International Journal of Latest Engineering and Management Research (IJLEMR)
ISSN: 2455-4847
www.ijlemr.com || Volume 02 - Issue 03 || March 2017 || PP. 59-71

REAL-TIME MULTI VIEW FACE DETECTION AND POSE ESTIMATION

AISHWARYA.S[1] , RATHNAPRIYA.K[1], SUKANYA SARGUNAR.V[2]
[1]*U. G STUDENTS, DEPT OF CSE, ALPHA COLLEGE OF ENGINEERING, CHENNAI,*
[2]*ASST PROF.DEPARTMENT OF CSE, ALPHA COLLEGE OF ENGINEERING, CHENNAI*

Abstract: In this paper we are going to handle multi view face detection and pose estimation in such a way that it works better in all environmental conditions and extreme positions. Face detection has been one of the fundamental technologies to enable natural human-computer interaction. Advancement in computer technology has made possible to evoke new image processing applications in field of face detection and recognizing human.it is important in variety of applications such as security cameras, video surveillance systems, robot vision.it can achieve impressive performance on face detection, a recent study shows that face detection can be further improved by using deep learning, leveraging the high capacity of deep convolution networks. That is the explicit mechanism to handle occlusions, the face detector therefore fails to detect faces with heavy occlusions and extreme poses. While frontal face detection has been largely considered as a solved problem, whereas multi view face detection remains a challenging task due to dramatic appearance changes under various poses, illuminations and expression conditions, various poses and appearance changes are the major tasks in multi view face detection.
Keywords: Multi View Face Detection; Face Recognition; Pose Estimation; Occlusions

I. INTRODUCTION

The objective of face detection is to find and locate faces in an image.it is the first step in automatic face recognition application. Face detection has been well studied for frontal and near frontal faces, but it also plays important role in multi view face detection also. However in unconstrained scenes such as faces in a crowd, state-of-the-art face detectors fail to perform well due to large pose variations , illuminations variation , occlusions , expression variation , low image resolution .Human face detection is an important processing in a variety of applications such as security cameras , video-surveillance systems , robot vision & so forth. A number of algorithms have been proposed so far. However , face detection is still a challenging task since it is difficult to achieve robustness under various situations such as different illuminations and occlusions. In image based approaches , on the other hand , face images are handled as a whole .Face detection and alignment are essential to many face applications , such as face recognition and facial expression analysis. However the large visual variations of faces , such as occlusions, large pose variations and extreme lightings, impose great challenges for these tasks in real world applications. Automatic human face detection is an challenging field of research with many useful real life applications. The use of computer vision in security applications and to minimize intervention of human beings has led the research in field of face biometrics. Face is vital part of human being that represents important information like expression, attention, identity etc. of an individual. The goal of face detection is to locate the occurrence of face in the frame and recognition system retrieves the identity of person for authorization. The main application of face recognition is "access control" that grants certain permissions to person detected.

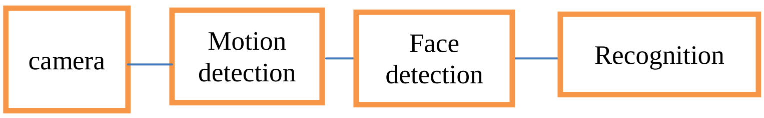


Fig. 1. System Architecture

A face recognition system automatically identifies a human face from database images. The face recognition problem is challenging as it has to account all possible variation in face caused by change in facial features, illumination, occlusions, etc. Face recognition systems requires high processing efficiency as well as reliability. Recognition stage uses face to identify a person and claim identity. Face recognition is time consuming process as it has to undergo large number of comparisons. Then the purpose of this paper is to further extend the capability of the system and make it applicable to multi-view face detection, the ultimate goal of this research is to develop a medical analysis system extracting structural characteristics of a human face from multi-view angles.

II. RELATED WORK

These research in face detection and recognition has reached to far extent. This process has started from 1960 when the first semiautomated algorithm was developed. In 1988 principle component analysis was applied for face recognition which boosted the research i recognition stage. In 1991, reliable real time automated recognition was reached by using residual error to detect faces where eigen face technique was used. In recent decade Viola and Jones [1] gave a new boost to real time face detection with the use of wavelet based method. The very first stage in the system is motion sensing or detection. The rest stages rely highly on this stage.

A. Background subtraction

Background model is established with absence of moving object. Adaptive background model is where a series of frames are used and an average is calculated from all of them over a period of time. This approach is useful if there are continuous changes in background scene. Non-adaptive background model is where a frame is taken and saved as the background. This approach is useful in static and indoor environments where there are very less variations in background over a time period. The background is then subtracted from current frame to segment moving objects. The decision of presence of moving object is based on the difference between the two images or frames. This paper focuses the temporal averaging method [2] for background subtraction. First frame is taken as background frame(Fig. 2).

Fig. 2. Background Image

For each frame new background model $B(x,y)$ is estimated as:

$Bt+1(x,) = \alpha It(x,y) + (1-\alpha)Bt(x,y)$ (1)

where $It(x,y)$ is current pixel value, t is frame number, (x,y) is pixel location in frame and α is learning rate (speed of updating background model).

The difference between current frame and background is given by

$D(x,y) = | \mathrm{l}t(x,y) - Bt(x,y) |$ (2)

The pixel whose difference value is greater than given threshold T are classified as foreground pixels given by Eq. 3.

$Mtx, y = \ 0\ Dt\ (x, y) \leq T$
$1\ Dt\ (x, y) > T$ (3)

Fig. 3 shows moving object in frame. The difference between background and current frame is shown in Fig. 4. The learning rate is made adaptive using the value of difference between current frame and background model. Threshold is made adaptive using learning rate and the value of difference between current frame and background model [2]

Fig. 3 moving object in current frame

B. Face detection

The next stage in succession is face detection module which is called only if motion is detected. The original Viola and Jones method intended for real time applications runs at 15 frames per second [1]. Within last decade the method has been used and improvised by many researchers for real-time applications. In our approach in order to speed up computations Violas face detection algorithm is applied only to the region which is segmented from background subtraction stage and identified as moving object. The detection method is based on the wavelet transform [3]. It represents shape of object in terms of subset of wavelet coefficients. Haar features with varying scale or location are computed in constant time using integral image [5],[4] , [6]. The four variance based Haar features are shown in Fig. 5. The variance of random variable X is calculated as follows:

$$Var\ (X)=E(X2)-\mu 2 \qquad (4)$$

where $E(X2)$ is expected value of $X2$ and μ is expected value of X.

The value of a rectangle feature is computed as the difference between sum of variance values in white region and sum of variance values in dark region. These features can be computed using integral image (x,y) and squared integral image $I2(x,y)$ obtained using Eq. 5 and Eq. 6 respectively [37].

$$(x,y)= \sum_{m=1}^{X} \sum_{n=1}^{y} f(m,n) \qquad (5)$$

$$I^2\ (x,)= \sum_{m=1}^{X} \sum_{n=1}^{y} f^{2}\ (m,n)\ \text{n} \qquad (6)$$

The value of the integral image at any location represents the sum of all pixel values to left and above it. Integral images are also used for calculating the values of $(X2)$, μ and variance at any position in an image given in Eq. 7.

$$\mu =1\ /N\ (I_1 + I_4 - I_2 - I_3) \qquad (7)$$
$$E\ (f\ (x,y)^2) = 1/\ N\ (I_1^{\ 2}\ + I_2^{\ 4} - I_2^{\ 2} - I_2^{\ 3}) \qquad (8)$$
$$Var\ (f\ (x,\text{y})) = E\ (f\ (x,y)^2) - \mu^2 \qquad (9)$$

where N is number of elements within region D.

The Haar classifier multiplies the weight of each rectangle by its area and adds to the results.

This method is not robust to face pose variation. It shows good results in illumination variations.

Fig . 4. Face detection using Haar features

C. Face Recognition

This is last and important stage of system. Recognition needs to project query image (unknown face) into face space. This image is then classified as known or unknown by comparing it with the faces of known individuals in database. This method decomposes face images into principal components that are characteristic features of images called ‘Eigenfaces’. The eigenface approach uses the Principle Component Analysis(PCA) to reduce the dimension of face image.

(a) Face Detection with Background Subtraction

(b) Face Detection without Background Subtraction

Fig 5: Face detection

III. EXISTING ALGORITHM.

A. Deep cascaded multitask framework

Face alignment also attracts extensive research interests. Research works in this area can be roughly divided into two categories, regression-based methods [7], [8], [9], and template fitting approaches [10], [11], [12]. Recently, Zhang et al. [13] proposed to use facial attribute recognition as an auxiliary task to enhance face alignment performance using deep CNN. However, most of previous face detection and face alignment methods ignore the inherent correlation between these two tasks.

Though several existing works attempt to jointly solve them, there are still limitations in these works. For example, Chen et al. [14] jointly conduct alignment and detection with random forest using features of pixel value difference. But, these handcraft features limit its performance a lot. Zhang et al. [15] use multitask CNN to improve the accuracy of multi view face detection but the detection recall is limited by the initial detection window produced by a weak face detector.

On the other hand, mining hard samples in training is critical to strengthen the power of detector. However, traditional hard sample mining usually performs in an offline manner, which significantly increases the manual operations. It is desirable to design an online hard sample mining method for face detection, which is adaptive to the current training status automatically. In this letter, we propose a new framework to integrate these two tasks using unified cascaded CNNs by multitask learning. The proposed CNNs consist of three stages.

In the first stage, it produces candidate windows quickly through a shallow CNN. Then, it refines the windows by rejecting a large number of nonfaces windows through a more complex CNN. Finally, it uses a more powerful CNN to refine the result again and output five facial landmarks positions. Thanks to this multi task learning framework, the performance of the algorithm can be notably improved. The major contributions of this letter are summarized as follows:

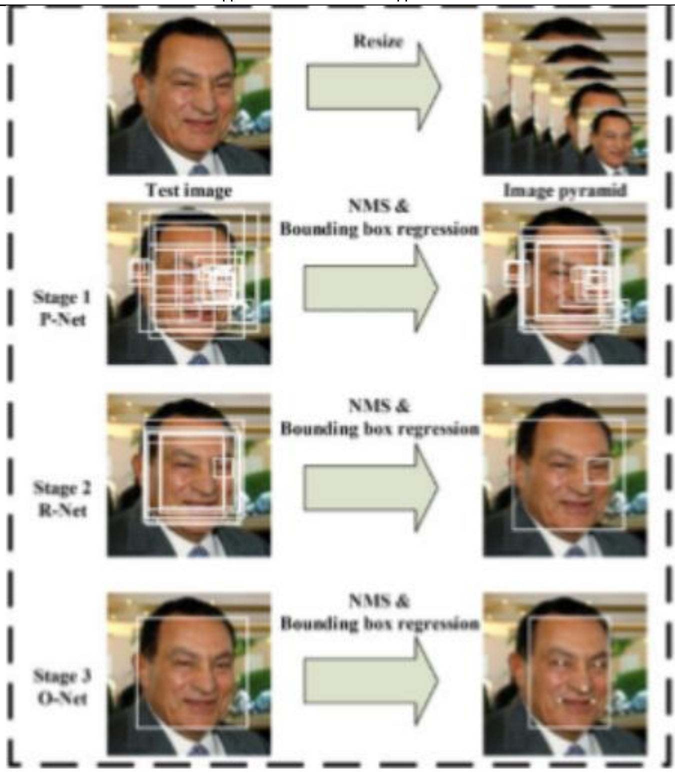


Fig.6. Pipeline of our cascaded framework that includes three-stage multitask deep convolutional networks. First, candidate windows are produced through a fast P-Net. After that, we refine these candidates in the next stage through a R-Net. In the third stage, the O-Net produces final bounding box and facial landmarks position.

1) We propose a new cascaded CNNs-based framework for joint face detection and alignment, and carefully design lightweight CNN architecture for real-time performance
2) We propose an effective method to conduct online hard sample mining to improve the performance.
3) Extensive experiments are conducted on challenging benchmarks to show significant performance improvement of the proposed approach compared to the state-of-the-art techniques in both face detection and face alignment tasks.

B. CNN Architectures

In [16], multiple CNNs have been designed for face detection. However, we notice its performance might be limited by the following facts:
1)Some filters in convolution layers lack diversity that may limit their discriminative ability; (2) compared to other multiclass objection detection and classification tasks, face detection is a challenging binary classification task, so it may need less numbers of filters per layer. To this end ,we reduce the number of filters and change the 5×5 filter to 3×3 filter to reduce the computing, while increase the depth to get better performance. With these improvements, compared to the previous architecture in [16] , we can get better performance with less runtime (the results in training phase are shown in Table I. For fair comparison, we use the same training and validation data in each group). We apply PReLU [17] as nonlinearity activation function after the convolution and fully connection layers (except output layers).

C. Training Data

Since we jointly perform face detection and alignment, here we use following four different kinds of data annotation in our training process:
1) negatives: regions whose the intersection-over-union (IoU) ratio is less than 0.3 to any ground-truth faces
2) positives: IoU above 0.65 to a ground truth face
3) part faces: IoU between 0.4 and 0.65 to a ground truth face; and
4) landmark faces: faces labeled five landmarks' positions.

There is an un clear gap between part faces and negatives ,and there are variances among different face annotations. So, we choose IoU gap between 0.3 and 0.4. Negatives and positives are used for face classification tasks , positives and part faces are used for bounding box regression, and landmark faces are used for facial landmark localization. Total training data are composed of 3:1:1:2 (negatives/positives/part face/landmark face) data. The training data collection for each network is described as follows:

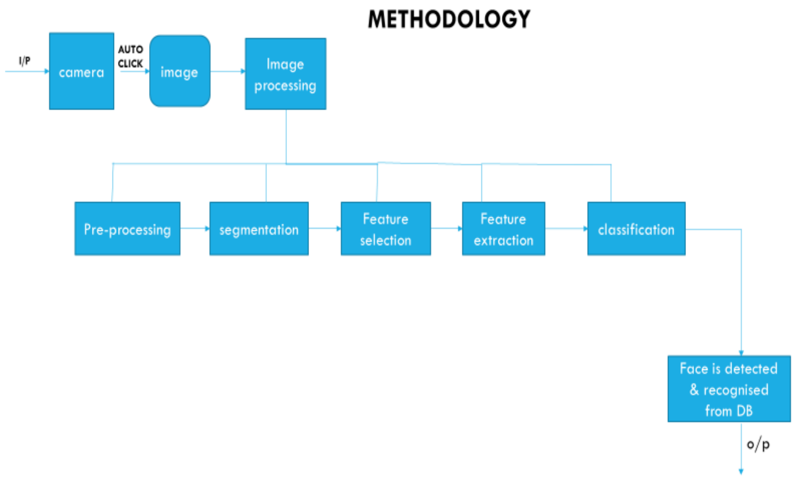
1) P-Net: We randomly crop several patches from WIDER FACE [24] to collect positives, negatives, and part face. Then , we crop faces from CelebA [18] as land mark faces.
2) R-Net: We use the first stage of our framework to detect faces from WIDER FACE [19] to collect positives, negatives, and part face while landmark faces are detected from CelebA [18].
3) O-Net : Similar to R-Net to collect data ,but we use the first two stages of our framework to detect faces and collect data.

D. Runtime Efficiency

Given the cascade structure, our method can achieve high speed in joint face detection and alignment. We compare our method with the state-of-the-art techniques on GPU and the results are shown in Table II. It is noted that our current implementation is based on unoptimized MATLAB codes.

IV. PROPOSED METHODOLOGY

Our proposed framework can achieve promising results for both face hallucination and recognition.



A. Pre-processing

Pre-processing is a common name for operations with images at the lowest level of abstraction both input and output are intensity images. The aim of pre-processing is an improvement of the image data that suppresses unwanted distortions or enhances some image features important for further processing.

Four categories of image pre-processing methods according to the size of the pixel neighborhood that is used for the calculation of new pixel brightness:

- Pixel brightness transformations,
- Geometric transformations,
- Pre-processing methods that use a local neighborhood of the processed pixel, and Image restoration

Image pre-processing methods use the considerable redundancy in images. Neighboring pixels corresponding to one object in real images have essentially the same or similar brightness value. Thus, distorted pixel can often be restored as an average value of neighboring pixels.

If pre-processing aims to correct some degradation in the image, the nature of a priori information is important.

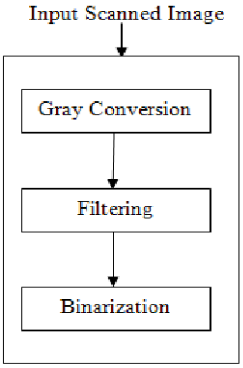


Fig.7. Block diagram of Image pre-processing

B. Grayscale Conversion

In photography and computing, a grayscale or grayscale digital image is an image in which the value of each pixel is a single sample, that is, it carries only intensity information. Images of this sort, also known as black-and-white, are composed exclusively of shades of gray, varying from black at the weakest intensity to white at the strongest . Grayscale images are distinct from one-bit bi-tonal black-and-white images, which in the context of computer imaging are images with only two colors, black and white (also called bi-level or binary images). Grayscale images have many shades of gray in between.

Grayscale images are often the result of measuring the intensity of light at each pixel in asingle band of the electromagnetic spectrum (e.g. infrared, visible light, ultraviolet, etc.), and in such cases they are monochromatic proper when only a given frequency is captured.

Converting color to grayscale:

Conversion of a color image to grayscale is not unique; different weighting of the color channels effectively represents the effect of shooting black-and-white film with different-colored photographic filters on the cameras

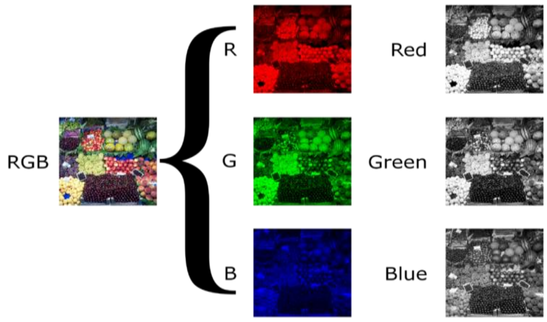
.

Colorimetric (luminance-preserving) conversion to grayscale:

A common strategy is to use the principles of photometry or, more broadly, colorimetric to match the luminance of the grayscale image to the luminance of the original color image. This also ensures that both images will have the same absolute luminance, as can be measured in its SI units of candelas per square meter, in any given area of the image, given equal white points. In addition, matching luminance provides matching perceptual lightness measures, such as L* (as in the 1976 CIE Lab color space) which is determined by the linear luminance Y (as in the CIE 1931 XYZ color space) which we will refer to here as Linear to avoid any ambiguity.

Grayscale as single channels of multichannel color images:

Color images are often built of several stacked color channels, each of them representing value levels of the given channel. For example, RGB images are composed of three independent channels for red, green and blue primary color components; CMYK images have four channels for cyan, magenta, yellow and black ink plates, etc.

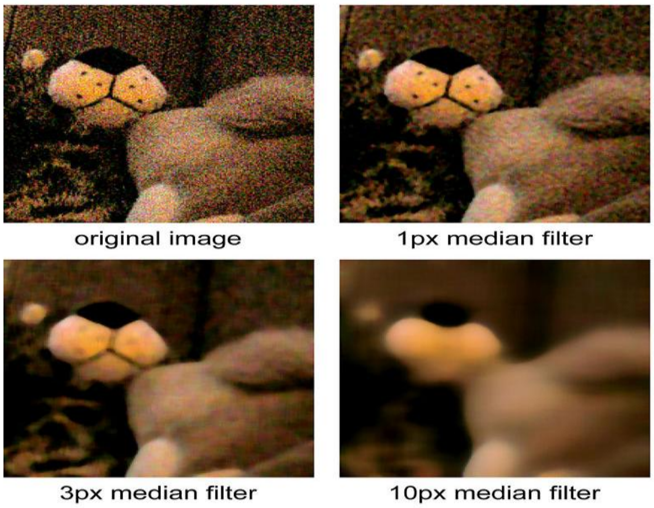


C. Median Filter

In signal processing, it is often desirable to be able to perform some kind of noise reduction on an image or signal.

The median filter is a nonlinear digital filtering technique, often used to remove noise. Such noise reduction is a typical pre-processing step to improve the results of later processing (for example, edge detection on an image).

Median filtering is very widely used in digital image processing because, under certain conditions, it preserves edges while removing noise



ALGORITHM DESCRIPTION:

The main idea of the median filter is to run through the signal entry by entry, replacing each entry with the median of neighboring entries.

The pattern of neighbors is called the "window", which slides, entry by entry, over the entire signal.

For 1D signal, the most obvious window is just the first few preceding and following entries, whereas for 2D (or higher-dimensional) signals such as images, more complex window patterns are possible (such as "box" or "cross" patterns).

D. Image Segmentation

In computer vision, image segmentation is the process of partitioning a digital image into multiple segments (sets of pixels, also known as super-pixels). The goal of segmentation is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze. Image segmentation is typically used to locate objects and boundaries (lines, curves, etc.) in images. More precisely, image segmentation is the process of assigning a label to every pixel in an image such that pixels with the same label share certain characteristics.

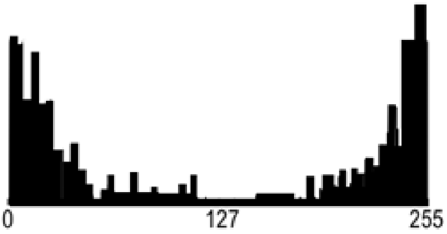
The result of image segmentation is a set of segments that collectively cover the entire image, or a set of contours extracted from the image (see edge detection). Segmentation techniques are either contextual or non-contextual. The latter take no account of spatial relationships between features in an image and group pixels together on the basis of some global attribute, e.g. grey level or color. Contextual techniques additionally exploit these relationships, e.g. group together pixels with similar grey levels and close spatial locations.

Simple thresholding

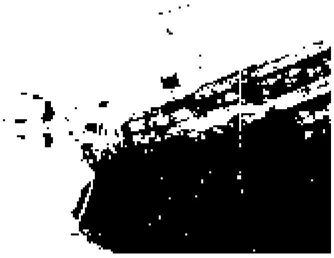
The most common image property to threshold is pixel grey level: $g(x,y) = 0$ if $f(x,y) < T$ and $g(x,y) = 1$ if $f(x,y) \geq T$, where T is the threshold. Using two thresholds, $T_1 < T_1$, a range of grey levels related to region 1 can be defined: $g(x,y) = 0$ if $f(x,y) < T_1$ OR $f(x,y) > T_2$ and $g(x,y) = 1$ if $T_1 \leq f(x,y) \leq T_2$.

International Journal of Latest Engineering and Management Research (IJLEMR)
ISSN: 2455-4847
www.ijlemr.com || Volume 02 - Issue 03 || March 2017 || PP. 59-71

Greyscale image



Its grey level histogram



Binary regions for $T = 26$

Binary regions for $T = 133$

Binary regions for $T = 235$

The Haar wavelet is also the simplest possible wavelet. The technical disadvantage of the Haar wavelet is that it is not continuous, and therefore not differentiable. This property can, however, be an advantage for the analysis of signals with sudden transitions, such as monitoring of tool failure in machines.

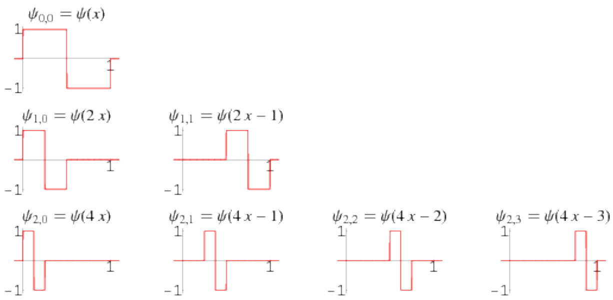
Define

$$\psi(x) \equiv \begin{cases} 1 & 0 \le x < \frac{1}{2} \\ -1 & \frac{1}{2} < x \le 1 \\ 0 & \text{otherwise} \end{cases} \qquad (1)$$

and

$$\psi_{jk}(x) \equiv \psi\left(2^j x - k\right) \qquad (2)$$

for j a nonnegative integer and $0 \le k \le 2^j - 1$.



So, for example, the first few values of $\psi_{jk}(x)$ are

ψ_{00}	=	$\psi(x)$	**(3)**
ψ_{10}	=	$\psi(2x)$	(4)
ψ_{11}	=	$\psi(2x-1)$	(5)
ψ_{20}	=	$\psi(4x)$	(6)
ψ_{21}	=	$\psi(4x-1)$	(7)

$\psi_{22} = \psi(4x-2)$ (8)

$\psi_{23} = \psi(4x-3).$ (9)

Then a function $f(x)$ can be written as a series expansion by

$$f(x) = c_0 + \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} c_{jk}\, \psi_{jk}(x). \quad \textbf{(10)}$$

The functions ψ_{jk} and ψ are all orthogonal in $[0, 1]$, with

$$\int_0^1 \psi(x)\, \psi_{jk}(x)\, dx = 0 \quad (11)$$

$$\int_0^1 \psi_{jk}(x)\, \psi_{lm}(x)\, dx = 0 \quad (12)$$

for $(j, k) \neq (0, 0)$ in the first case and $(j, k) \neq (l, m)$ in the second.

These functions can be used to define wavelets. Let a function be defined on n intervals, with n a power of 2. Then an arbitrary function can be considered as an n-vector $\mathbf{f}$, and the coefficients in the expansion $\mathbf{b}$ can be determined by solving the matrix equation

$$\mathbf{f} = \mathsf{W}_n\, \mathbf{b} \quad \textbf{(13)}$$

for $\mathbf{b}$, where W is the matrix of ψ basis functions. For example, the fourth-order Haar function wavelet matrix is given by

$$W_4 = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & -1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & 0 & -1 \end{bmatrix} \quad \textbf{(14)}$$

$$= \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

E. Feature Extraction

In machine learning, pattern recognition and in image processing, featureextraction starts from an initial set of measured data and builds derived values (features) intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps, and in some cases leading to better human interpretations. Feature extraction is related to dimensionality reduction.

When the input data to an algorithm is too large to be processed and it is suspected to be redundant (e.g. the same measurement in both feet and meters, or the repetitiveness of images presented as pixels), then it can be transformed into a reduced set of features (also named a feature vector). Determining a subset of the initial features is called *feature selection.*

PCA for Feature extraction.

Covariance Matrix

Covariance is the measure of how two different variables change together. The covariance between two variables, X and Y, can be given by the following formula.

$cov=\sum i=1n(Xi-\bar{X})(Yi-\bar{Y})(n-1)$

Now, if we wanted to look at all the possible covariance's in a dataset, we can compute the covariance matrix, which has this form –

$C=\left(cov(x,x)cov(y,x)cov(z,x)cov(x,y)cov(y,y)cov(z,y)cov(x,z)cov(y,z)cov(z,z) \right)$

Notice that this matrix will be symmetric ($A=AT$), and will have a diagonal of just variances, because $cov(x,x)$ is the same thing as the variance of x. If you understand the covariance matrix and eigenvalues/vectors, you're ready to learn about PCA.

F. Support vector machine (SVM)

In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

V. ADVANTAGES OF PROPOSED SYSTEM

1. Performance has been improved.
2. Experimental results show that SVMs achieve significantly higher search accuracy than traditional query refinement schemes after just three to four rounds of relevance feedback.
3. Images can be captured and detected in extreme positions , occlusions, various angles and illuminations .
4. Suits for well-lit faces.

VI. EXPERIMENTAL RESULTS AND DISCUSSION

The Samples used in this work were 200 face images of 10 people taken from 20 directions between 0 ^{o} to 360 $^{o.}$ The sample photos were prepared as a preliminary database to use in the multi dimensional face detection. Face detection was carried out on face images angled at 0^{o} , 30^{o} , 60^{o} , 90^{0},
120^{o}, 150 ^{o}, 180^{o} and upto 360 ^{o}. The detection rate was evaluated by the cross validation . Namely all face images except for one person were utilized as templates and the face detection carried out for the face images of the person excluded from the template.

VII. CONCLUSION

A multi-view face detection and pose estimation system has been developed successfully. Approximately 85% detection rates were obtained by employing the multi clue derived from our original feature vectors. To further improve the accuracy and the performance , the SVM algorithm has been used in this paper. As the result , more than 90% detection rate has been achieved for profile and the performance of the system has been successfully improved.

ACKNOWLEDGEMENT

Author would like to give sincere gratitude especially to Mrs.Sukanya Sargunar.V (Asst Prof CSE) for their guidance and support to pursue this work.

REFERENCES:

[1]. P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, 2001, pp. 511–518.
[2]. N. Kostov and B. Nikolov, “Motion detection using adaptive temporal averaging method,” Radio Engineering, vol. 23, no. 2, pp. 652–658, 2014.
[3]. S. Gundimada and V. Asari, “Face detection technique based on rotation invariant wavelet features,” in Proceedings of International Conference on Information Technology: Coding and Computing, vol. 2, April 2004, pp. 157–158.
[4]. L. Zhang and Y. Liang, “A fast method of face detection in video images,” 2nd International Conference on Advanced Computer Control ICACC, vol. 4, pp. 490–494, 2010.
[5]. P. Viola and M. J. Jones, “Robust real-time face detection,” International Journal of Computer Vision, vol. 57, no. 2, pp. 137–154, May 2004.
[6]. Z. Mei, J. Liu, Z. Li, and L. Yang, “Study of the eye-tracking methods based on video,” in Third International Conference on Computational Intelligence, Communication Systems and Networks, 2011, pp. 1–5.
[7]. X. P. Burgos-Artizzu, P. Perona, and P. Dollar, “Robust face landmark estimation under occlusion,” in IEEE Int. Conf. Computer. Vis., 2013, pp. 1513–1520.
[8]. X. Cao, Y. Wei, F. Wen, and J. Sun, “Face alignment by explicit shape regression,” Int. J. Comput. Vis., vol. 107, no. 2, pp. 177–190, 2012.

[9]. J. Zhang, S. Shan, M. Kan, and X. Chen, "Coarse-to-fine auto-encoder networks (CFAN) for real-time face alignment," in Eur. Conf. Comput. Vis., 2014, pp. 1–16.
[10]. X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in IEEE Conf. Comput. Vis. Pattern Recognit., 2012, pp. 2879–2886.
[11]. T.F.Cootes,G.J.Edwards,andC.J.Taylor,"Activeappearancemodels," IEEE Trans. Pattern Anal. Mach. Intell., vol. 23, no. 6, pp. 681–685, Jun. 2001.
[12]. X. Yu , J.Huang , S. Zhang , w.Yan and D.Metaxus , "Pose – free facial landmark fitting via optimized part mixtures and cascaded deformable shape model," in IEEE Int. Conf. Comput. Vis., 2013, pp. 1944–1951.
[13]. Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in Eur. Conf. Comput. Vis., 2014, pp. 94–108.
[14]. D.Chen,S.Ren,Y.Wei,X.Cao,andJ.Sun,"Joint cascade face detection and alignment," in Eur. Conf. Comput. Vis., 2014, pp. 109–122.
[15]. C. Zhang and Z. Zhang, "Improving multiview face detection with multitask deep convolutional neural networks," in IEEE Winter Conf. Appl. Comput. Vis., 2014, pp. 1036–1041.
[16]. H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," in IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 5325–5334.
[17]. K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in IEEE Int. Conf. Comput. Vis., 2015, pp. 1026–1034.
[18]. Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in IEEE Int. Conf. Comput. Vis., 2015, pp. 3730–3738.
[19]. S.Yang,P.Luo,C.C.Loy,andX.Tang,"WIDER FACE:A Face detection benchmark," arXiv:1511.06523.