A Survey on Face Detection Methods

Ming-Hsuan Yang, Narendra Ahuja, and David Kriegman

Abstract

Human faces provide enormous information and a friendly interface in intelligent human computer interaction. This has motivated a very active research area on, among others, face recognition, face tracking, pose estimation, expression recognition and gesture recognition. However, most existing methods on these topics assume human faces in an image or a image sequence have been identified and localized. To build a fully automated system that analyzes information of human faces, it is essential to develop robust and efficient algorithms to detect human faces.

Given a single or a sequence of images, the goal of face detection is to identify and locate human faces regardless of their positions, scales, orientations and lighting conditions. Such problem is challenging because human faces are highly non-rigid objects with a high degree of variability in size, shape, color and texture. The purpose of this paper is to give a critical survey of existing techniques on face detection which has attracted a high degree of interest in recent years. We conclude this paper with several promising directions for future research.

I. Introduction

With the advent of new technology and media into our society, human computer interaction (HCI) has become an active research area in that more friendly and effective interfaces are being developed. Among all the interfaces between humans and computers, it is commonly believed that human faces is one of the most effective media since it carries enormous information that computers can react accordingly. For example, computers can adjust its behavior by knowing the user's feeling through his or her facial expression. Visual attention is another instance that computers can react based on the user's interests. Towards this goal, face and facial expression recognition has attracted more attention recently although it has been studied for more than twenty years by psychophysicists, neuroscientists and engineers. Many interesting and useful applications have been developed from these efforts. Most existing methods assume that human faces have been extracted from a static image or from a sequence of images and focus on the recognition algorithms. However, face detection from a single image or a image sequence is a very challenging task and is no easier than face recognition. Face detection is considerably difficult because it involves locating faces with no prior knowledge about their scales, locations, orientations (up-right, rotated) with or without occlusions, with different poses (frontal, profile). Facial expressions and lighting conditions also change the overall appearances of faces, thereby making it difficult to detect them. Furthermore, the appearance of human faces in an image depends on the poses of humans and the viewpoints of the acquisition devices. The challenges associated with face detection problem can be attributed to the following factors:

- Poses : Faces can appear in different poses (frontal, 45 degree, profile, upside down) that make their appearances vary in the images. Some poses may occlude facial features such as eyes and noses.
- Presence or absence of common structural features : Human faces have different facial features such as beard, mustache or glasses. Moreover, such features have drastic differences in appearances because of locations and sizes.
- Facial expressions : The appearance human faces are affected by facial expressions.
- Occlusions : Faces may be occluded by other objects. In an image with a group of people, some faces may partially occlude other faces.

• Imaging formulation conditions : Imaging factors affect the resulting appearance of human faces when the image is formulated, thereby causing problems such as scale, orientation, viewpoint, and lighting conditions.

It is clear that these unknown factors make face detection a very interesting and challenging problem.

We now give a definition of face detection: Given an arbitrary image or image sequence, which can come from a digitized image or a scanned photograph, the goal of face detection is to determine whether or not there is any human face in the image, and, if present, return its location and spatial extent.

There are many closely related problems to face detection. *Face localization* [1] aims to localize a single face in an image. In other words, this is a simplified face detection problem with the assumption that the input contains only one face. *Facial Feature Detection* [2] is a problem to detect facial features in a face with the assumption that there is only one face in an image. *Face recognition* or *face identification* [3] [4] [5]. is to compare an input image against a library and report a match, if any. *Face authentication* [6] is to verify the claim of the identify of an input image and *face tracking* is to find the locations of an face or faces in an image sequence in real time. *Facial expression recognition* concerns with identifying the expression of humans from their facial expressions [7] [8] [9] [10]. It is evident that face detection is the first step for any automated system to solve the above problems.

It is worth noticing that many existing literature use the term "face detection" in their papers, but the methods and the experimental results only show that a single face is localized in an input image. In this paper, we differentiate face detection from face localization since the later is a simplified problem of the former problem.

In recent years, many methods have been proposed to detect human faces in a single image [11] [12] [13] or a sequence of images [14] based on gray scale [11] [12] [13] or color [15] [14]. Although many exciting and interesting results have resulted from the ongoing research efforts on face detection, we have not seen any survey on this particular topic. A survey of early methods before 1991 for face recognition and face detection can be found in survey on face recognition [5]. Another excellent survey on face recognition and several face detection methods before 1994 is in [4].

This paper aims to give a comprehensive and critical survey on current face detection methods. In Section II we give a detailed review of techniques to detect faces in still images. Section III presents a discussion on face detection from image sequences. We conclude this paper with discussion of several promising directions for face detection in Section IV.

II. Face Detection From Still Images

In this section, we review existing techniques to detect faces from a still image. Most face detection methods can be categorized into:

1. Knowledge-based top-down method
2. Feature invariant approach

3. Integration of multiple feature
4. Appearance-based approach

We first discuss the representative methods in each category, then address on the problems with these approaches. Possible improvements are discussed in Section IV.

A. Knowledge-based top-down model

In this approach, face detection methods are developed based on the rules derived from our knowledge of human faces. It is easy to come up with simple rules to describe the features of a face and their relationships. For example, a face usually appears with two eyes that are symmetric to each other, a nose and a mouth. The relationships between these features can be represented by their relative distances and positions. Faces are then detected by applying these codes rules to find features. Face candidates are detected after applying all the rules. A verification process is optionally used to increase avoid false detections.

One problem with this approach is that it is difficult to translate our knowledge about faces into rules effectively. If the rules are detailed (i.e. strict), it may fail to detect faces that do not pass all the rules. If the rules are too general, it may give many false positives. Moreover, it is difficult to extend this approach to multiple views since it is impossible to enumerate all the possible cases. On the other hand, heuristics about human faces works well in detecting frontal human faces in uncluttered scenes.

Yang and Huang [16] use a hierarchical knowledge-based method to detect faces. Their system consists of three levels of rules. At the highest level, all possible candidates of human faces are found by scanning over the input image based some rules. The rules for higher level are general descriptions of what a face looks like, and the rules for lower levels are established based on details of facial features. In this method, the rules of levels 1 and 2 are established using mosaic images which are constructed Simple coded rules such as "the eyes should be darker than the rest of the face" are used to locate face candidates. by decreasing the resolution of the original image. A mosaic image consists of cells where each cell is a window of $n \times n$ pixels. Level 1 (highest level) uses coarsest mosaics to search face candidates which are further processed in level 2 uses finer mosaics. In level 3, local histogram equalization is performed on the the face candidates received from level 2, followed by edge detections. These candidates are then examined at next level by another set of rules where all the candidates are classified as face or non-face based a edge-detection method to search for facial features such as eye and mouth regions. Although this method does not generate high detection rate, the ideas of mosaicing, multiple levels, and rules to guide face searches have been used in later works on face detection.

Lanitis, Taylor and Cootes describe a face recognition method using shape and grey-level parameters [17] in that the face is located using active shape model search, thereby generating shape parameters. The face patch is then deformed to the average shape and they grey level parameters are extracted. The shape and grey level parameters are used together for classification.

In [18], a rule based method that extends [16] is developed. In this method, the horizontal profile of an input image is obtained by averaging over all pixel intensities in each image column. Two local minima, determined by detecting abrupt changes in intensities, correspond to the left and right side of the head. Similarly, the vertical profile is obtained and the local minima are determined for the locations of mouth lips, nose tip and eyes. These detected features constitute a facial candidate. Subsequently, eyebrow/eyes, nostrils/nose and the mouth detection rules are used to validate these candidates. Their system reports a success rate of 86.5%.

B. Bottom-up feature model

Contrast to knowledge-based top-down model, researchers have been trying to find feature invariants of faces for detection. The underlying assumption is based on the observation that human can detect faces effortless because facial features always exist in the image, regardless of poses, viewpoints or lighting conditions. The assumptions, if hold, are also the advantages of this approach since faces of different poses can be detected. On the other hand, the major problem with feature-based algorithm is that the image features can be badly corrupted due to illumination, noise or occlusion. Feature boundaries can be weakened by illumination, and shadows can cause numerous strong edges that render perceptual grouping algorithms useless.

B.1 Facial features

Govindaraju et. al. [19] [20] [21] present a two stage face detection method in that face hypothesis is generated and tested. A face model is defined in terms of features based on the edges of the frontal view of a face. The features of the model describe the curves of the left side, the hair-line and the right side of a face contour in the frontal view. They use Marr-Hildreth edge operator to obtain an edge-image of an input image where the edges are further processed by a thinning algorithm. A filtering processed is used to remove objects whose contours are unlikely to be parts of a human face. Pairs of fragmented contours are linked based on the proximity of fragments and their relative orientations. Corners in the contours are detected to segment them into feature curves. These feature curves are then labeled by examing their geometric properties and the relative positions in the neighborhood. Pairs of feature curves are joined by edges if their attributes are compatible (i.e. these features have a possibility to be parts of the same face). The ratios of the pairs of features forming an edge is compared with the golden ratio and a cost is assigned to the edge. If the cost of a group of three feature curves (with different labels) is low, the group becomes a hypothesis. Collateral information, which indicates the number of persons in the image, is obtained from the caption of the input image to select the best hypotheses. Their system reports a success rate about 70% based on a test of 50 photographs.

In [22] facial features are extracted first. Constellations are then formed from the pool of candidates and the most face-like constellation is determined. Finding the best constellation is formulated as a random

graph matching problem in which the nodes of the graph correspond to features on the face and the arcs represent the distances between the different features. Computation of the matching process is reduced by using a controlled search. The basic idea is that given the positions of several features, we can estimate the other features and the covariance of the estimates. Ranking of constellations is based on a probability density that a constellation corresponds to a face versus the probability it was generated by an alternative mechanism.

Yow et. al. [23] propose a feature-based method which uses a large amount of evidence from the visual image and contextual evidence. The first stage operates on the raw image data and produce a list of interest points from the image, indicating likely location of facial features. The second stage examines these interest point, group them based on Gestalt principles and label them accordingly to knowledge acquired from training data. The labeled features are further grouped based on model knowledge of where they should occur with respect to each other. Each facial feature and grouping is then evaluated using a Bayesian network.

Sinha [24] uses a small set of spatial image invariants to describe the space of face patterns. The underlying observation is that the local structure of brightness distribution of a human face remains largely unchanged although illumination and other changes can significantly alter brightness levels at different parts of the face. His scheme encodes these observed brightness regularities as a ratio template and uses it for pattern match. A ratio template is a coarse spatial template of a face with a few appropriately chosen subregions that roughly correspond to key facial features such as eyes, cheeks and foreheads. The brightness constraints between facial parts are captured by an appropriate set of pairwise brighter-darker relationships between corresponding subregions. A face is detected if an image satisfies all the pairwise brighter-darker constraints.

In [25] a morphology-based technique is developed to extract eye-analogue segments for face detection. First, morphological operations such as closing, clipped difference, and or are applied to locate eye-analogue pixels in the original image. Then, a labeling process is performed to generate the eye-analogue segments. These segments are used to guide the search for potential face regions. Potential face regions are obtained if a possible geometrical combinations of eyes, nose, eyebrows and mouth exists. These candidates are further verified by a neural network similar to [26]. Their experiments demonstrate an approximately 94% accuracy rate.

B.2 Texture

In [27], a method based on the feature parameters of space grey-level dependence matrix (SGLD) [28] is developed. A face-texture model composed by a set of inequalities is derived and a face area is defined as such a region where these inequalities hold. Color information is also incorporated with the face-texture model for face detection.

B.3 Color

Human skin color has been used and proved to be an effective feature in many applications from human face detection to hand tracking. Although different people have different color in appearance, several studies have showed the major difference between lies in intensity rather than color itself [15] [14]. To build a skin color model, we can use a single Gaussian or a mixture of Gaussians to model the skin color distribution.

Most recently, several modular systems using a combination of shape analysis, color segmentation and motion information for locating or tracking heads and faces in an image sequence have been developed [15] [29] [14] [30] [31]. See II-B.4 for more detail.

B.4 Multiple Features

Recently, methods that combine several facial features are proposed to detect faces. Most methods utilize skin color, size and shape global features to find face candidates and then verify these candidates using local, detailed features such as eye bows, hairs. The general approach begins with detection of skin-like regions, since it has been shown that human skin color fall in a small manifold in color space. Several color spaces have been utilized to find the manifold of skin colors including normalized RGB [32], [33], HSV [31], CIE XYZ [34], to CIE LUV [29] color space. Next, skin-like pixels are grouped together using component analysis or clustering. If the shape of a connected region is similar to ellipses or oval shape, then it becomes a face candidate. Finally, several verification techniques based on local features are used to determine whether the detected region is a face or not.

In [34] [35], Yachida et. al. build a skin color distribution function in CIE XYZ color space. They assume skin and hair regions appear in each face. Each pixel is classified as hair, face, hair/face, hair/background cell based on the distribution and luminance, thereby generating skin-like and hair-like regions. Three models (one frontal and two sideview), which describe relationships of these cells, are used to compare with the extracted skin-like and hair-like regions. If the match of cells of the extracted against any model is above a threshold, the detected region becomes a face candidate. Eye-eyebrow and nose-mouth features are extracted from a face candidate, using horizontal edges, to verify whether it is a face or not. The disadvantage of this approach is that it has many assumptions about what is a face. It is not clear that this method can detect human faces of different poses or hairs.

Sobottka and Pitas [31] propose a method for face localization and facial feature extraction based on shape and color information. First, color segmentation in HSV space is performed to locate skin-like regions. Connected components are then determined by a region growing algorithm at a coarse resolution of the segmented image. For each connected component, the best-fit ellipse is computed using the moments. The ellipses that are good approximations of connected components are selected and considered as face candidates. These candidates are subsequently determined by searching for facial features inside of the connected components. Facial features, such as eyes and mouths, are extracted based

on the observation that these features have lower intensity than the rest regions of the face. Contrast to the methods that detect skin-like regions based on pixels, [29] uses a segmentation scheme to better extract facial regions in complex background.

The blob representation, developed by Pentland and Kauth, is a way to extract a compact, structurally meaningful description of multispectral satellite imagery [36] [37]. The feature vectors at each pixel are formed by adding spatial coordinates to the spectral (or textural) components of the imagery and then clustered so that image properties such as color and spatial similarity form coherent connected regions, or "blobs." This method has been applied to detect faces where each feature vector consists of spatial position and normalized chromaticity ,i.e. $X = (x, y, \frac{r}{r+g+b}, \frac{g}{r+g+b})$ [32], [33]. A connectivity algorithm is then used to grow blobs and the resulting skin blob whose size and shape is closest to the canonical face size and shape is considered as a face.

C. Template Matching

In template matching, a standard pattern of a human face (usually frontal) is stored first. Given an input image, the correlation values with several sizes of the standard pattern are calculated for the face contour, eyes, nose and mouth independently. From these correlation values of the face portions, the final determination is done for the existence of a face. This approach has the advantage of being simplistic. However, it has proved to be inadequate for face detection since it cannot deal with variation of faces in scale, pose and shape. Multiresolution, multiscale and deformable templates have been proposed to achieve scale and shape invariances.

Early attempt to detect frontal faces in photographs is reported by Sakai, Nagao and Fujibayashi [38]. They use several subtemplates which correspond to facial features such as head line, eyes, mouth to model a human face. Lines in the input image are extracted based on greatest gradient descent and then matched against the subtemplates. When a matching is obtained on the pixel (i, j) of th input image, the pixels within a region $(i + D, j + D)$ are also searched to find the extent of a face.

Yuille [39] uses deformable templates to model facial features that fit an a priori elastic model to the elastic features of the face. In this approach, facial features of interest are described by parameterized templates. An energy function is defined to link edges, peaks and valleys in the input image to corresponding parameters in the template. The best fit of the elastic model is found by energy minimization which is achieved by altering the parameter values. One drawback of this approach is that it requires the deformable template be placed in proximity of the object of interest before the process begins.

D. Appearance-based methods

D.1 Neural Network

The basic assumption of this approach is based on the observation that faces are a highly structured class of image patterns and can be detected by examine only local image information within a spatially

well-defined boundary. In other words, it is possible to train a system to capture complex face patterns from the examples which may otherwise be difficult to parameterize by other techniques. The detection paradigm works by testing candidate image locations for local patterns that appear like faces. The crux of this approach is a classification procedure that determines whether or not a given local image pattern is a face and the classification problem is posed as learning to identify faces from training examples of face and nonface patterns. The lure of such a approach is we can build a system without defining "features" of a face manually and such approach can potentially be extended to detect other spatially well-defined objects since the system learns from the appearance of the training examples without any prior knowledge. In fact, such example-based approach has been applied to recognize 3D objects [40], [41], prototype learning [42], pedestrian detection [43], etc. However, the disadvantage of this approach is that it is difficult to collect enough training examples for such tasks. Moreover, there is no theory about the number of examples needed to train the system to achieve a certain detect rate.

Sung and Poggio develop a clustering and distribution-based system for face detection [44], [45], [12]. Their method extends Sinha's [24] invariance-based system by using a learning approach, rather than manually encoding the invariants. Their system consists of two major components, a distribution-based face model and a multilayer perceptron classifier. Each face and nonface example is first normalized and processed to a 19×19 pixel patterns. In other words, each face and nonface example can be considered as a multidimensional vector. Next, they use six face clusters and six nonface clusters to piecewise approximate the multidimensional distribution of these face and nonface patterns. Each face pattern cluster is a multidimensional Gaussian function with a centroid location and a covariance matrix that describes the local data distribution around the centroid. Two distance metrics measure the distance of an input image to the prototype clusters are used. The first distance component is a normalized Mahalanobis distance between the test pattern and the cluster centroid, measured within a lower-dimensional subspace spanned by the cluster's 75 largest eigenvectors. It ignores pattern differences in the cluster's smaller eigenvector directions, because the eigenvalues that can be recovered in these directions may be significantly inaccurate due to insufficient data. The second distance component is a standard Euclidean distance between the test pattern and its projection in the 75-dimensional subspace. This distance component accounts for pattern differences not captured by the first distance component. Each test pattern is represented by a feature vector of 12 two-value distance measurements. The last step in their system is to use a multilayer perceptron network to classify face window patterns from nonface patterns using two distances to each of the clusters. The train their classifier on feature distance vectors from a database of 47,316 window patterns. There are 4,150 positive examples of face patterns and the rest are nonface patterns. The network is trained with a standard backpropagation learning algorithm. Note that it is easy to get a representative sample of images which contain faces, but much more difficult to get a representative sample of those which do not. This problem is avoided a bootstrap method that electively adds images to the training set as training progress. They start with a small set of nonface examples in the training example

and train the MLP classifier with the current database of examples. Then, they run the face detector on a sequence of random images and collect all the nonface patterns that the current system wrongly classifies as faces. These nonface patterns are then added to the training database as new nonface examples. This bootstrap method avoids the problem of collecting representative sample of nonface examples.

Several methods that use neural network to detect human faces have been proposed. The main feature of this method is to train a neural network to detect the presence or absence of faces by scanning over all possible positions at different scales. One of the earliest neural network literature on face detection is by Soulie, Vinnet and Lamy [46] in that they apply a time-delay neural network (TDNN) [47] to detect faces. The implemented principle is to scan an input image with a TDNN retina of 20×25 pixels. For each position in the image, the network tells faces from the background. To cope size variation, they use a multiresolution decomposition on the input image using wavelet transforms. Their method reports false negative rate of 2.7%, false positive rate of 0.5% from a test of 120 images. In [48], Vaillant, Monrocq and Le Cun use neural networks to detect human faces in images. Examples of face and nonface images of $20x20$ pixels are first created. One neural network is trained to find rough localizations of human faces at some scale. Another network is trained to determine the exact position of faces at some scale. Given an image, ares which may contain faces can be selected as face candidates by the first network. These candidates are verified by the second network. Burel and Carel [49] propose an example-based method for face detection using neural networks. The large number of training examples of faces and backgrounds are compressed into fewer examples using a Kohonen algorithm, i.e. taking profit of the vector quantization properties of this algorithm. A multi-layer perceptron (MLP) is used to learn these examples for face/background classification. The detection phase consists of scanning each image at various resolution. For each location and size of the scanning window, the window content is normalized to a standard size, and its mean and variance to reduce of the intensity to reduce the effects of lighting conditions. Each normalized window is then propagated through a MLP for classification.

Perhaps the most significant neural network based approach for face detection is developed by Rowley, Baluja and Kanade [26], [50], [11]. Their method has some similarities to the system by Sung and Poggio [45] and [49]. A multilayer neural network is used to learn the face and nonface patterns from face/nonface images (i.e. the intensities and spatial relationships of pixels) while Sung and Poggio [45] use a neural network to find a discriminant function to classify face and nonfaces using distance measures. They also use multiple neural networks and several arbitration methods to improve performance while Burel et. al. [49] use a single network and Vaillant et. al [48] use two networks for classification. The first component of their method is a filter that receives a $20x20$ pixel region of an image and greater an output ranging from -1 to 1, signifying the presence of a face. To detect faces anywhere in an image, the filter is applied at every location in the image. To detect faces larger than the window size, the input image is repeated reduced in size by subsampling and the filter is applied at each size. To train the neural networks to be accurate filters, a large number of face and nonface images are collected. Nearly 1,050 face examples are

gathered and these images contain faces of various sizes, orientations, positions and intensities. The eyes, tip of nose and corners and center of the moth of each face is labeled manually. These points are used to normalize each face to the same scale, orientation, and position. The second component is to merge overlapping detection and arbitrate between the outputs of multiple networks. They use simple arbitration schemes such as ANDing, ORing and voting and report improved performance. Rowley [26] reports several systems with different arbitration are less computationally expensive than Sung and Poggio's system , and has higher detect rate based on a test set of 24 images containing 144 faces. One limitation of these two systems is that they can only detect upright, frontal faces. Recently, Rowley [51] extend their system to detect rotated faces using router network which process each input window to determine its orientation and then rotates the window to an upright position before presenting to the neural networks as described above. However, the new system has a lower detection rate on upright faces than the upright detector. Nevertheless, the system is able to detect 76.9% of faces over two large test sets with a small number of false positives (those who are nonface objects but classified as faces by the trained network).

D.2 Support Vector Machine

Support Vector Machine (SVM) has also been applied to face detection by Osuna, Freud and Girosi [13] in that they develop a decomposition algorithm that guarantees global optimality. SVM can be considered as a new paradigm to train polynomial, neural networks, or radial basis function (RBF). While most of the techniques used to train the classifiers (such as neural networks and RBF) are based on the idea of minimizing the training error, i.e. *empirical risk*, SVM operate on another induction principle, called *structural risk minimization*, which minimizes an upper bound on the generalization error. Training a SVM is equivalent to solving a linearly constrained quadratic programming problem. Based on two test sets of 10,000,000 pattern windows, their system performs slightly better and is approximately 30 times faster than the system by Sung and Poggio's [44].

D.3 Eigenface Approach

A probabilistic visual learning method, based on density estimation in high-dimensional space using an eigenspace decomposition, is developed by Moghaddam and Pentland [1]. In principal component analysis, the largest-eigenvalue eigenvectors are identified and selected as principal components to form a subspace. These principal components preserve the major linear correlations in the data and discard the minor ones. In contrast, the form an orthogonal decomposition of the vector space into two mutually exclusive and complementary subspaces: the principal subspace (or feature space) and its orthogonal complement. Therefore, the target density is decomposed into two components: the density in the principal subspace (spanned by the principal components) and its orthogonal complement (which is discarded in standard principal component analysis). A multivariate Gaussian and a mixture of Gaussians are used to learn the statistics of the local features of a face. These probability densities are then used for object detection based

on a maximum likelihood estimation. The proposed method has been applied for face localization, coding and recognition. Compared with the classic eigenface approach [3], the proposed method shows better performance in face recognition. In terms of face detection, their technique has only been demonstrated on localization (i.e. assuming an input image contains only one face).

D.4 Statistical Approach

In [52], Schneiderman and Kanade describe an probabilistic model for object recognition based primarily on local appearance, which differs significantly from appearance-based method that emphasize global appearance. This approach contrasts to the methods in [45] [13] which model the full, global extent of the object, in this case human face, at once. The reason that they emphasize on local appearance is that some local patterns on the object are more unique than others. For human faces, the intensity patterns around the eyes of a human face are much more unique than the pattern found on the cheeks. To represent the uniqueness of local appearance, the statistics of local appearance need to be modeled. The reason that they use a functional form of the posterior probability function is to capture the joint statistics of local appearance and position on the object as well as the statistics of local appearance. This probabilistic model of local appearance and spatial relationships shows comparable performance with [11].

In [53], a higher order statistics based clustering algorithm and a hidden Markov model (HMM) scheme are proposed for face detection. In the first method, the unknown distributions of face and face-like manifolds are modeled using higher order statistics. The conjecture is that face pattern distribution is unlikely to be governed by multidimensional Gaussian functions as used in [45]. A multilayer perceptron is used for classification, as in [45]. The second method uses a HMM to learn the face to nonface and nonface to face transitions. The observation sequence is generated in the transform domain by comparing each masked subimage with a knowledge-base consisting of 6 face and 6 face-like centroids, similar to the distance metric used in [45].

D.5 Information Theory Approach

Lew applies Kullback relative information for face detection by associating hypothesis' H_1 to the event that the template is a face and H_0 to the event that the template is not a face [54]. A face training database consisting of 9 views of 100 individuals is used to estimate the face distribution. The nonface probability density function is estimated from a set of 143,000nonface templates. From the training sets of face and nonface, the most informative pixels (MIP) are selected to maximize the Kullback relative information (i.e. to give the maximum class separation). It turns out the MIP distribution focuses on the eye and mouth regions and avoids avoids the nose area. The MIP are then used to obtain linear features for classification and representation using the method of Fukunaga and Koontz [55]. To detect faces, a window is passed over the input image and the DFFS (distance from face space) as defined in [56] is calculated. If the DFFS to the face cluster is lower than the DFFS to the nonface cluster, a face is

assumed to exist within the window.

Kullback relative information is also employed by Colmenarez and Huang to maximize the discrimination between positive and negative examples of faces [57]. They use a family of discrete Markov processes to model the face and background patterns and estimate the probability model. The learning process is converted into an optimization problem to select the Markov process that maximizes the information based discrimination between the two classes. Face detection is carried by computing the likelihood ratio using the trained probability model.

Qian and Huang [58] presents a method that employs the strategies of both view-based and model-based methods. First, a visual attention algorithm which uses high level domain knowledge is applied to reduce the search space. This is achieved by selecting image areas in which targets may appear based on the region maps generated by a region detection algorithm (water shed method). Within the attention regions, they apply a detection algorithm that combines template matching methods with feature based method via hierarchical Markov random field (MRF) and maximum a posterior (MAP) estimation. Their approach assumes that the probability measure of the occurrence of some targets and their features is a hierarchical MRF and then seek the maximum probability of the occurrence of the targets given an input image using MAP estimation.

Induction learning is also used to detect human faces. In [59], Quinlan's C4.5 algorithm is used to induce a decision tree from the positive and negative examples. Each example is a feature vector of thirty feature values (entropy, mean and standard deviation) to represent a face. From these examples, C4.5 then builds a classifier as a decision tree whose leaves indicate class identity and nodes specify some tests to perform on a single attribute to branch the current unclassified inputs. Once the decision tree has been built, a face is detected if it falls into a face class.

III. Detecting and Tracking Human Faces from Image Sequence

Hunke and Waibel [60] develop a neural network based system that manipulates camera orientation and zoom to keep a person's face located at all times. In the first step, all areas containing face-like colors are extracted using a face color classifier which is able to classify colors as face colors or background colors. The second step detects objects using the combined color and movement information that is detected by determining the difference of corresponding pixels in two following images. The combined color and movement information is fed into two multilayer perceptron to determine the position and size of the detected face. This system has later been extended to track several human faces and facial features [14], [61].

Jebara and Pentland proposes a tracking system based on skin color, symmetry, 3D model and eigenface [62]. One feature of their system is that a 3D range data model of an average human face is formed off-line from a database of range data. They system is initialized by using skin classification, symmetry operation, 3D warping and eignefaces to find a face. They use Gaussian mixture to model the skin color

distribution where the parameters are estimated by the EM algorithm. Any pixel is labeled as skin if the probability is above a threshold, followed by a connected component analysis to determine regions of skin pixels in the image. The dark symmetry transform [63] is used to determine the possible positions of eyes. Additionally, the positions of nose and mouth are determined by some heuristics. These four points are used to align them via a 3D mapping and a vertical stretch into a standard pose. The 2D image's intensity data is then mapped onto the 3D structure and then rotated into a normalized frontal view, followed by a projection to form a segmented, mug shot image of the face. The mug shot is projected into an eigenspace of color face images to measure the "faceness" of the image. The detected four facial features and motion information is used for face tracking.

IV. Conclusion

Although significant progress has been made in the last two decades, we believe that a robust face detection system should address the following problems:

- Utilizing mixtures of different classifiers
- Effects of lighting conditions
- Detecting faces of Different poses
- Empirical experiments for comparison
- A database for benchmarking

Recently, several approaches that uses a mixture of Gaussians [12], [64] or a mixture of subspaces [65] to recognize human faces. The argument is that human face is a high manifold that can not be appropriately represented by a single cluster. Moreover, most existing face detection use faces with neutral expressions and few, if any, facial features. To build a detection system that can identify faces with different expressions and facial features, it is natural to expect that different clusters will be best represented by different subspaces.

Several promising classification methods have been proposed and applied in different domains. Support vector machine has been applied to face detection [13], pedestrian detection [43] and 3D object recognition [41]. Factor analysis has been applied to face recognition [65]. Other methods such as sensible principal component analysis [66], hierarchical mixtures of experts [67], mixtures of probabilistic principal component analysis [68], factor analysis with wake-sleep learning [69] [70], and linear generative methods [66] have been proposed. These methods can potentially be applied in face detection and improve current methods.

The effects of lighting condition have attracted more attention in building robust face recognition [71] [72] [73]. The approach in [71] are based on two observations. First, the images of a Lambertian surface, taken from a fixed viewpoint but under varying illumination, lie in a 3D linear subspace of the high-dimensional image space. Second, the first observation is not exactly correct because of the effects shadowing, specularities and facial expressions. Therefore, certain regions of the face may have variability

from image to image that often deviates significantly from the linear subspace, thereby making the first observation less reliable. To cope with these problems, they propose a method called Fisherface that combines principal component analysis (PCA) and Fisher linear discriminant (FLD) for face recognition. PCA is applied on the input images to reduce the dimensions of the feature space. This is followed by FLD to further reduce the dimensions for classification. Zhao, Chellappa and Krishnaswamy also apply similar approach for face recognition and demonstrate that linear discriminant analysis (LDA) with PCA perform better than pure LDA and PCA in face recognition [74].

Although there have been numerous existing methods to detect human faces, it is not clear which method perform better in which situation. Therefore, it is important to conduct empirical experiments on these methods, thereby gaining more insight to improve existing methods. Towards this end, it is equally important to build a database for benchmarking on face detection. Although there exist several databases for experiments on face recognition (such as FERET and Olivetti), there is no major database for face detection. Recently, a database has been built for face detection [75]. It is important that more efforts be put in empirical experiments and building database for benchmarking.

Face detection is a very challenging and interesting problem. In the meanwhile, it is practical and important since it is the first step in any fully automated face recognition system. With the growing interest in human-computer interactions, it is important to investigate fast and robust method. This paper gives a critical review of existing methods and discuss several promising directions for future research. We believe that a good face detection system should utilize prowess of different classifiers. and address the above problems.

References

[1] B. Moghaddam and A. Pentland, "Probabilistic visual learning for object recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 696–710, 1997.

[2] I. Craw, D. Tock, and A. Bennett, "Finding face features," in *Proceedings of European Conference on Computer Vision*, 1992, pp. 92–96.

[3] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[4] R. Chellappa, C. L. Wilson, and S. Sirohey, "Human and machine recognition of faces: A survey," *Proceedings of the IEEE*, vol. 83, no. 5, pp. 705–740, 1995.

[5] A. Samal and P. Iyengar, "Automatic recognition and analysis of human faces and facial expressions: A survey," *Pattern Recognition*, vol. 25, no. 1, pp. 65–77, 1992.

[6] C. Kotropoulos, A. Tefas, and I. Pitas, "Frontal face authentication using variants of dynamic link matching based on mathematical morphology," in *Proceedings of International Conference on Image Processing*, 1998, pp. 122–126.

[7] K. Matsuno, C.-W. Lee, S. Kimura, and S. Tsuji, "Automatic recognition of human facial expressions," in *Proceedings of the Fifth International Conference on Computer Vision*, 1995, pp. 352–359.

[8] I. A. Essa and A. Pentland, "Facial expression recognition using a dynamic model and motion energy," in *Proceedings of the Fifth International Conference on Computer Vision*, 1995, pp. ?–?

[9] Marian Stewart Bartlett, Paul A. Viola, T. J. Sejnowski, B. Golomb, J. Larsen, J. C. Hager, and P. Ekman, "Classifying facial action," in *Advances in Neural Information Processing Systems*, 1996, vol. 8, pp. 823–829.

[10] J. J.-J. Lien, T. Kanade, J. F. Cohn, and C.-C. Li, "Subtly different facial expression recognition and expression

intensity estimation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1998, pp. 853–859.

[11] Henry Rowley, Shumeet Baluja, and Takeo Kanade, "Neural network-based face detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 1, pp. 23–38, 1998.

[12] Kah-Kay Sung and Tomaso Poggio, "Example-based learning for view-based human face detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 1, pp. 39–51, 1998.

[13] E. Osuna, R. Freund, and F. Girosi, "Training support vector machines: an application to face detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1997, pp. 130–136.

[14] J. Yang and A. Waibel, "A real-time face tracker," in *Proceedings of the Third Workshop on Applications of Computer Vision*, 1996, pp. 142–147.

[15] H. P. Graf, E. Cosatto, D. Gibbon, M. Kocheisen, and E. Petajan, "Multimodal system for locating heads and faces," in *Proceedings of the Second IEEE International Conference on Automatic Face and Gesture Recognition*, 1996, pp. 88–93.

[16] G. Yang and T. S. Huang, "Human face detection in complex background," *Pattern Recognition*, vol. 27, no. 1, pp. 53–63, 1994.

[17] A. Lanitis, C. J. Taylor, and T. F. Cootes, "An automatic face identification system using flexible appearance models," *Image and Vision Computing*, vol. 13, no. 5, pp. 393–401, 1995.

[18] C. Kotropoulos and I. Pitas, "Rule-based face detection in frontal views," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, 1997, vol. 4, pp. 21–24.

[19] V. Govindaraju, D. B. Sher, and R. K. Srihari, "Locating human face in newspaper photographs," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1989, pp. 549–554.

[20] V. Govindaraju, S. N. Srihari, and D. B. Sher, "A computational model for face location," in *Proceedings of the Third International Conference on Computer Vision*, 1990, pp. 718–721.

[21] V. Govindaraju, "Locating human faces in photographs," *International Journal of Computer Vision*, vol. 19, no. 2, pp. 129–146, 1996.

[22] T.K. Leung, M.C. Burl, and P. Perona, "Finding faces in cluttered scenes using random labeled graph matching," in *Proceedings of the Fifth International Conference on Computer Vision*, 1995, pp. 637–644.

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

[24] P. Sinha, "Object recognition via image invariants: a case study," *Investigative Ophthalmology and Visual Science*, vol. 35, no. 4, pp. 1735–1740, 1994.

[25] C.-C. Han, H.-Y. M. Liao, K.-C. Yu, and L.-H. Chen, "Fast face detection via morphology-based pre-processing," in *Proceedings of the Ninth International Conference on Image Analysis and Processing*, 1998, pp. 469–476.

[26] H. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1996, pp. 203–208.

[27] Y. Dai and Y. Nakano, "Extraction for facial images from complex background using color information and sgld matrices," in *Proceedings of the International Workshop on Automatic Face and Gesture Recognition*, 1995, pp. 238–242.

[28] R. M. Haralick, "Textural features for image classification," *IEEE Trans. System, Man and Cybernetics*, vol. 3, no. 6, pp. 610–621, 1973.

[29] Ming-Hsuan Yang and Narendra Ahuja, "Detection human faces in color images," in *Proceedings of the International Conference on Image Processing*, 1998, pp. 127–130.

[30] S. McKenna, Y. Raja, and S. Gong, "Tracking colour objects using adaptive mixture models," *Image and Vision Computing*, 1998.

[31] J. Sobottka and I. Pitas, "Segmentation and tracking of faces in color images," in *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*, 1996, pp. 236–241.

[32] T. Starner and A. Pentland, "Real-time asl recognition from video using hmm's," Tech. Rep. Technical Report 375, Media Lab, MIT, 1996.

[33] N. Oliver, A. Pentland, and F. Berard, "Later: Lips and face real time tracker," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1997, pp. ?–?

[34] Q. Chen, H. Wu, and M. Yachida, "Face detection by fuzzy matching," in *Proceedings of the International Conference on Computer Vision*, 1995, pp. 591–596.

[35] Haiyuan Wu, T. Yokoyama, D. Pramadihanto, and M. Yachida, "Face and facial feature extraction from color image," in *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*, 1996, pp. 345–350.

[36] A. Pentland, "Classification by clustering," in *Proceedings of the IEEE Symposium on Machine Processing and Remotely Sensed Data*, 1976, pp. ?–?

[37] R. Kauth, A. Pentland, and G. Thomas, "Blob: An unsupervised clustering approach to spatial preprocessing of mss imagery," in *Proceedings of the Eleventh International Symposium on Remote Sensing of the Environment*, 1977, pp. ?–?

[38] T. Sakai, M. Nagao, and S. Fujibayashi, "Line extraction and pattern detection in a photograph," *Pattern Recognition*, vol. 1, pp. 233–248, 1969.

[39] A. Yuille, P. Hallinan, and D. Cohen, "Feature extraction from faces using deformable templates," *International Journal of Computer Vision*, vol. 8, no. 2, pp. 99–111, 1992.

[40] H. Murase and S. Nayar, "Visual learning and recognition of 3d objects from appearance," *International Journal of Computer Vision*, vol. 14, no. 1, pp. 5–24, 1995.

[41] M. Pontil and A. Verri, "Support vector machines for 3d object recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 6, pp. 637–646, 1998.

[42] P. Benoit, N. Ahuja, and N. Srinivasa, "Learning multiscale image models of 2d object classes," in *Proceedings of the Third Asian Conference on Computer Vision*, 1998, pp. 323–331.

[43] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio, "Pedestrian detection using wavelet templates," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1997, pp. 193–199.

[44] K. Sung and T. Poggio, "Example-based learning for view-based human face detection," Tech. Rep. AIM-1521, MIT AI Lab, 1994.

[45] K.-K. Sung, *Learning and Example Selection for Object and Pattern Detection*, Ph.D. thesis, MIT AI Lab, 1996.

[46] F. Soulie, E. Viennet, and B. Lamy, "Multi-modular neural network architectures: Pattern recognition applications in optical character recognition and human face recognition," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 7, no. 4, pp. 721–755, 1993.

[47] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 37, no. 3, pp. 328–?, 1989.

[48] R. Vaillant, C. Monrocq, and Y. Le Cun, "An original approach for the localisation of objects in images," in *IEE Proceedings of Vision, Image and Signal Processing*, 1994, vol. 141, pp. 245–250.

[49] G. Burel and D. Carel, "Detection and localization of faces on digital images," *Pattern Recognition Letters*, vol. 15, no. ?, pp. 963–967, 1994.

[50] H. Rowley, S. Baluja, and T. Kanade, "Human face detection in visual scenes," in *Advances in Neural Information Processing Systems*, 1996, vol. 8, pp. 875–881.

[51] H. Rowley, S. Baluja, and T. Kanade, "Rotation invariant neural network-based face detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1998, pp. 38–44.

[52] H. Schneiderman and T. Kanade, "Probabilistic modeling of local appearance and spatial relationships for object recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1998, pp. 45–51.

[53] A. N. Rajagopalan, K. S. Kumar, J. Karlekar, R. Manivasakan, and M. M Patil, "Finding faces in photographs," in *Proceedings of the Sixth International Conference on Computer Vision*, 1998?, pp. ?–?

[54] M. S. Lew, "Information theoretic view-based and modular face detection," in *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*, 1996, pp. 198–203.

[55] F. Fukunaga and W. Koontz, "Applications of the karhunen-loeve expansion to feature selection and ordering," *IEEE Trans. Neural Computers*, vol. 19, no. ?, pp. 917–923, 1970.

[56] A. Pentland, B. Moghaddam, and T. Starner, "View-based and modular eigenspaces for face recognition," in *Proceedings of the Fourth International Conference on Computer Vision*, 1994, pp. ?–?

[57] A. J. Colmenarez and T. S. Huang, "Face detection with information-based maximum discrimination," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1997, pp. 782–787.

[58] R. J. Qian and T. S. Huang, "Object detection using hierarchical mrf and map estimation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1997, pp. 186–192.

[59] J. Huang, S. Gutta, and H. Wechsler, "Detection of human faces using decision trees," in *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*, 1996, pp. 248–252.

[60] H. M. Hunke, "Locating and tracking of human faces with neural networks," Tech. Rep. CMU-CS-94-155, School of Computer Science, Carnegie Mellon University, 1994.

[61] J. Yang, R. Stiefelhagen, U. Meier, and A. Waibel, "Visual tracking for multimodal human computer interaction," in *Proceedings of CHI 98*, 1996, pp. ?–?

[62] T. S. Jebara and A. Pentland, "Parameterized structure from motion fro 3d adaptive feedback tracking of faces," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1997, pp. 144–150.

[63] M. F. Kelly and M. D. Levine, "Annular symmetry operators: A method for locating and describing objects," in *Proceedings of the Fifth International Conference on Computer Vision*, 1995, pp. 1016–1021.

[64] B. Moghaddam, W. Wahid, and A. Pentland, "Beyond eigenfaces: Probabilistic matching for face recognition," in *Proceedings of the Third International Conference on Automatic Face and Gesture Recognition*, 1998, pp. ?–?

[65] B. J. Frey, A. Colmenarez, and T. S. Huang, "Mixtures of local subspaces for face recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1998, pp. 32–37.

[66] S. Roweis and Z. Ghahramani, "A unifying review of linear gaussian methods," *Neural Computation*, vol. 11, no. 2, 199.

[67] G. E. Hinton, B. Sallans, and Z. Ghahramani, "A hierarchical community of experts," in *Learning in Graphical Models*, M. I. Jordan, Ed. Kluwer Academic Publishers, 1997, To appear.

[68] M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analysis," *Neural Computation*, vol. 6, pp. 181–214, 1997.

[69] R. M. Neal and P. Dayan, "Factor analysis using delta-rule wake-sleep learning," Tech. Rep., Dept. of Statistics, University of Toronto, 1996.

[70] B. J. Frey, G. E. Hinton, and P. Dayan, "Does the wake-sleep algorithm produce good density estimators?," in *Advances in Neural Information Processing Systems*, 1996, vol. 8, pp. ?–?

[71] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, 1997.

[72] Y. Adini, Y. Moses, and S. Ullman, "Face recognition: The problem of compensating for changes in illumination direction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 721–732, 1997.

[73] A. S. Georghiades and D. J. Kriegman, "Illumination cones for recognition under variable lighting: faces," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1998, pp. 52–58.

[74] W. Zhao, R. Chellappa, and A. Krishnaswamy, "Discriminant analysis of principal components for face recognition," in *Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition*, 1998, pp. 336–341.

[75] A. C. Loui, C. N. Judice, and S. Liu, "An image database for benchmarking of automatic face detection and recognition algorithms," in *Proceedings of International Conference on Image Processing*, 1998, pp. ?–?