CERIAS Tech Report 2001-110
Face Detection For Pseudo-Semantic Labeling in Video Databases
by A Albiol, C Bouman, E Delp
Center for Education and Research
Information Assurance and Security
Purdue University, West Lafayette, IN 47907-2086

FACE DETECTION FOR PSEUDO-SEMANTIC LABELING IN VIDEO DATABASES

Alberto Albiol

Departamento de Comunicaciones
Universidad Politécnica de Valencia
Valencia, Spain
email: alalbiol@dcom.upv.es

Charles A. Bouman and Edward J. Delp

School of Electrical and Computer Engineering
1285 Electrical Engineering Building
Purdue University
West Lafayette, IN 47907-1285,USA
email: {bouman,ace}@ecn.purdue.edu

ABSTRACT

Pseudo-semantic labeling represents a novel approach for automatic content description of video. This information can be used in the context of a video database to improve browsing and searching. In this paper we describe our work on using face detection techniques for pseudo-semantic labeling. We present our results using a database of MPEG sequences.

1. INTRODUCTION

The amount of digital video available for many applications has undergone explosive growth in recent years. However the technologies for organizing and searching video data are in their infancy. Many of the issues in the development of video databases are being addressed in the emerging MPEG-7 standard [1]. In order to develop and test new algorithms for browsing and searching in video databases, complete systems must be implemented. An example of one of these systems is *ViBE* (Video Indexing and Browsing Environment) [2, 3, 4] which is being developed at Purdue University. *ViBE* is a browseable/searchable paradigm for organizing video data containing a large number of sequences. One of the techniques that *ViBE* uses for content description and retrieval is known as pseudo-semantic labeling.

Pseudo-semantic labeling bridges the gap between low-level and truly semantic labels for the description of video content. We are investigating in *ViBE* the use of several pseudo-semantic labels. These include labels such as "face," "indoor/outdoor," "high action," "man made," and "natural." Since pseudo-semantic labels are generally associated with some level of uncertainty, each label is assumed to be a continuous value in $[0, 1]$ where 1 indicates that the video frame has the associated property with high confidence, and 0 indicates that it does not. This values will be used in other subsystems of *ViBE* to improve the browsing and searching in the video database.

In this paper, we describe our work on the "face" label. The goal here is to detect whether a frame in the sequence contains a face. Different approaches have been developed in recent years for the face detection problem. Some of the most representative work includes shape-feature approaches [5]. Neural network approaches are used in [6, 7], and template matching approaches are used in [8, 9, 10]. These approaches have tended to focus on still gray-scale images. While they report good performance, they are often computationally expensive. This is especially true of template matching and neural network approaches, since they require processing for each possible position and scaling of the image. Also, the results may still be excessively sensitive to the rotation and pose of the face. Recently, a number of authors have used color and shape as important cues for face detection. In [11, 12, 13] both color and spatial information is exploited to detect segmented regions corresponding to faces. However, in [12, 13] the color and spatial information is also used to merge image regions in the segmentation process.

2. ALGORITHM DESCRIPTION

Our method for face labeling is designed to be robust for variations that can occur in face illumination, shape, color, pose, and orientation. To achieve this goal, our method integrates information we have regarding face color, shape, position and texture to identify the regions which are most likely to contain a face. It does this in a computationally efficient way which avoids explicit pattern matching.

This research has been partially supported by the grants TIC 98-0442 and TIC 98-0335 of the Spanish Government

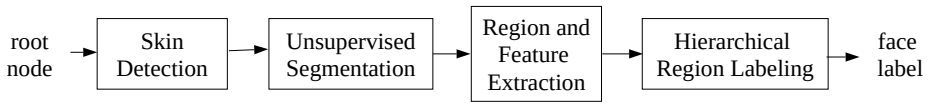

Figure 1: Block diagram illustrating the processing steps used to obtain the pseudo-semantic "face" label.

ViBE incorporates Shot Boundary Detection and Shot-Tree representation [2] that is used to temporally segment the video sequence and organize it into an efficient data structure. This representation greatly reduces the task of pseudo-semantic labeling. For the work presented in this paper, only one frame per shot is labeled. We should note that all processing is done on full frames extracted from the MPEG data.

Figure 1 illustrates the processing steps that we use to obtain the face label for a given frame. In the following subsections we will describe each of these blocks.

2.1. Skin Detection

The first step, skin detection, is used to segment regions of the image which potentially correspond to face regions based on pixel color. Under normal illumination conditions skin colors fall into a small region of the color space [14], so pixels falling outside this region are unlikely to be of interest.

The skin detection works by modeling the skin colors using a Gaussian mixture distribution, and then segmenting the image with a multiscale Bayesian segmentation algorithm known as sequential maximum a posteriori (SMAP) [15]. We use the expectation maximization (EM) algorithm to estimate the parameters of the Gaussian mixture distribution from a training set of skin pixels [16]. The training set is obtained from more than 200 images randomly selected from our database which were manually segmented into skin and no skin regions.

The white pixels of the images of the second column of Figure 2 represent the pixels classified as skin for three different examples.

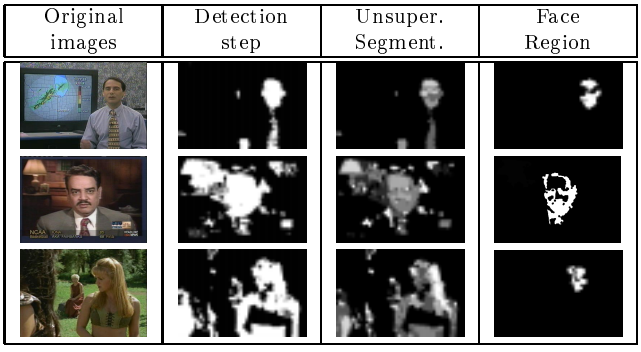

Figure 2: Results of each of the four steps used in obtaining the face label.

2.2. Unsupervised Segmentation

Once the pixels of interest are identified, unsupervised segmentation is used to separate these pixels into smaller regions which are homogeneous in color. This is important because the skin detection will produce non homogeneous regions often containing more than a single object. For example, skin colored materials may be erroneously included in the skin detection. These undesired regions are separated from true faces using the subsequent unsupervised segmentation step.

To do this, we again use the EM algorithm to cluster the detected skin pixels using a multivariate Gaussian mixture distribution. Each component of the Gaussian mixture is then treated as a distinct subclass of the skin color space. Each pixel of the detected skin region is then individually classified to its highest probability subclass (i.e. the MAP classification), and the resulting segmentation is smoothed using a morphological opening and closing operation with a 5×5 circular kernel. This morphological smoothing eliminates many small spurious regions produced by the MAP classification. As we can see in the third column of Figure 2, the unsupervised segmentation process further partitions the skin detected regions into smaller more homogeneous regions.

2.3. Region Extraction and Hierarchical Region Labeling

The next step is to extract connected regions and features vectors which characterize those regions. For each connected region, we compute a set of features which we call a state vector that describes the color and shape of the region. Let Ω_i be the i^{th} region of pixels, and let x_r be a color pixel in YUV coordinates at position $r \in \Omega_i$. Here $r \in \Omega_i$ is assumed to be a two dimensional column vector $r = [r_1, r_2]^t$ where r_1 and r_2 index the row and column of the pixel. For each region, we will extract a state vector of information defined by $q_i = [N_i, \bar{x}_i, \gamma_i, \mu_i, R_i, max_i, min_i]$. Each feature of q_i describes a characteristic of the color or shape of the region. The elements of the state vector are given by

the following expressions:

$$\begin{aligned} N_i &= \textstyle\sum_{r\in\Omega_i} 1 \\ \bar{x}_i &= \frac{1}{N_i}\textstyle\sum_{r\in\Omega_i} x_r \\ \gamma_i &= \mathrm{Diag}\left\{\frac{1}{N_i}\textstyle\sum_{r\in\Omega_i}(x_r-\bar{x}_i)(x_r-\bar{x}_i)^t\right\} \\ \mu_i &= \frac{1}{N_i}\textstyle\sum_{r\in\Omega_i} r \\ R_i &= \frac{1}{N_i}\textstyle\sum_{r\in\Omega_i}(r-\mu_i)(r-\mu_i)^t \\ max_i &= (\max_{r\in\Omega_i} r_1, \max_{r\in\Omega_i} r_2) \\ min_i &= (\min_{r\in\Omega_i} r_1, \min_{r\in\Omega_i} r_2) \end{aligned} \quad (1)$$

where γ_i is a vector containing the variance for each of the Y, U, and V color components. An important property of the state vector is that it can be recursively updated for merged regions. For example, let Ω_i and Ω_j be two regions that are merged to form a new region Ω_h. The new state vector for the region h may be expressed as $q_h = U(q_i, q_j)$ where $U(\cdot,\cdot)$ is a function of the two original state vectors.

From the state vectors q_i, we obtain a feature vector v_i which contains the specific features we will use for labeling regions. The elements of v_i are $\bar{x}_i$, γ_i, μ_i, the area of the bounding box derived from max_i and min_i and normalized by N_i, and three more features obtained from the eigenvalues and eigenvectors of R_i. Let $\lambda_1 > \lambda_2$ be the two eigenvalues of R_i with corresponding eigenvectors e_1 and e_2. Then the three remaining features are $\sqrt{\lambda_1\lambda_2}/N_i$, the orientation of e_1, and λ_2/λ_1.

A multivariate Gaussian distribution is used to model the behavior of v_i for face areas. Let $\bar{v}$ and Σ_v be the mean and covariance of the vector v_i for regions corresponding to face areas. Then the dissimilarity between v_i and the mean behavior for a face region is given by:

$$C_i = (v_i - \bar{v})^t \Sigma_v^{-1} (v_i - \bar{v}) \quad (2)$$

The smaller the value of C_i the greater the likelihood that the region is a face area. In our experiments, $\bar{v}$ and Σ_v are extracted as the sample statistics from a set of 100 manually segmented frames, each containing at least one face.

As discussed earlier, the unsupervised segmentation is likely to partition face areas into more than one connected region. Therefore, we must merge regions to form a new region which best matches the behavior expected for face areas. We do this by searching each pair-wise merging of regions to find the new region which minimizes (2). We first form a matrix $C_{i,j}$ defined by

$$C_{i,j} = (v_{i,j} - \bar{v})^t \Sigma_v^{-1} (v_{i,j} - \bar{v}) \quad (3)$$

where $v_{i,j}$ is the feature vector computed by merging regions Ω_i and Ω_j. We then search for the minimum

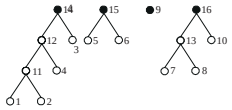

Figure 3: Binary tree structure resulting from skin-like region merging.

entry of the matrix, $C_{i^*,j^*} = \min_{(i,j)\in P} C_{i,j}$, where P is the set of entries of the matrix defined by

$$P = \{(i,j) : C_{i,j} < C_{i,i} \text{ and } C_{i,j} < C_{j,j}\}$$

The restriction on the set P insures that the minimum occurs between a pair of distinct regions which, when merged, will match the face hypothesis better then any existing individual region. Once the two distinct regions Ω_{i^*} and Ω_{j^*} are identified, these two regions are merged to form a new region $\Omega_h = \Omega_{i^*} \cup \Omega_{j^*}$. After merging, the original regions Ω_{i^*} and Ω_{j^*} must be removed from the matrix and replaced by the new region Ω_h, and the entries of $C_{i,j}$ must be updated.

This process of merging is repeated until the set P is empty. At this point, the merging of any two regions will only reduce the quality of the match to the face hypothesis. Figure 3 illustrates how this recursive merging process progresses. Each node represents a region of the image with the internal nodes representing regions that result from merging. The merging process terminates in a set of nodes, in this example nodes 9, 14, 15, and 16. Any of these nodes which contains less then 600 pixels are removed, and the region that best fits the face hypothesis is used to compute the face label. This is done using the equation:

$$p = \begin{cases} -\frac{C^*}{40} + 1 & C^* < 40 \\ 0 & C^* > 40 \end{cases} \quad (4)$$

where $C^* = \min_i C_i$ where i indexes the remaining nodes. Notice that $p \in [0,1]$ is larger if the frame is more likely to contain a face, and smaller if it is less likely. The fourth column of Figure 2 shows the composite region which best fits the face hypothesis for each example.

3. RESULTS

The algorithm described was tested on 7 sequences belonging to the $ViBE$ video database. Ground truth information relative to "face" and "no face" was determined for each frame used for our tests. These images were then used for classification. The results were evaluated in the following way:

- A False Alarm results if the frame contains no faces, but our algorithm says that there is at least one face.

Seq.	Shots	Faces	DE (%)	FA (%)	CC (%)
news1	231	76	73.68	16.13	80.51
news2	72	29	93.1	23.25	83.33
news3	78	33	87.87	15.55	85.89
news4	103	42	90.47	13.11	88.34
news5	188	51	76.47	13.86	83.51
movie	142	92	84.78	28	80.28
drama	100	90	94.4	20	93
total	914	413	85.23	16.96	84.02

Table 1: Results for face detection using the ground truth sequences. Faces indicates the number shots containing at least one face. DE indicates the Detection rate. FA indicates the False Alarm rate and CC the Correct Classification rate.

- A Detection results if the frame contains at least one face, and our algorithm says that there is at least one face.
- A Correct Classification results when a Detection is achieved or when there is no face in the frame and our algorithm finds no face in the frame.

Once we have merged skin-like regions, we apply a threshold to the pseudo-semantic label value given by Equation 4 to the remaining regions. If there exist any value greater than the threshold, the frame is said to contain at least one face. Thus an optimal threshold can be obtained that maximizes the Classification rate.

It is important to emphasize that no thresholds are used when similarity measurements are performed. Our algorithm produces confidence values that are then used by the *ViBE* browsing environment. Thresholds are used only in this paper to show the performance of the face labeling algorithm.

In Figure 4, a bounding box is drawn in each image for regions whose pseudo-semantic label value is greater than 0.5. As shown, we are able to detect faces under many different illumination conditions and poses. In Figure 4g, we can see that two boxes are drawn. In the classical face-detection problem this would have been classified as a false alarm. In our application, we do not care about such false alarms because our objective is to determine whether a face is presented or not. Figures 4l, 4m, 4n and 4o show false alarms where faces are detected in frames containing no human faces. In Table 1, the results for each ground truth sequence are shown separately.

4. REFERENCES

[1] MPEG Requirements Group, "Applications for MPEG-7," in *Doc. ISO/MPEG N2462*, MPEG Atlantic City Meeting, October 1998.

[2] Jau-Yuen Chen, Cuneyt Taskiran, Alberto Albiol, Charles A. Bouman, and Edward J. Delp, "ViBE: A compressed video database structured for active browsing and search," *Submitted to IEEE Transactions on Image Processing*, 1999. This paper is available at ftp://skynet.ecn.purdue.edu/pub/dist/delp/vibe-ip/

[3] Jau-Yuen Chen, Cuneyt Taskiran, Edward J. Delp, and Charles A. Bouman, "ViBE: A new paradigm for video database browsing and search," in *IEEE Workshop on Content-Based Image and Video Databases*, Santa Barbara, CA, June 21 1998, pp. 96–100.

[4] Cuneyt Taskiran, Jau-Yuen Chen, Charles A. Bouman, and Edward J. Delp, "A compressed video database structured for active browsing and search," in *Proceedings of IEEE Int'l Conference on Image Processing*, Chicago, IL, October 4-7 1998, pp. 133–137.

[5] Kin Choong Yow and Roberto Cipolla, "Feature-based human face detection," *Image and Vision Computing*, vol. 15, no. 9, pp. 713–735, 1997.

[6] Henry A. Rowley, Shumeet Baluja, and Takeo Kanade, "Human face detection in visual scenes," *Advances in Neural Information Processing Systems*, vol. 8, pp. 875–881, 1996.

[7] Kah-Kay Sung and Tomaso Poggio, "Example-based learning for view-based human face detection," Technical Report AI Memo 1521, MIT, 1994.

[8] Baback Moghaddam and Alex Pentland, "Probabilistic visual learning for object detection," in *Proceedings of the fifth International Conference on Computer Vision*, Cambridge, MA, 1995, pp. 786–793.

[9] Antonio J. Colmenarez and Thomas S. Huang, "Face detection with information-based maximum discrimination," in *Proceeding of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Juan, PR, 1997, pp. 782–787.

[10] Michael S. Lew, "Information theoretic view-based and modular face detection," in *2nd. International Conference on Automatic Face and Gesture Recognition*, Killington, VT, October 1996, pp. 198–203.

Figure 4: The boxes indicate the faces detected (threshold $p = 0.5$).

[11] N. Herodotou, K. Plataniotis, and A. Venetsanopoulos, "A color segmentation and classification scheme for facial image and video retrieval," in *IX European Signal Processing Conference*, Rhodes, Greece, September 1998.

[12] Veronica Vilaplana, Ferran Marques, Philippe Salembier, and Luis Garrido, "Region-based segmentation and tracking of human faces," in *European Signal Processing*, Rhodes, September 1998, pp. 593–602.

[13] Ming-Hsuan Yang and Narendra Ahuja, "Detecting human faces in color images," in *Proceedings of IEEE Int'l Conference on Image Processing*, Chicago, IL, October 4-7 1998, pp. 127–130.

[14] Hualu Wang and Shih-Fu Chang, "A highly efficient system for automatic face region detection in MPEG video," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 7, no. 4, pp. 615–628, August 1997.

[15] C. A. Bouman and M. Shapiro, "A multiscale random field model for Bayesian image segmentation," *IEEE Trans. on Image Processing*, vol. 3, no. 2, pp. 162–177, March 1994.

[16] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society B*, vol. 39, no. 1, pp. 1–38, 1977.