
[14] A. Abdallah, M. Abou El-Nasr, and A. Lynn Abbott, "A new color image database for benchmarking of automatic face detection and human skin segmentation techniques", to appear in *Proc. of the Fourth International Conference on Machine Learning and Pattern Recognition* (MLPR 2007), Barcelona, Spain, Apr. 2007.