Combining Skin-Color Detector and Evidence Aggregated Random Field Models Towards Validating Face Detection Results

Author List
Institute Name

Abstract

In this paper, a framework for validating any generic face detection algorithm's result is proposed. A two stage cascaded face validation filter is described that relies on a skin-color detector and on a face silhouette structure modeler towards increasing face detection capacity of any face detection algorithm. While the skin-color detector combines a static skin-color and a dynamic background-color modeler, the face silhouette structure modeler incorporates an aggregate of random field models combined through a Demspter-Shafer framework of evidence merging. Together, the two modelers validate any face subimage generated by face detection algorithms. Experiments conducted on FERET and on an in-house face database supports the claim for improved face detection results using the proposed filter. An extension of the same framework towards head pose estimation is also suggested.

1. Introduction

Face detection has become an important first step towards solving plethora of other computer vision problems like face recognition, face tracking, pose estimation, intent monitoring and other face related processing. Over the years many researchers have come up with algorithms, that have over time, become very effective in detecting faces in complex backgrounds. Currently, the most popular face detection algorithm is the Viola-Jones [12] face detection algorithm whose popularity is boosted of by its availability in the open source computer vision library, OpenCV. Other popular face detection algorithms are identified in [3] and [15].

Most face detection algorithms learn faces by modeling the intensity distributions in upright face images. These algorithms tend to respond to face-like intensity distributions in image regions that do not depict any face as they are not contextually aware of the presence or absence of a human face. These spurious responses make the results unsuitable for further processing that requires accurate face images as inputs, such as the ones mentioned above. Figure 1 shows an example where a face detection algorithm detects two faces - one true and the other false.

Figure 1: An example false face detection.

The problem of false face detection has motivated some researchers to develop heuristic approaches aimed for validating the face detection results. Most of these heuristics integrate primitive context into the problem by searching for skin tone in the output subimages. However, this simple approach often fails to distinguish faces from non-faces, because face detectors often fail to center the cropping box precisely around the detected face. This produces a significant patch of skin colored pixels, but only a partial face. This centering problem can be dealt with by extracting the skin colored regions and comparing their shape to an ellipse. While such heuristics, are simple, and somewhat effective, their validation is not reliable enough to meet the needs of higher level face processing tasks. Further, they do not provide a confidence metric for their validation.

This paper treats the problem of face detection validation in a systematic manner, and proposes a learning framework that incorporates both contextual and structural knowledge of human faces. A face validation filter is designed by combining two statistical modelers, 1) a human skin-tone detector with a dynamic background modeler (Module 1), and 2) an evidence-aggregating human face silhouette random

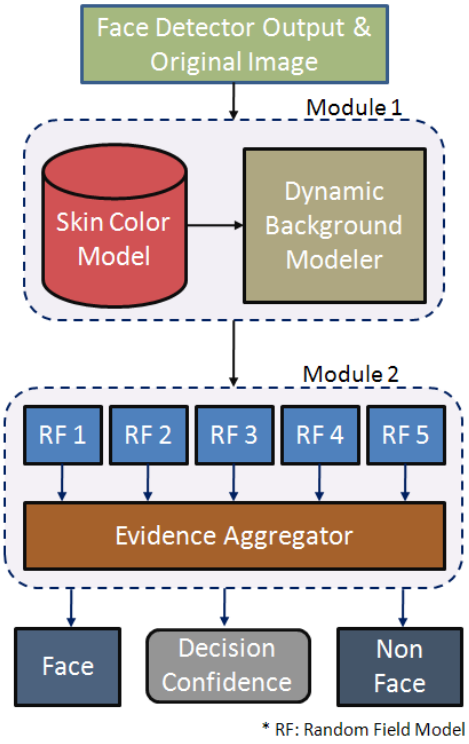

Figure 2: Block diagram.

field modeler (Module 2), which provides a confidence metric on its validation task. The block diagram in Figure 2 shows the functional flow of data through the two modules in the proposed framework. The details of the statistical models and their learning will be presented later in the paper, which is organized as follows. Section 2 reviews some of the earlier research. Section 3 introduces the proposed framework, with details on the learning process. Section 4 discusses the experiments carried out to test the proposed framework. Section 5 presents the results while Section 6 discusses them. Section 7 concludes the paper and discusses future work.

2. Related Work

As mentioned earlier, the problem of face detection validation has not been treated methodically before, though the problem has been handled by many as an integral component of face detection algorithms. All the past work in this area can be broadly characterized into two groups: a) Low level image feature models mostly based on skin color such as [2], [5] and [13], and b) High level facial feature models such as [4], [10] and [14].

The low level skin color based approaches try to reduce computational complexity by first identifying skin color in images so that search can be reduced. Most of the times, simple geometrical properties of the retained skin regions are used to determine if the region is a face. Such simplification of faces into trivial geometrical structures results in false detections. The facial feature based methods achieves face detection by individually identifying the integral components of a face image such as eyes, nose, etc. Though these schemes could be robust, the associated computational load is high. Interested readers could find more related references in [15] and [3]. The framework proposed in this paper uses statistically learnt knowledge about human faces to overcome computational complexity thereby augmenting face validation to existing face detection algorithms seamlessly.

3. Proposed Framework

As Shown in Figure 2, the framework essentially has two statistically learnt models, Module 1 and Module 2, that are cascaded to form the face detection validation filter. The output from a face detector is sent to Module 1, which distinguishes the skin pixels in the face region from the background pixels, thereby constucting a skin region mask. This skin region mask then becomes the input to Module 2, which is essentially an aggregate of random field models learnt from manually labeled (*true*) face detection outputs. The results of each random field model within the aggregate are then combined, using rules of Dempster-Shafer Theory of Evidence [9]. This *combining of evidence* provides a metric for the belief (i.e. confidence) of the system in its final validation. The two modules are detailed in the following subsections.

3.1. Module 1: Human Skin Tone Detector with Dynamic Background Modeler

Most of the skin tone detectors used for human skin color classification use prior knowledge, which is provided in the form of a parametric or non-parametric model of skin samples that are extracted from images - either manually, or through a semiautomated process. In this paper we employ such an a priori model, in combination with a dynamic background modeler, so that the skin vs. non-skin boundary is accurately determined. Accurate skin region extraction is essential for Module 2, as it validates images based on their structural properties. The two functional components of Module 1 are:

3.1.1 *a-priori* Bi-modal Gaussian Mixture Model for Human Skin Classification

A normalized RGB color space has been a popular choice among researchers for parametric modeling of human skin color. The normalized RGB (typically represented as nRGB) of a pixel X with X_r, X_g, X_b as its red, green and

blue components respectively, is defined as:

$$X^{nRGB}_{i|i\in\{r;g;b\}} = \frac{X_i}{\left(\sum_{\forall_{i|i\in\{r,g,b\}}} X_i\right)} \quad (1)$$

Normalized RGB space has the advantage that only two of the three components, nR, nG or nB, is required at any one time to describe the color. The third component can be derived from the other two as:

$$X^{nRGB}_{i|i\in\{nR;nG;nB\}} = 1 - \left(\sum_{\forall_{k|(k\in\{nR,nG,nB\},k\neq i)}} X_k\right) \quad (2)$$

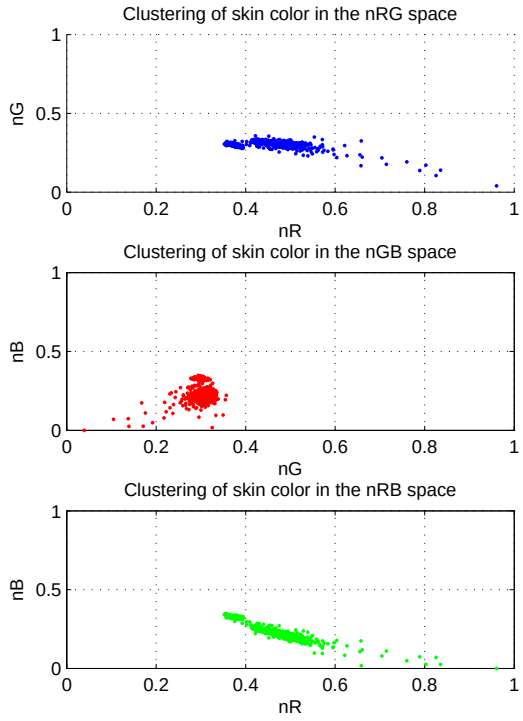

Figure 3: Skin pixels in nRGB space.

In our experiments, we found that skin pixels form a tight cluster when projected on nG and nB space as shown in the Figure 3. The study was based on a skin pixel database, consisting of nearly $150,000$ samples, built by randomly sampling skin regions from 1040 face images collected on the web as well as from FERET face database [8]. Further analysis also showed that the cluster formed on the 2D nG-nB space had two prominent density peaks which motivated the modeling of skin pixels with a Bi-modal Gaussian mixture model learnt using Expectation Maximization (EM) with a k-means initialization algorithm [1]. The Bi-modal Gaussian mixture model is represented as.

$$f^{skin}_{X|X=[nG;nB]}(x) = w_1 f_{Y_1}(x;\Theta_1=[\mu_1,\Sigma_1]) + w_2 f_{Y_2}(x;\Theta_2=[\mu_2,\Sigma_2]) \quad (3)$$

3.1.2 Dynamically Learnt Multi-modal Gaussian Model for Background Pixel Classification

As mentioned earlier, classification of regions into face or non-face requires accurate skin vs. non-skin classification. In order to achieve this, we learn the background color surrounding each face detector output dynamically. To this end we extract an extra region of the original image around the face detector's output, as shown in Figure 4. Since the size of the face detector output varies from image to image, it is necessary to normalize the size. This is done by down-sampling the size of the original image to produce a face detector output region containing 90x90 pixels. The extra region pixels surrounding the face are then extracted from the 100x100 region around this 90x90 normalized face region.

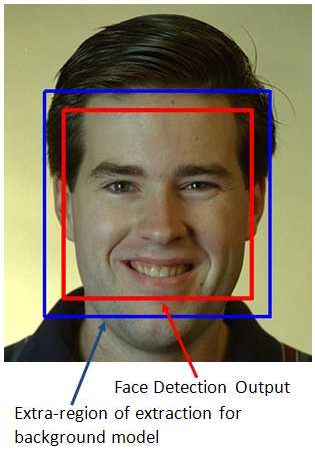

Figure 4: Extra region for background modeling.

Once the outer pixels are extracted, a Multi-modal Gaussian Mixture is trained using EM with k-means initialization, similar to the earlier case with skin pixel model. The resultant model can be represented as.

$$f^{non-skin}_{X|X=[R;G;B]}(x) = \sum_{i=1}^{m} w(i) f_{Y_i}(x;\Theta_i=[\mu_i,\Sigma_i]) \quad (4)$$

where, m is the number of mixtures in the model. We found empirically that a value of $m=2$ or $m=3$ modeled the backgrounds with sufficient accuracy.

3.1.3 Skin and Background Classification using the learnt Multi-modal Gaussian Models

The skin and non-skin models, $f^{skin}_{X|X=[nG;nB]}(x)$ and $f^{non-skin}_{X|X=[R;G;B]}(x)$ respectively, are used for classifying every pixel in the scaled face image obtained as explained in the Section 3.1.2. Example skin-masks are shown in Figure 5. This example shows two sets of images - one corresponding to a *true* face detection result, and another *false* face detection result.

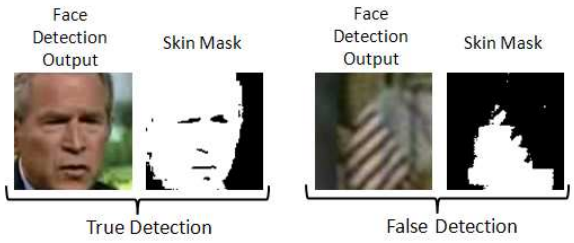

Figure 5: Example of *true* and *false* face detection.

The structural analysis through Random Field models explained in the next section will describe the design concepts that will help distinguish between *true* and *false* face detections shown in Figure 5.

3.2. Module 2: Evidence-Aggregating Human Face Silhouette Random Field Modeler

In order to validate the skin region extracted as explained in Section 3.1, we build statistical models from examples of faces. We developed statistical learners inspired by Markov Random Fields (MRF) to capture the variations possible in *true* skin masks (face silhouette). The following subsections describes MRF models and the variant we created for our experiments.

3.2.1 Random Field (RF) Models

In this work, we used a minor variant of MRFs to learn the structure of a *true* face skin mask. MRFs encompass a class of probabilistic image analysis techniques that rely on modeling the intensity variations and interactions among the image pixels. MRFs have been widely used in low level image processing including, image reconstruction, texture classification and image segmentation [7].

In an MRF, the sites in a set, $\mathcal{S}$, are related to one another via a neighborhood system, which is defined as $\mathcal{N} = \{\mathcal{N}_i, i \in \mathcal{S}\}$, where $\mathcal{N}_i$ is the set of sites neighboring i, $i \notin \mathcal{N}_i$ and $i \in \mathcal{N}_j \iff j \in \mathcal{N}_i$.

A random field X said to be an MRF on $\mathcal{S}$ with respect to a neighborhood system $\mathcal{N}$, if and only if,

$$P(\mathbf{x}) > 0, \forall \mathbf{x} \in \mathcal{X} \tag{5}$$

$$P(x_i|x_{\mathcal{S}-\{i\}}) = P(x_i|x_{\mathcal{N}_i}) \tag{6}$$

where, $P(x_i|x_{\mathcal{S}-\{i\}})$ represents a Local Conditional Probability Density function defined over the neighborhood $\mathcal{N}$. The variant of MRF that we created for our experiments relaxed the constraints imposed by MRFs on $\mathcal{N}$. Typically, MRFs requires that sites in set $\mathcal{S}$ be contiguous neighbors. The relaxation in our case allows for distant sites to be grouped into the same model.

We empirically found out that modeling the skin-region validation problem into one single RF gave poor results. We devised 5 unique RF models with an Dempster-Shafer Evidence aggregating framework that could not only validate the face detection outputs, but also provide a metric of confidence. Thus, Equation 6 could be alternatively seen as a set $P(\mathbf{x}) = \{P^1(\mathbf{x}), \ldots, P^5(\mathbf{x})\}$, each having their own neighborhood system $\mathcal{N}^k = \{\mathcal{N}^1, \mathcal{N}^2, \ldots, \mathcal{N}^5\}$, such that

$$P^k(x_i|x_{\mathcal{S}-\{i\}}) = P(x_i|x_{\mathcal{N}_i^k}) \tag{7}$$

3.2.2 Pre-processing

As described earlier, each face detector output is normalized and expanded to produce a 100x100 pixel image, from which a binary skin mask is generated. A morphological opening and closing operation is then performed on the skin mask (to eliminate isolated skin pixels), and the mask is then partitioned into one hundred 10x10 blocks, as shown in Figure 6. The number of mask pixels (which represent skin pixels) are counted in each block, and a 10x10 matrix is constructed, where each element of this matrix could contain a number between 0 and 100. This 10x10 matrix is then used as the basis for determining whether the face detector output is indeed a face.

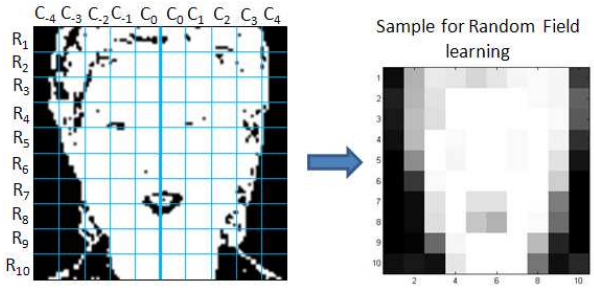

Figure 6: Pre-processing.

3.2.3 The Neighborhood System

The determination of whether the face detector output is actually a face is based on heuristics that are derived from anthropological human face models [11] and through our own statistical analysis. These include:

1. Human faces are horizontally symmetrical (i.e. along any row of blocks R_i) about a central vertical line joining the nose bridge, the tip of the nose and the chin cleft, as shown in Figure 6. In particular, our analysis of a large set of frontal face images showed that the counts of skin pixels in the 10 blocks that form each row in Figure 6 were roughly symmetrical across this central line.

2. The variations along the verticals (C_i's) are negligible enough that in building a Local Conditional Probability Density function, each R_i can be considered independent of the other. That is, for example, modeling

variations of C_0 w.r.t C_1 on R_1 is similar to modeling variations of C_0 w.r.t C_1 on any other $R_{i|i\neq 1}$. Thus, analysis of Local Conditional Probability could be restricted to single R_i at a time, as shown in Figure 7.

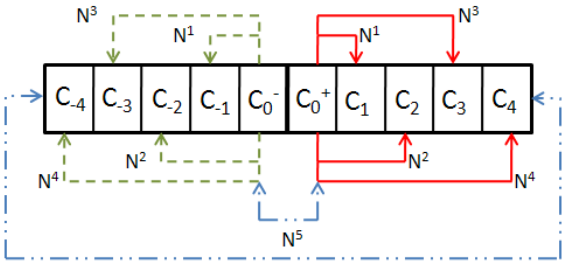

Figure 7: Neighborhood System.

The different neighborhood systems $\mathcal{N}^k$, used in the RF models, $P^k(x|x_{\mathcal{N}^k})$, can be defined as (Refer Figure 7):

$$\mathcal{N}^k = \{C_{j|j\in\{|k|;0^-;0^+\}}\} \tag{8}$$

3.2.4 Local Conditional Probability Density (LCPD)

To model the variations on the skin-region mask, we choose to build 2D histogram for each of the 5 RF over their unique neighborhood system. The design of the dimensions were such that they captured the various structural properties of *true* skin masks. The two dimensions (represented in a histogram pool $\mathbf{H}^k$) with individual element of the pool, $\mathbf{z}$, can be defined as:

- $\mathbf{H}^{k|k=\{1;2;3;4\}} = \{\mathbf{z}\}$, where,

$$\mathbf{z} = [x_{C_{0\pm}}, \delta(x_{C_{0\pm}}, x_{C_{\pm k}})], \forall R_j \tag{9}$$

- $\mathbf{H}^{k=5} = \{\mathbf{z}\}$, where,

$$\mathbf{z} = [\mu(x_{C_{0^+}}, x_{C_{0^-}}), \mu(x_{C_{-4}}, x_{C_{+4}})], \forall R_j \tag{10}$$

where, x_{C_k} is the count of skin pixels in the block C_k. The two functions $\delta(.,.)$ and $\mu(.,.)$ are defined as

$$\delta(x_{C_{0\pm}}, x_{C_{\pm i}}) = \begin{cases} x_{C_{0^+}} - x_{C_{+i}}, & i > 0 \\ x_{C_{-i}} - x_{C_{0^-}}, & i < 0 \end{cases} \tag{11}$$

$$\mu(a,b) = \frac{a+b}{2} \tag{12}$$

In order to estimate the LCPD on these 5 histogram pools, we use Parzen Window Density Estimation (PWDE) technique, similar to [6], with a 2D Gaussian window. Thus, each of LCPD can now be defined as

$$P^k(\mathbf{z}) = \frac{1}{(2\pi)^{\frac{d}{2}} n h_{opt}^d} \sum_{j=1}^{n} \exp \left[-\frac{1}{2h_{opt}^2} \left(\mathbf{z} - \mathbf{H}_j^k\right)^T \Sigma^{-1} \left(\mathbf{z} - \mathbf{H}_j^k\right) \right]$$

where, n is the number of samples in the histogram pool $\mathbf{H}^k$, d is number of dimensions (in our case 2), Σ and h_{opt} are the covariance matrix over $\mathbf{H}^k$ and the optimal window width, respectively, defined as:

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix}, \quad h_{opt} = \frac{\sigma_1 + \sigma_2}{2} \left\{ \frac{4}{n(2d+1)} \right\}^{1=(d+4)}$$

Figure 8 shows the 5 LCPDs learnt over a set of 390 training frontal face images.

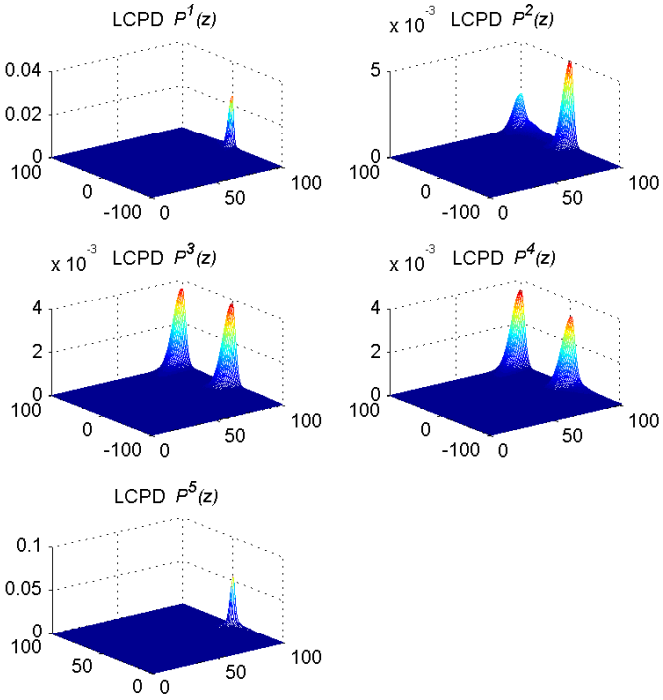

Figure 8: Frontal face Local Conditional Probability Density (LCPD) models.

3.2.5 Human Face Pose

During our studies we discovered that the structure of the skin-region varies based on the pose of detected face as shown in Figure 9. Combining face examples from different pose into one set of RFs seemed to dilute the LCPDs and hence the discriminating capability. This motivated us to design three different sets of RFs, one for each pose. This was accomplished by grouping *true* face detections into three piles, Turned right (r), Facing front (f), and, Turned Left (l). Thus, the final set of LCPDs could be described by the super set.

$$P(\mathbf{z}) = \left\{ P_{m|m=\{r;f;l\}}^{k|k=\{1;:::;5\}}(\mathbf{z}) \right\} \tag{13}$$

3.3 Combining Evidence

Given any test face detection output, $\mathbf{z}$ is extracted (as described in Equation 9 and 10) and projected on the LCPD

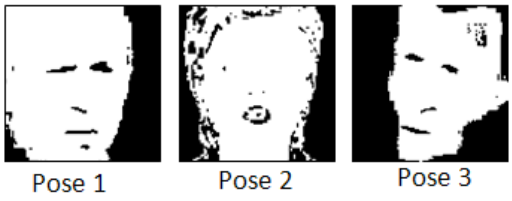

Figure 9: Skin-region masks.

set $P(\mathbf{z})$ to get a set of likelihoods l_m^k. As in the case of any likelihood analysis, we combined the joint likelihood of multiple projections using log-likelihood function, $L_m^k = \ln\left(l_m^k\right)$, such that,

$$\prod_{\forall \mathbf{z} \in \mathbf{H}_m^k} \ln\left(l_m^k(\mathbf{z})\right) = \sum_{\forall \mathbf{z} \in \mathbf{H}_m^k} L_m^k(\mathbf{z}) \tag{14}$$

Given these log-likelihood values, one can set hard thresholds on each one of them to validate a face subimage discretely as *true* or *false*. We incorporated a piece-wise linear decision model (soft threshold) instead of a hard threshold on the acceptance of a face subimage. This is illustrated in the Figure 10. Each LCPD $P^k(\mathbf{z})$ was provided with an upper and lower threshold of acceptance and rejection respectively. The upper and lower bounds were obtained by observing $P^k(\mathbf{z})$ for the three face poses $P_{r;f;l}^k(\mathbf{z})$. Thus, any log-likelihood values lesser than the lower threshold (L_L) would result in a decision against the test input (Probability 0), while any log-likelihood value greater that the upper threshold (L_U) would be a certain accept (probability 1). Anything in between would be assigned a probability of acceptance. In order to combine the decisions from the five

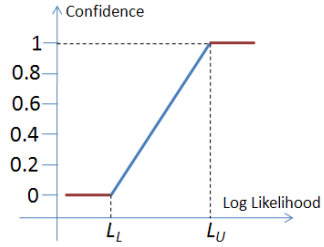

Figure 10: Soft threshold.

LCPD $P^k(\mathbf{Z})$, we resort to Dempster-Shafer Theory of Evidence.

3.3.1 Dempster-Shafer Theory of Evidence (DST)

The Dempster-Shafer theory is a mathematical theory of evidence [9] which is a generalization of probability theory with probabilities assigned to sets rather than single entities.

If X is an universal set with power set, $\mathbf{P}(X)$ (Power set is the set of all possible sub-sets of X, including the empty set $\emptyset$), then the theory of evidence assigns a belief mass to each subset of the power set through a function called the basic belief assignment (BBA), $m : \mathbf{P}(X) \rightarrow [0, 1]$, when it complies with the two axioms. a) $m(\emptyset) = 0$ and b) $\sum_{\mathbf{A} \in \mathbf{P}(X)} m(\mathbf{A}) = 1$. The mass, $m(A)$, of a given member of the power set expresses the proportion of all relevant and available evidence that supports the claim that the actual state belongs to A and to no particular subset of A. In our case, $m(A)$ correlates to the probability assigned by each of LCPDs towards the subimage being a face or not.

The true use of DST in our application becomes clear with the *rules of combining evidences* which was proposed as an immediate extension of DST. According to the rule, the combined mass (evidence) of any two expert's opinions, m_1 and m_2, can be represented as:

$$m_{1;2}(A) = \frac{1}{1-K} \sum_{B \cap C = A; A \neq \emptyset} m_1(B) m_2(C) \tag{15}$$

where,

$$K = \sum_{B \cup C = \emptyset} m_1(B) m_2(C) \tag{16}$$

is a measure of the conflict in the experts opinions. The normalization factor, $(1 - K)$, has the effect of completely ignoring conflict and attributing any mass associated with conflict to a null set.

The 5 LCPDs, $P^k(\mathbf{z})$, were considered as experts towards voting on the test input as a face or non-face. In order to use these mapped values in Equation 15 - 16, we normalized evidences generated by the experts to map between $[0, 1]$, and any conflict of opinions were added into the conflict factor, K. For the sake of clarity, we show an example of combining two expert opinions in Figure 11. The same idea could be extended to multiple experts.

		Expert 1's opinion	
		Face $m_1(B)$	Non-Face $m_1(C)$
Expert 2's Opinion	Face $m_2(B)$	Opinion Intersect $[m_1(B) * m_2(B)]$ (Sum in Numerator)	Opinion Conflict $[m_1(C) * m_2(B)]$ (Sum into K)
	Non-face $m_2(C)$	Opinion Conflict $[m_1(B) * m_2(C)]$ (Sum into K)	Opinion Intersect $[m_1(C) * m_2(C)]$ (Sum in Numerator)

Figure 11: An example of combining evidence from two experts under Dempster-Shafer Theory.

3.4 Coarse Pose estimation

Since the RF models were biased with pose information, we also investigated the possibility of determining the pose of the face based on the evidences obtained from the

LCPDs. We noticed that the LCPDs $P^3(\mathbf{z})$, $P^4(\mathbf{z})$ and $P^5(\mathbf{z})$ were capable of not only discriminating faces from non-faces, but were also capable of voting towards one of 3 pose classes, Looking right, Frontal, and Looking Left along with a confidence metric. Due to space constraints, the procedure is not explained in detail, but it is similar to what was followed for face versus non-face discrimination as explained in Section 3.3.

4. Experiments

In all our experiments, Viola-Jones face detection algorithm [12] was used for extracting face subimages. The proposed face validation filter was tested on two face image data sets, 1. *The FERET Color Face Database*, and 2. *An in-house face image database* created from interview videos of famous personalities.

In order to prepare the data for processing, face detection was performed on all the images in both the data sets. The number of face detections do not directly correlate to the number of unique face images as there are plenty of false detections. We manually identified each and every face detection to be *true* or *false* so that ground truth could be established. The details of this manual labeling is shown below:

1. *FERET*
   - Number of actual face images: $14,051$
   - Number of faces detected using Viola-Jones algorithm: $6,208$
   - Number of *true* detections: $4,420$
   - Number of *false* detections: $1,788$ (28.8%)
2. In-house database
   - Number of actual face images: $2,597$
   - Number of faces detected using Viola-Jones algorithm: $2,324$
   - Number of *true* detections: $2,074$
   - Number of *false* detections: 250 $(10.7\ \%)$

5. Results

In order to compare the performance of the proposed face validation filter, we defined four parameters:

1. **Number of false detections (NFD)**

$$\text{NFD} = \text{Count of false detections}$$

2. **False detection rate (FDR):**

$$\text{FDR} = \frac{\text{\# of false detections}}{\text{Total \# of face detections}} \text{x}100$$

3. **Precision (P)**

$$\text{P} = \frac{\text{\# of true detections}}{\text{\# of true detections} + \text{\# of false detections}}$$

4. **Capacity (C)**

$$\text{C} = \left(\frac{\text{\# of true detections}}{\text{\# of actual faces in database}} \right) - \text{FDR}$$

	Before Validation	After Validation
NFD	$1,788$	208
FDR	$28.8\ \%$	$3.35\ \%$
P	0.7120	0.9551
C	0.026	0.281

Table 1: Face detection validation results on FERET database.

	Before Validation	After Validation
NFB	250	2
FDR	$10.76\ \%$	$0.01\ \%$
P	0.892	0.999
C	0.691	0.798

Table 2: Face detection validation results on the in-house face database.

As explained in Section 3.4, the framework was extensible to perform coarse pose estimation. Figure 12 shows the result of passing two frames of a video sequence as input the face validation filter. The frames were extracted from a video of the same individual exhibiting arbitrary facial motion. The frames were 0.55 seconds apart. As can be noticed, the head pose is slightly different between the two frames. The pose estimation results are shown below the two frames.

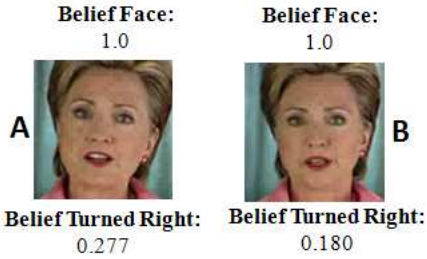

Figure 12: Coarse pose estimation.

6 Discussion of Results

Performance analysis of the proposed face validation filter can be understood through the four parameters defined in Section 5. **NFB** and **FDR** are direct measurements of the

number of mistakes (naming non-faces as faces) made by the face detection algorithm on the two data sets. As can be verified from Table 1 and 2, there is a significant reduction in the false detections through the introduction of the filter.

The precision parameter, **P**, can be perceived as the probability that a face detection result retrieved at random will truly contain a face. It can be seen that the precision of the system drastically improves with the introduction of the face validation filter thereby assuring a *true* face subimage at the output.

The capacity parameter, **C**, measures the relative difference between face detection and false detection rates of a face detection system. Alternately, **C** can be considered to measure the net *true* face detection ability of any algorithm on a specific face data set. **C** ranges from -1 to 1. -1 when none of the faces in the database are detected with all reported detections being wrong. 1 when all the faces in the database are detected with no false detections. It can be seen from Tables 1 and 2 that the capacity of the face detection system, when combined with face validation filter, is significantly higher and moves towards 1. One can thus infer that the combined system has better *true* face detection ability.

Finally, Figure 12 shows the coarse pose estimation results. The two frames in the figure shows cases when the face is slightly turned right, with one (**A**) turned more right than the other (**B**). The face validation filter verifies that the faces are actually turned right and the belief values represent a scale on the amount of rotation. Since we did not do any specific mapping of the belief values to pose angle, we could not confirm quantitatively how accurate the pose estimations were. Through visual consort, one can verify that the labeling is meaningful.

7. Conclusions and Future Work

In this paper a face detection validation filter is proposed that effectively combines a contextual skin color modeler and a structural face silhouette modeler towards increased *true* face detection ability. The proposed framework is tested with Viola-Jones face detection algorithm on two face databases (FERET and in-house face database) that confirm the increased *true* face detection. We also show how the proposed framework can be used towards coarse head pose estimation.

The work presented in this paper is part of a larger framework that could be used to improve skin region extraction on any face image by feeding back knowledge from the structural random field modeler into the skin detection module. As part of future work, we plan to migrate the static skin color model in Module 1 (see Figure 2) into a dynamic skin color model which will learn alongside the background modeler

References

[1] J. Bilmes. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. Berkeley CA, April, 1998. International Computer Science Institute, U.C. Berkeley.

[2] A. Hadid and M. Pietikainen. A hybrid approach to face detection under unconstrained environments. *18th International Conference on Pattern Recognition*, 1:227–230, 2006.

[3] E. Hjelmas and B. K. Low. Face detection: A survey. *Computer Vision and Image Understanding*, 83:236–274, Sep, 2001.

[4] M. B. Hmid and Y. B. Jemaa. Fuzzy classification, image segmentation and shape analysis for human face detection. *8th International Conference on Signal Processing*, 4, 2006.

[5] I. Naseem and M. Deriche. Robust human face detection in complex color images. *IEEE International Conference on Image Processing*, 2:338–41, 2005.

[6] R. Paget, I. D. Longstaff, and B. Lovell. Texture classification using nonparametric markov random fields. *13th International Conference on Digital Signal Processing Proceedings*, 1:67–70, 1997.

[7] P. Perez. Markov random fields and images. *CWI Quarterly*, 11:413–437, 1998.

[8] P. J. Phillips, H. Moon, P. Rauss, and S. A. Rizvi. The feret evaluation methodology for face-recognition algorithms. *Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition*, page 137, 1997.

[9] K. Sentz and S. Ferson. Combination of evidence in dempster-shafer theory. Sandia National Laboratories, April, 2002.

[10] U. Tariq, H. Jamal, M. Shahid, and M. Malik. Face detection in color images, a robust and fast statistical approach. *Proceedings of INMIC 2004. 8th International Multitopic Conference*, pages 73–78, 2004.

[11] M. Vezjak and M. Stephancic. An anthropological model for automatic recognition of the male human face. *Annals of Human Biology*, 21:363–380, 1994.

[12] P. Viola and M. J. Jones. Robust real-time face detection. *Int. J. Comput. Vision*, 57:137–154, 2004.

[13] M. Wimmer, B. Radig, and M. Beetz. A person and context specific approach for skin color classification. *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, 2:39–42, 2006.

[14] Y.-W. Wu and X.-Y. Ai. Face detection in color images using adaboost algorithm based on skin color information. *International Workshop on Knowledge Discovery and Data Mining*, pages 339–342, 2008.

[15] M.-H. Yang, D. Kriegman, and N. Ahuja. Detecting faces in images: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:34–58, 2002.