Real Time Face Tracking and Head Pose Estimation for Augmented Reality in Mobile Platforms

2

Introduction

3

Background

Augmented Reality (AR)[1]

"A technology that superimposes computer-generated images and graphics onto real world environments"

Mobile Augmented Reality (MAR)

Environment and self augmentation through mobile devices

LiveRoom[2]

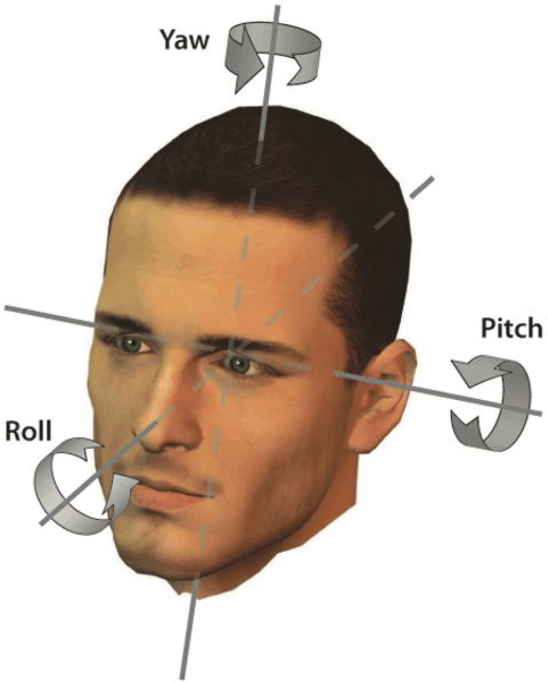
An application that augments the environment allowing users to virtually try out online products before purchasing

4

Background ctd.

Head pose

Translational and rotational movements (pitch, roll and yaw) of the head along the 3 axes.[3]



5

Problem

LiveRoom supports only **environment augmentation** through the back facing camera.

It does not have support for **self augmentation** through the front facing camera.

Environment Augmentation[2]

Self Augmentation[4]

6

Problem ctd.

Need

Placing virtual accessories realistically on the human face providing a seamless AR experience

Limitations

- Low processing power
- Memory limitations
- Low quality video stream captured by the front facing camera

Requirement

An approach that allows real time tracking of face and estimation of head pose for mobile platforms

7

Project output

- A real time face tracking and head pose estimation algorithm optimized for the front facing camera in mobile devices.
- A plugin for the Unity game engine which can be integrated with the LiveRoom AR application.

8

SpeculAR

An AR based demo mobile application which allows users to try on virtual models of eyewear by augmenting them on human faces.

9

Literature Review

10

Methods

Aim, design and setting of the study and description of all processes and methodologies employed.

Methods

- **Aim of the study:**
  - Reveal the existing face detection, tracking and head pose estimation algorithms and related APIs and SDKs.
- **Methods followed:**
  - Related peer reviewed journals and conference papers were searched mainly on the Google Scholar database.
  - Additional conference papers and journal articles were found going through the reference sections.
  - Websites which contained information on latest APIs, SDKs and commercial AR applications in the mobile AR domain were searched.

56

Peer-reviewed conference papers and journal articles

From 1992 - 2016

Covering a wide range of both fairly old and new techniques

7

Websites on related APIs, SDKs and mobile applications

13

Workflow

14

Face Detection and Feature Extraction

15

Viola-Jones Object Detection Framework

- Based on a cascade of weak classifiers that can detect separate Haar-like features identified in a facial image[5].
- A set of classifiers is trained using supervised learning to detect examples of a particular class.

Examples where the Haar-like features can be applied to differentiate certain unique features in the human face[5].

16

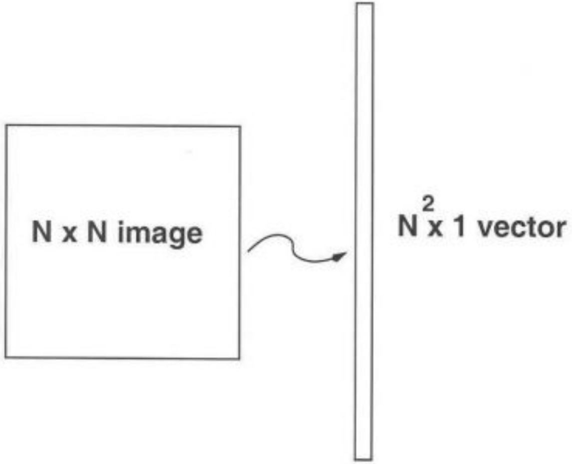
Viola-Jones Object Detection Framework ctd.

- Within an enclosing detection window of resolution 24x24, 160,000+ Haar-like features found.
- A variant of the Adaboost algorithm is used to select a small set of features and use in a cascade of weak classifiers to identify unseen faces.
- Experiments have yielded
  - A detection rate of 95%
    - With a false positive rate of 1 in 14084 for a frontal face classifier constructed from 200 features
  - 0.7 seconds to scan a 384 x 288 pixel image
- **Disadvantage**
  - Can only detect frontal faces and can hardly detect even faces rotated 45^0 around.

17

Using Eigenfaces

- A template matching problem.
- Obtain face images I1 , I2 , ..., IM (training faces).
- Represent every image Ii as a vector.
- Compute the average face vector.
- Subtract the mean face.
- Represent image as a linear combination of the best K eigenvectors.
- If distance from face space is less than a threshold, then that is a face.[6]



18

Skin based face detection

- Using human skin as a mode of detecting a human face and its enclosing features.[7]
- A geometrical facial model to detect the facial region and extract its features like eyes, mouth, nostrils and eyebrows.
- Successful results with near frontal facial images under proper lighting conditions.
- Has enough performance to be run on low computational power devices like mobile platforms.
- **Disadvantage**
  - Lacks robustness to perform accurately in varying lighting conditions and in presence of face occlusions.

19

Neural Networks based face detection

- Henry A. Rowley, Shumeet Baluja, and Takeo Kanade developed a method for recognizing faces in grayscale images utilizing neural networks.[8]
- An upright frontal face identification framework
- Utilizes a retinally associated neural network system to inspect small windows of a picture and choose whether each window contains a face.
- Mediates between multiple networks to enhance execution.
- Direct strategy for training positive face examples. False discoveries at the training stage are used as negative face examples.

20

Neural Networks based face detection ctd.

- Has 2 stages[8].
  - **Stage1:**
    - Gets a 20x20 pixel area of the image as input, and produces a result of 1 to -1, meaning the presence or absence of a face in that area.
    - Subsampling.
  - **Stage2:**
    - The arbitrator then merges detections from individual filters and eliminates overlapping detections.
- Experiments have yielded,
    - A detection rate between 77.9% and 90.3%.

21

Support Vector Machines (SVMs)

- The hyperplane of the SVM is found by solving a linearly constrained quadratic programming problem[9].
- Present a decomposition algorithm that can be used to train SVMs over very large datasets guaranteeing optimality.
- **Steps**
  - Scans an image for faces at many possible scales by dividing the original image into overlapping sub images.
  - The sub images are then classified using the SVM as a face or a non-face.
- **Advantages**
  - Can capture non-frontal facial images up to a small degree

22

Support Vector Machines (SVMs)

- Their implementation has the ability to deal with about **2,500 support vectors** on a machine with **128MB of RAM**.
- Detection rate of **97.1%** in high quality images and a detection rate of **74.2%** in mixed quality images.

Results captured by the SVM face detection system, detecting non-frontal faces[9]

23

Facial feature extraction

- Identifying and locating interesting features in a facial image like eyes, eyebrows, nose and mouth or any other interesting facial features that do not correspond to a sensible human organ.
- Locations of some features might help in narrowing down the search for other features [10].
- In general, feature based facial feature extraction methods can be put under two broad categories as low level methods and template based methods.

24

Low Level Methods

Uses color or the grey levels in the image

25

Color based

- The basic approach is skin color analysis.
- Good candidate for systems that require real time performance [7] making them suitable for mobile AR applications of our interest.
- Images need to be converted into NCC color space to enhance the human skin present in an image [11].
- Then a histogram is drawn. Skin color found in the image will occupy a small cluster in that histogram.
- Identify the cluster and check if a given point lies in it or not.
- Authors of this method have focussed more on better accuracy and robustness. No enough evidence found on suitability for mobile platforms.

26

Color based ctd.

- R.L. Hsu et al. proposed a method that involves constructing feature maps for eyes, mouth and face boundary[12].
- To detect the eyes, two separate eye maps are built where one corresponds to the chrominance component and the other corresponds to the luminance component.
- Has given around 80% average accuracy but the average computational time has ranged from around 10 - 40 seconds when run on a 860 MHz CPU.
- Quite improbable for it to give real time results when run on a general mobile device that does not have a high computational power.

Edge detection based

- Very commonly used technique in detecting features.
- As Hjelmas and Low state in their survey[13], the Sobel operator is the most common edge detection approach used in facial feature extraction methods.
- Marr-Hildreth edge operator[14] and a variety of first and second derivatives of Gaussians have also been utilized in some other approaches[15].
- K.C. Yow and R. Cipolla[15] talks about a method of detecting facial features using a simple facial model.

28

Edge detection based ctd.

- Smoothing the image and then filtering it with a second derivative Gaussian filter elongated at an aspect ratio of 3:1.
- Local maxima in the resulting image would mark the presence of interesting points that agree to some extent with the facial feature model.

The facial model suggested by R. Cipolla[12]

29

Edge detection based ctd.

- Focus of the authors had been to have their method as universal as possible and their future work is also targeted at making the method more scale invariant and more robust.

a) Facial image b) Interest points

Interesting points found by use of spatial filtering with second derivative Gaussian filter[7]

30

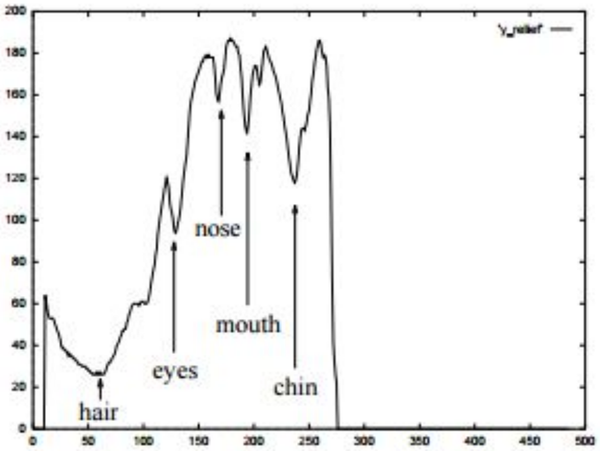
Projections

- Project the image into some plot with respect to some property or quality in the image and analyzes the plot.
- Sobottka and Pitas conducted one research[16] on using grey scale intensities on an enhanced facial image to detect facial features.
- preprocessing step where the dark regions are enhanced by applying a grayscale erosion followed by an extremum sharpening operation.

31

Projections ctd.

- The y-projection (taking means of the rows) would show significant minima for hair, eyes, nose, mouth and chin.



y-projection of a facial image[16]

32

Projections ctd.

- Detection rate varies between 74% to 97% for different facial features and will show poor performance with images of people having different hairstyles and beards.
- Not good for mobile applications since the user base of the application cannot be restricted.

33

Gabor Filters

- Image analysis with Gabor filters is considered to be quite similar to the perception in the human visual system.
- Sirohey and Rosenfeld[17] have used oriented Gabor wavelets that form an approximation to the human eye as the filter for eye detection.
- Their results have given 80% detection rate with Aberdeen database and a 95% detection rate on a certain video frame sequence[17].
- Requires high computational power and hence not suitable for running on mobile platforms.

34

Template Based Methods

Searches for features adhering to a template

35

Correlation

- Involves finding the minimum error between a certain template and a patch in the image.
- Convolute the template with all possible positions in the image and find the locations that gives the maximum value.
- Brunelli and Poggio[18] have used correlation as a step in detecting the facial features.
- First the image is normalized as a preprocessing step.
- Pre registered facial images are kept in a database along with a set of masks.
- A new image is matched with all the masks in the database using the correlation technique and a cumulative score is calculated by comparison.

36

Correlation ctd.

- The highest score corresponds to the correct match.
- Template matching correlation technique gives better accuracy than low level feature based methods[18].
- Issues
  - Cannot handle variations in illumination and object orientation.
  - Relatively high computation time.
- Good candidate for one time feature extraction needs in mobile devices.

Hough transform

- The classical Hough transform technique was focussed on identifying lines in images.
- Later on various extensions of the same technique came out to existence that have the capability of identifying positions of arbitrary shapes like circles or ellipses.
- G. Chow and X. Li[19] have suggested a way of extracting the shapes and locations of eyes and mouth from a relatively unposed head and shoulder image.
- Approach consists of 3 modules
  - Context module: looks for possible faces
  - Eyes module: detects eyes
  - Mouth module: detects mouth

Hough transform ctd.

- Eye and mouth modules uses a combined approach of the Hough transform and deformable template techniques.
- Eye module locates the irises which are modeled as a pair of circles using the circle Hough transformation.
- Accuracy
  - 55% and 45% fits in eye and mouth modules respectively.
- Performance
  - Mouth detection - average time of 3.23 seconds.
  - Eye detection - average time of 11.545 seconds.
- Possible candidate to use in mobile platforms.

39

Face Tracking

Determining where the user’s face is in a predefined space.

40

Kanade-Lucas-Tomasi (KLT) tracking algorithm

- Detects a set of object points across the video frames.
- Once the algorithm detects the face, the next step detects feature points that can be constantly tracked.
- When features are lost, the algorithm replaces the lost features by finding new features.
- Good features are located by examining the minimum eigenvalue of each 2x2 gradient matrix.
- Features are tracked using Newton-Raphson method of minimizing the difference between the two windows
- Remarkably robust for tracking facial images.

KLT tracking algorithm ctd.

- Algorithm has 2 main steps[20].
  - **KLT feature selection:**
    - Find the coordinates in the image which have a varying texture.
  - **KLT feature tracking:**
    - Select features to track based on texture (intensity information).
    - Select a number of fixed-sized feature windows on the first image of a sequence.
    - Use **Sum of Squared intensity Differences (SSD)** as the measurement criterion to realize the tracking of feature points. The target is to calculate the translation distance **d = (Δx, Δy)**.

Active Shape Models (ASMs)

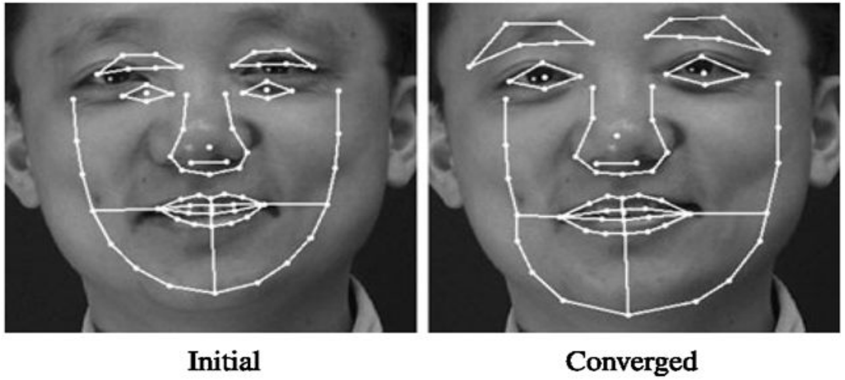
- **A statistical model of shape** that captures the variability of a particular object class given an annotated set of training images[21].
- Constructs a mean shape and derives main modes of shape variation in terms of eigenvectors.

$$\mathbf{x = x' + Q_s p} \rightarrow \mathbf{(1)}$$

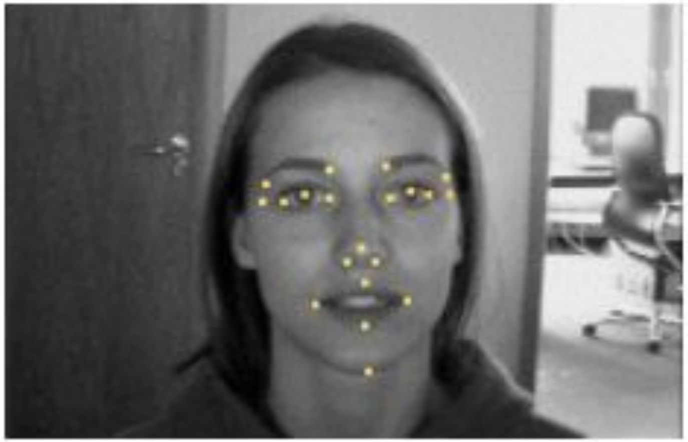
- Each of these modes change the shape by moving landmarks along straight lines normal to the shape contours through mean landmark positions.
- New shapes are generated by modifying the mean shape with weighted sum of these modes.

43

ASMs ctd.



ASM Face alignment [22]



A face from BioID database with correctly positioned landmarks by improved ASM model fitting [22]

4

ASMs ctd.

4

- Stephen Milborrow and Fred Nicolls proposed 6 improvements for locating facial features and tracking human faces with ASMs.
  - Fitting more landmarks than actually needed.
  - Selectively using two instead of one dimensional landmark templates.
  - Adding noise to the training set.
  - Relaxing the shape model where advantageous.
  - Trimming covariance matrices by setting most entries to zero.
  - Stacking two ASMs in series.

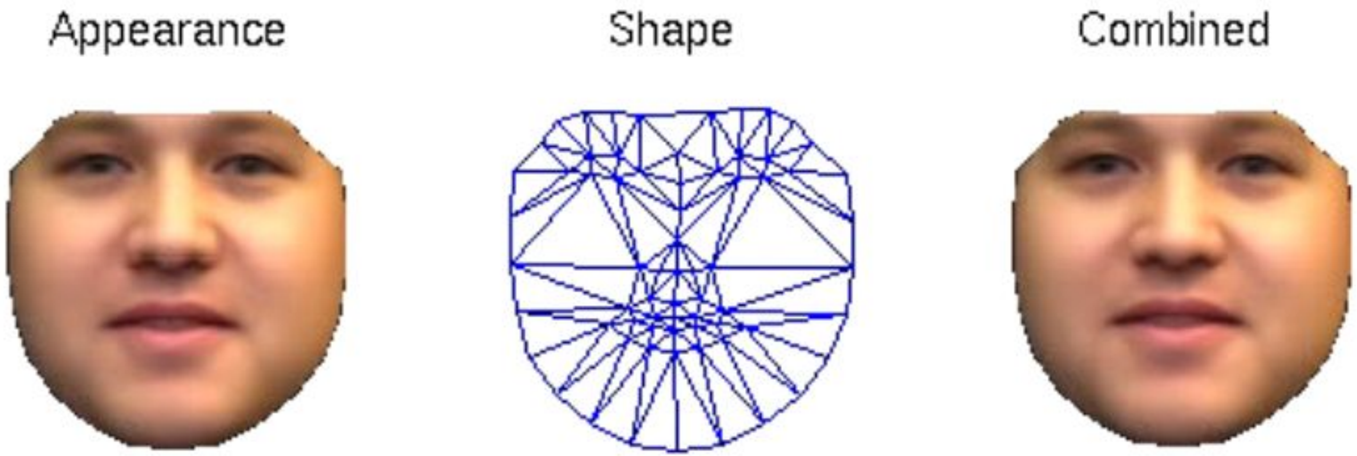
45

Active Appearance Models (AAMs)

- A way of matching a statistical model of shape and appearance to facial images.
- First, the model is built by combining a model of shape variation with a model of texture variation.
- **Shape model**
  - Built by applying **Procrustes Analysis** to align the set of landmark points marked in the images and **PCA** to generate a set of eigenvectors and their corresponding eigenvalues.
- **Appearance model**
  - Trained by warping each training image to the mean shape and calculating a mean appearance model along with appearance eigenvectors.

46

AAMs ctd.

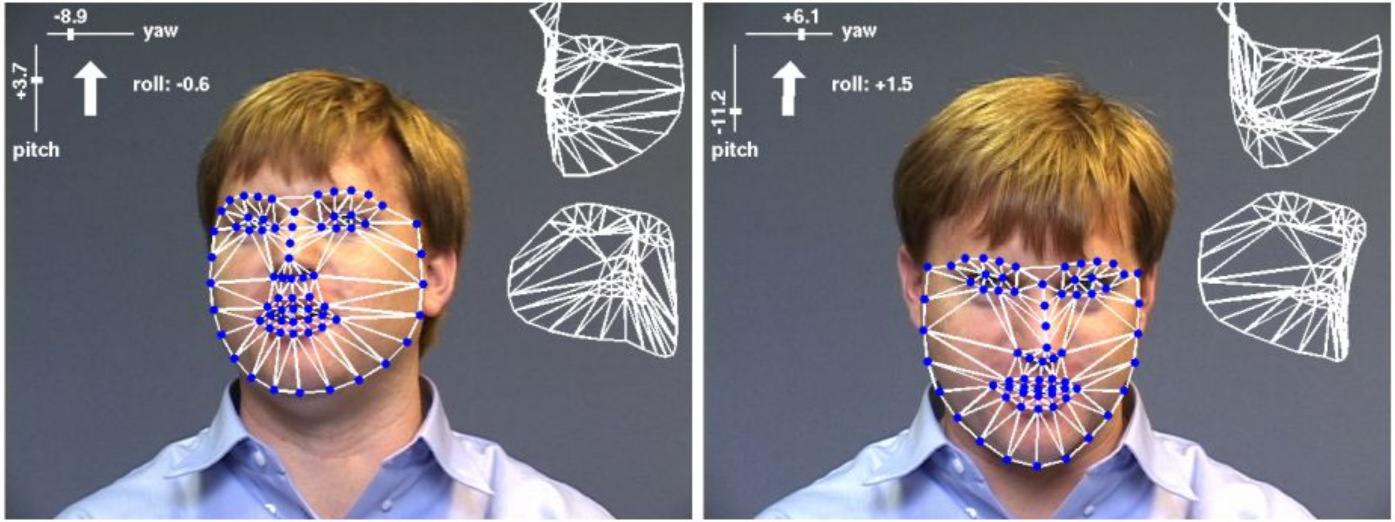


AAM Shape and Appearance models [22]

- Then for fitting the model, various algorithms including **regression**, **classification** and **nonlinear optimization methods** have been proposed.

47

Active Appearance Models (AAMs)



AAM Facial Fitting [22]

- In the original paper of AAM proposed by Cootes et al. [23], it follows an iterative matching algorithm by learning the perturbations in the model parameters and the induced image errors to calculate the shape and appearance parameters.

4

Constrained Local Models (CLMs)

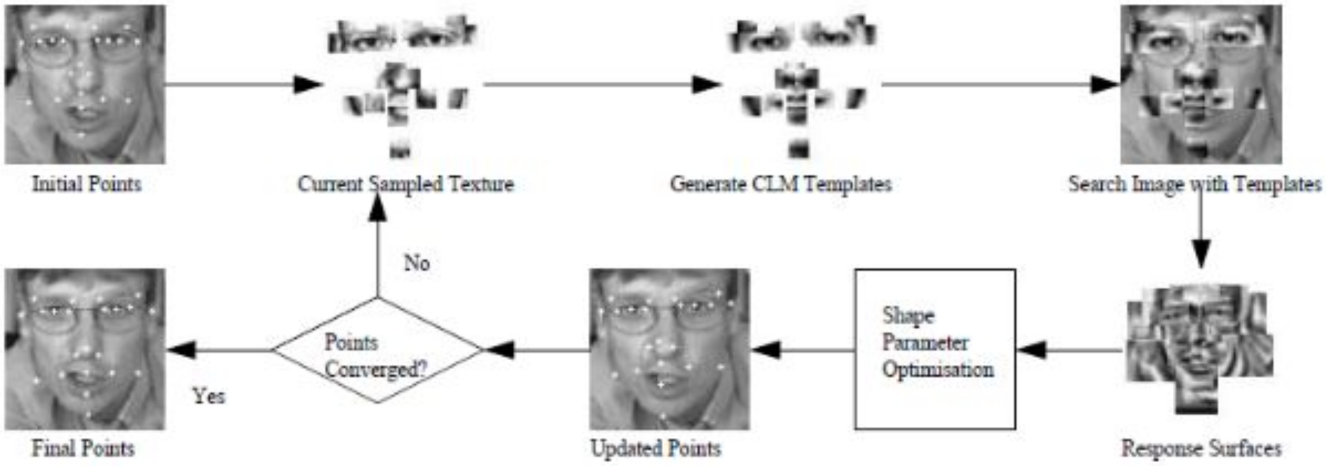
8

- Uses a joint shape and texture appearance model to generate a set of region template detectors.
- The appearance model used here is quite similar to that used in AAM.
  - But instead of using a global appearance model built by triangulation of the shape model of the face, CLM uses the variation of appearance of a set of template regions or patches surrounding individual features.
- **Fitting procedure:**
  - Accurate initialization of the facial feature points.
  - Generate a set of region templates around the points.
  - Apply the templates to the image and generate a set of response image patches for each feature point.

49

CLMs ctd.

- Find the shape parameters so as to maximise this sum of responses.



CLM Search [22]

50

Deformable 3D face models

- Fitting of a generic, deformable 3D face model on facial images.
- Detects features based on local image gradient analysis and backproject a deformable 3D face model through optimization of its deformation parameters.

Experimental results obtained for generic face model for images [23]

51

Deformable 3D face models

- Disadvantages arising from training based approaches:
  - Requiring the new images not to differ too much in illumination conditions, shape and appearance from the training images.
- This approach avoids the above disadvantages.
- Outperform many other face fitting alternatives under challenging illumination conditions and on devices with low hardware specifications.

Dynamic Templates and Re-Registration Techniques

- Recover the full motion of the head using a cylindrical head model created from an initial reference template of the head image and the corresponding head pose.
- 3 main steps[24]
  - Uses the **Iteratively Reweighted Least Squares (IRLS)** and image gradient to accommodate non-rigid motion and occlusion.
  - Dynamically updates the templates while tracking and maintain accurate tracking even when the face moves out of view of the camera.
  - Minimize error accumulation in the use of dynamic templates, this system re-register images to a reference template.
- 98% accuracy in automatic blink recognition.

53

Head Pose Estimation

54

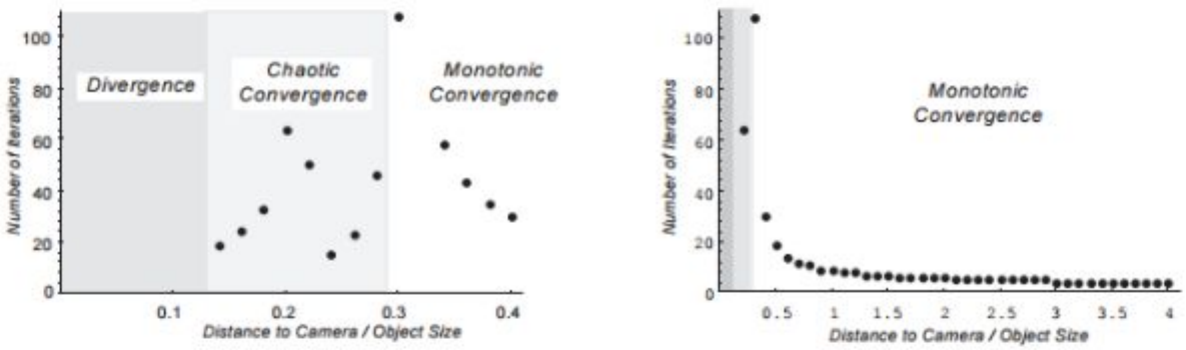
POSIT Algorithm

- **P**ose from **O**rthography and **S**caling with **IT**erations.
- For estimation of the 3D pose of an object from a 2D image.
- Requires four or more detected features in the 2D image with known relative geometry of the object which they represent.
- Combines 2 algorithms[25]
  - **P**ose from **O**rthography and **S**caling **(POS)**
    - Approximates the perspective projection with a scaled orthographic projection.
  - **POSIT**
    - Uses approximate pose found by POS to compute better scaled orthographic projections of the feature points.

55

POSIT Algorithm ctd.

- **Advantages**[25]
  - Converges to accurate pose measurements in a few iterations with low computational cost.
  - Does not require starting from an initial guess.
  - Computes the pose using fewer floating point operations. So can be considered useful for real time pose estimation.



Number of iterations for POSIT as a function of distance to camera[25].

56

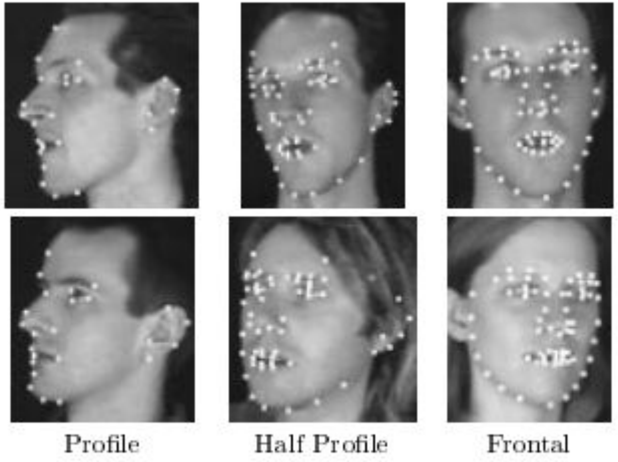
SFM

- Non-rigid **S**tructure **F**rom **M**otion and Point Correspondence.
- **Disadvantages** in classical model based head pose estimation methods[26].
  - Always requiring a 3D head model or a reference frame to estimate the pose.
  - Camera requiring to be calibrated in advance.
  - The difficulty of the models to deal with non-rigid motion of human faces.
- Utilizes a 2D AAM in each video frame to track the changing pose and expressions of the human face in 2D image frames.
- Accurately recover the pose using robust statistics **(RANSAC)** and point correspondences.

57

View Based AAMs

- Capture the full 180^0 rotation using only 5 models, roughly centered at viewpoints at 90^0, -45^0, 0^0, 45^0 and 90^0.



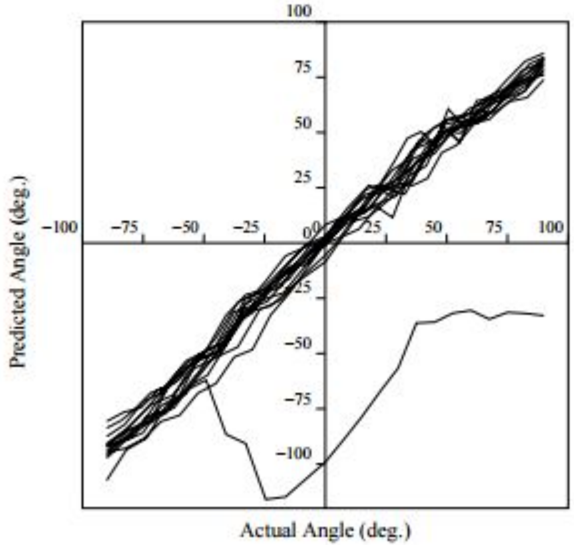
Examples from the training sets for the models [27]

- By using the AAM algorithm, we can match any of these individual models to new facial images.

58

View Based AAMs ctd.

- Tested tracking performance on 15 new video sequences of people, each comprising of 20 to 30 video frames
- 3 frames per second on a 450MHz Pentium III machine.



Comparison of angle derived from AAM tracking with actual angle [27]

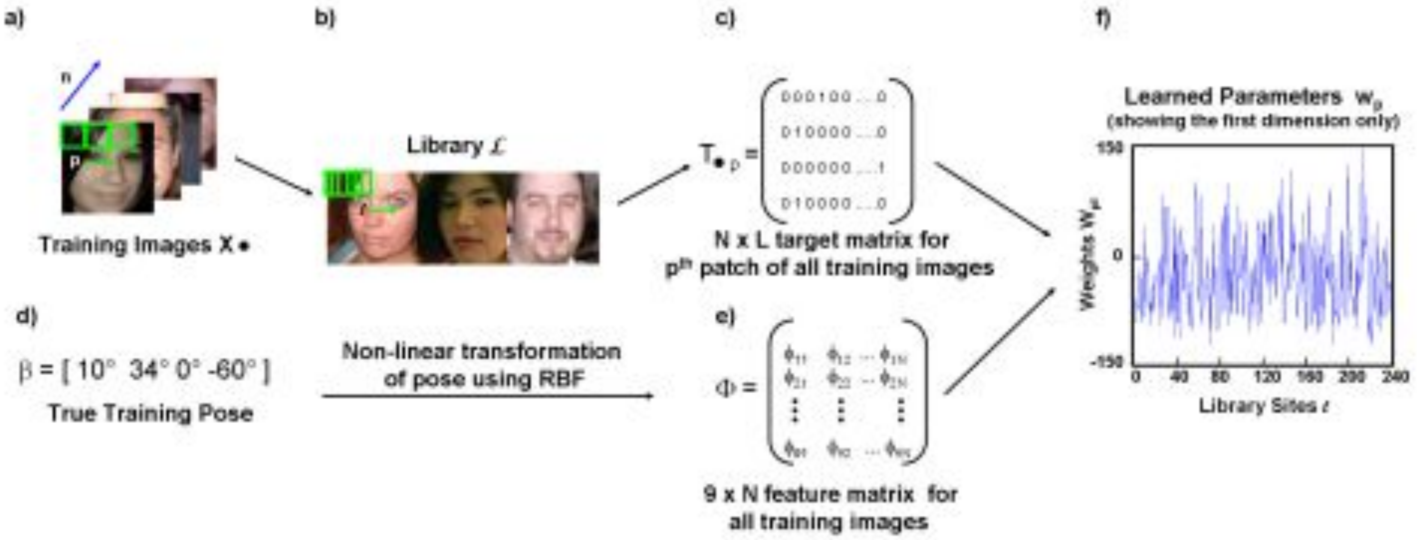
59

Patch based pose estimation with continuous regression

- Breaks the test image into a non overlapping regular grid of patches where each is treated separately.
- Contains a predefined library of object instances, which can be considered as a palette from which image patches can be taken.
- The test image patch is approximated by a patch from the library, which can be thought of as having a different affinity with each pose.
- The relative affinity of the selected patch from the library for each pose is used to determine a posterior probability over the pose[29].

60

Patch based pose estimation with continuous regression



Overview of system [28]

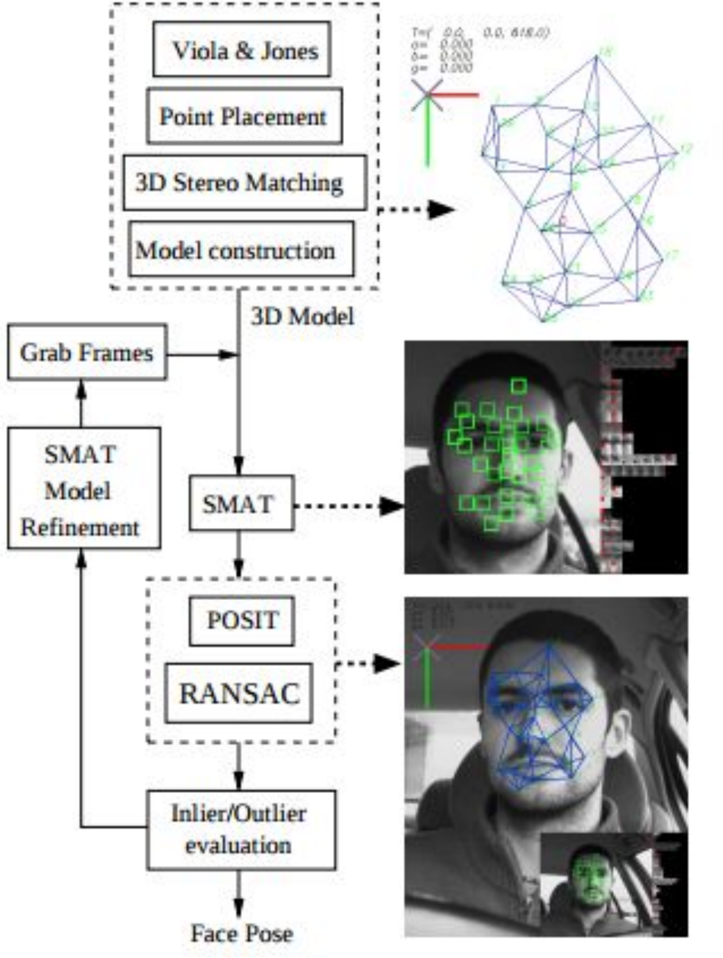
61

Automatic 3D model construction using SMAT, RANSAC and POSIT

- Used as a basis of a driver monitoring system. A face model is initialised and constructed on line.
- When the face comes frontal to the camera, a fixed point distribution is superposed over the face, then several points close to those locations are selected for tracking[29].
- The 2D projections of the model points are tracked separately for left and right images using SMAT model refinement.
- Incorrectly tracked points are rejected using RANSAC and finally the 3D pose is recovered from the 2D points set using POSIT.
- Head rotations up to $\pm 45^{\circ}$ are correctly estimated in real time.

62

Automatic 3D model construction using SMAT, RANSAC and POSIT



Overview of system [29]

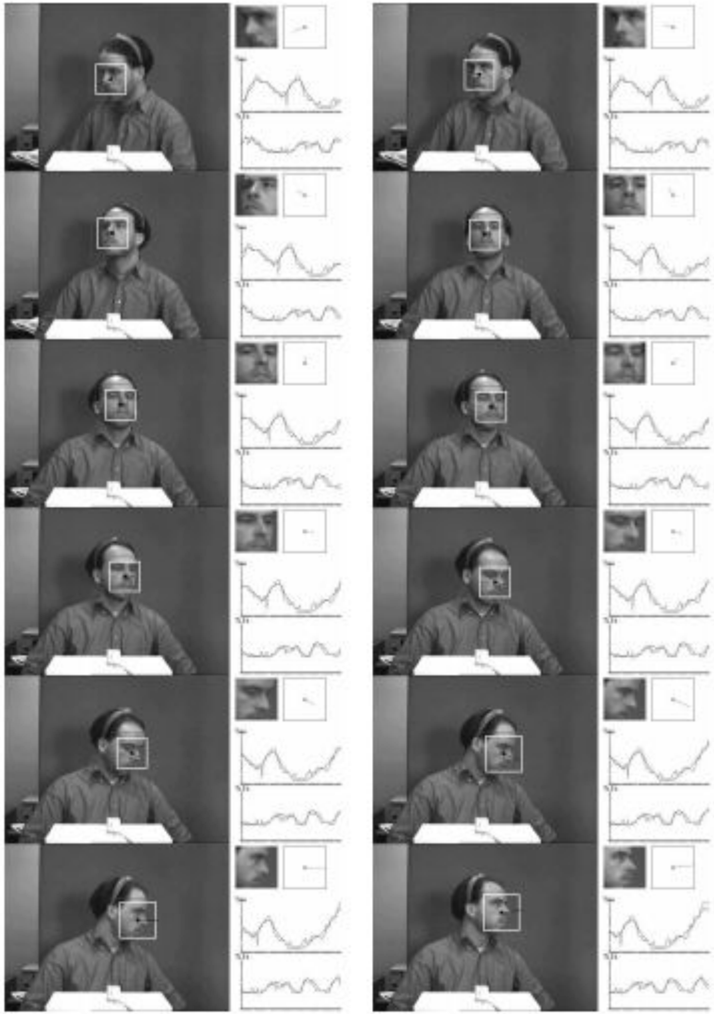
63

Support Vector Machines (SVMs)

- Generalises and transforms a generic 2D facial appearance model across the view sphere, to investigate face detection and pose estimation efficiently[30].
- Performs automatic feature extraction by constructing complex nonlinear decision boundaries for learning the distribution of a given data set using Structural Risk Minimisation (SRM).
- Pose estimation is automatically performed by Gaussian kernel functions.

64

SVMs ctd.



Test Results[30]

Pose estimation from prototype images

- Jamie Sherrah, Shaogang Gong and Eng-Jon Ong have developed a view based statistical learning technique based on the similarity of images to prototype real time identity independent head pose estimation from a single 2D view[31].
- In order to build the appearance model example, views labelled with 3D pose angles are required.
- Estimates an average face template at each pose.
- Uses Orientation selective Gabor filters to detect features for pose discrimination.
- Use PCA for dimensionality reduction and it was found to provide invariance to identity while accurately describing pose changes.

66

Pose estimation from prototype images ctd.

An example labelled head image set. The image of labelled views are from +90 to -90 in Yaw and +30 to -30 in tilt at 10° intervals[31]

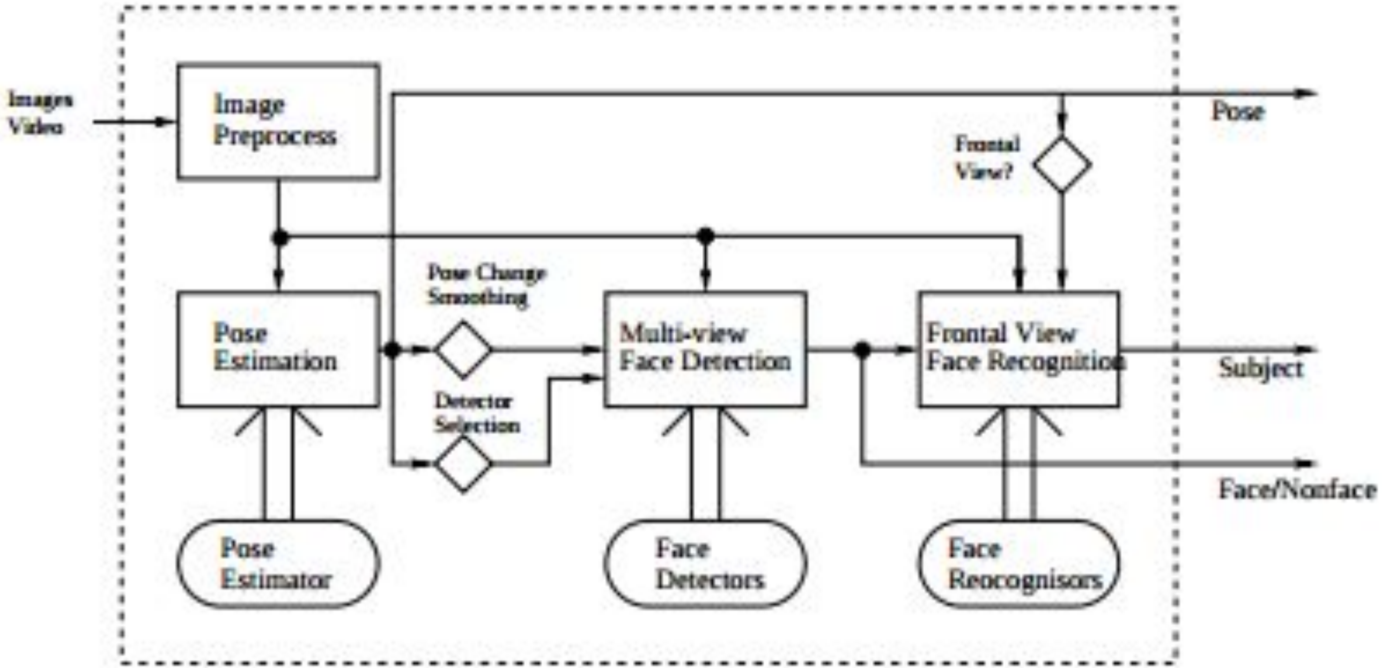
67

Support Vector Regression (SVR) and classification based pose estimation

- Uses SVR to construct pose estimators.
- Trains a set of **Support Vector Classification (SVC)** classifiers for multi-view face detection.
- Uses pose information to select face detectors to improve the detection accuracy.
- Uses pose change smoothing technique for further computational reduction.
- Activates an SVC multiclass classifier to perform face recognition when faces are in frontal view.
- All the above are integrated in an SVM based framework.
- Detection rate of above 95%, recognition accuracy of above 90%, average pose estimation error around 10^0[32].

68

Support Vector Regression (SVR) and classification based pose estimation ctd.



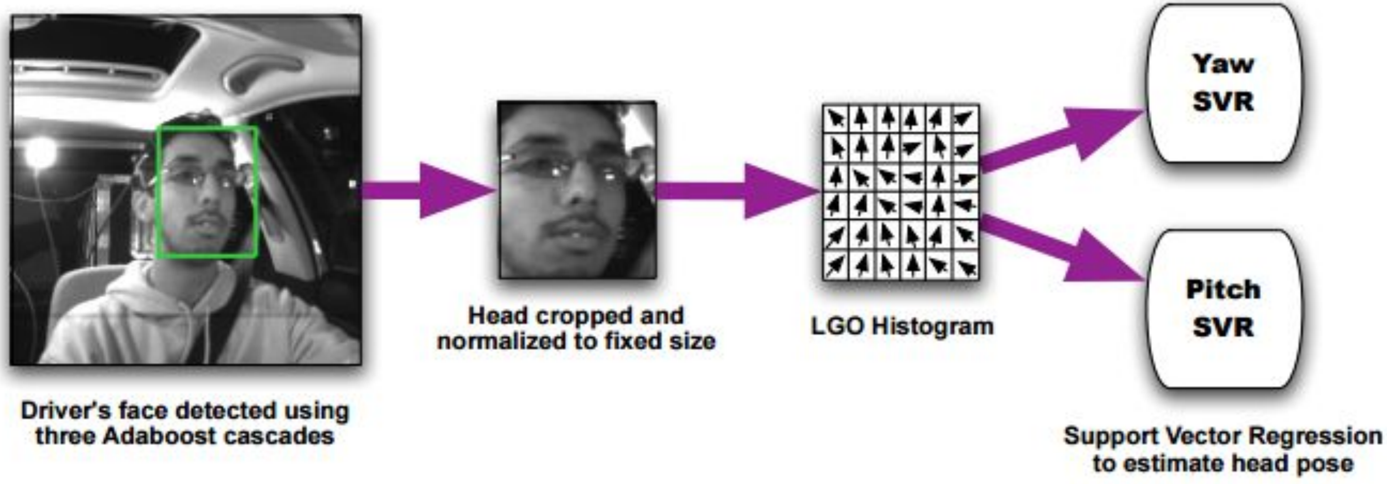
Multi-view face detection, recognition and pose estimation[32]

SVR with Localized Gradient Histograms

- Erik Murphy-Chutorian, Anup Doshi, and Mohan Manubhai Trivedi have developed an identity and lighting invariant system to estimate a driver's head pose[33].
- Consists of following steps.
  - Facial regions are found by three cascaded Adaboost face detectors applied to the grayscale video images.
  - The detected facial region is scale normalized to a fixed size and used to compute a Localized Gradient Orientation histogram.
  - The histogram is passed to two SVRs trained for head pitch and yaw.

70

SVR with Localized Gradient Histograms ctd.

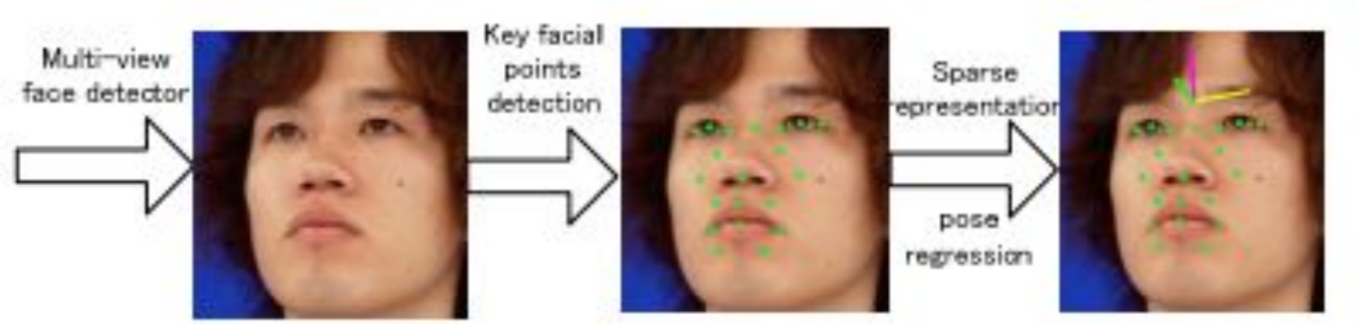


Overview of the proposed Head Pose Estimation System[33]

71

SVR with sparse representations

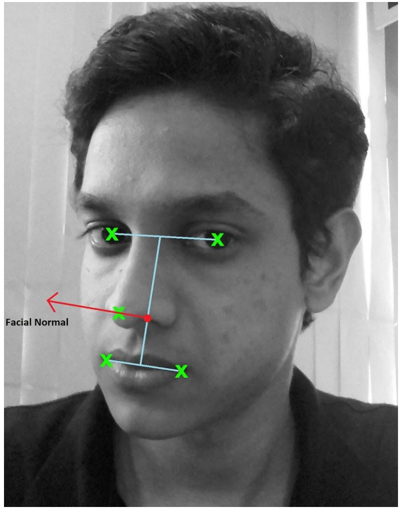
- Develops a high performance JGCF pose estimation system based on sparse Bayesian regression technique, which is also known as **Relevance Vector Machine (RVM)** and sparse representation of facial patterns[34].
- Facial properties are represented using sparse features of 20 key localized facial points.
- RVM is used to find the relationship between the sparse representation and the yaw and the pitch angles.



Flow chart of the Pose Estimation System[34]

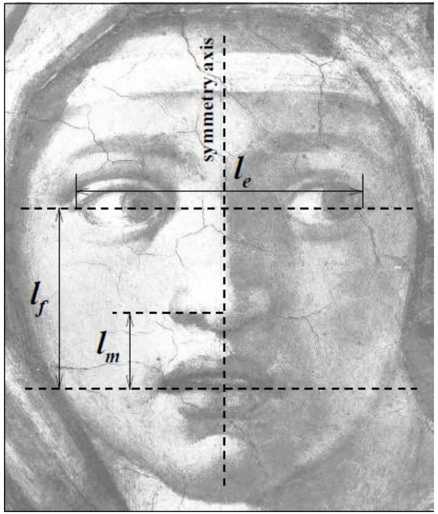
Geometric Head Pose Estimation

- Estimating the head pose or the gaze direction based on the geometric positioning of features extracted from a human face.
- Cipolla and Gee [35] suggests that once the positions of the two eyes, two corners of the mouth and the nose tip is correctly found, the facial normal can be easily computed from them.



Geometric Head Pose Estimation ctd.

- Assumption - In every human being, the ratio between lengths l_m and l_f, stays roughly the same.



- Consumes very low computation time and can give real-time results on mobile platforms.
- **Issue:** Necessary to have both the eyes visible in the given image.

74

Neural networks

- Example system[36] used two neural networks which have been trained to approximate the functions that map an image of a head to the orientation of the head.
- Combines head tracking and localization with a neural network based head orientation estimation system.
- The head localization subsystem detects and extracts the image region which corresponds to a human head.
- This head region is then applied as an input to a neural network system for head orientation estimation.
- An example system of an Artificial Neural Network[37] had used three neural networks; first for color segmentation, second for localization of the face, and third for final recognition of the head orientation.

75

Gabor Wavelet Networks

- The use of Gabor filters enables the coding of geometrical and textural object features.
- Gabor filters can be used as a model for object features to ensure data reduction and allow any desired precision of the object representation ranging from sparse to photo realistic representation.
- The example work has encoded feature information in the wavelet coefficients.
- Then an artificial neural network has been used to compute the head pose from the wavelet coefficients.
- GWNs have been used successfully for wavelet based affine real time face tracking and pose invariant face recognition[38].

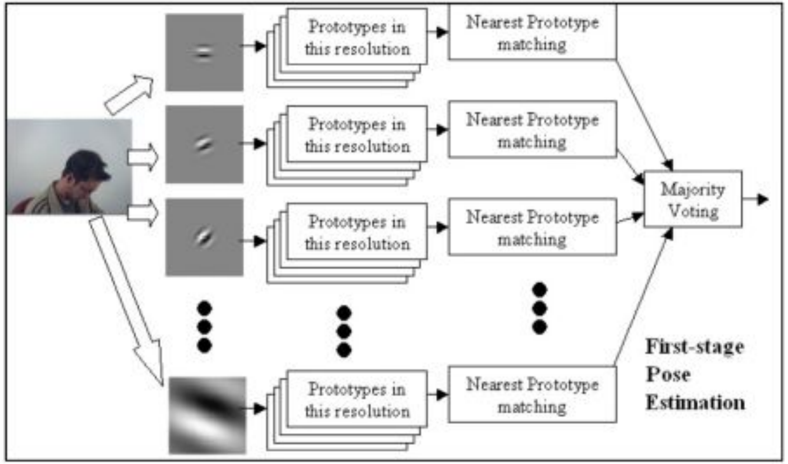
Convolutional networks

- An example system[39] have used a convolutional network which integrates detection and pose estimation.
- This network is capable of mapping images of faces to points on a low dimensional manifold, parametrized by pose.
- The images of non faces have been mapped to points far away from that manifold.
- The network is trained by optimizing a loss function of three variables called image, pose, and face/non-face label.
- When these three variables match, the energy function is trained to have a small value. When they do not match, it is trained to have a large value.

77

Subspace analysis with the topography method

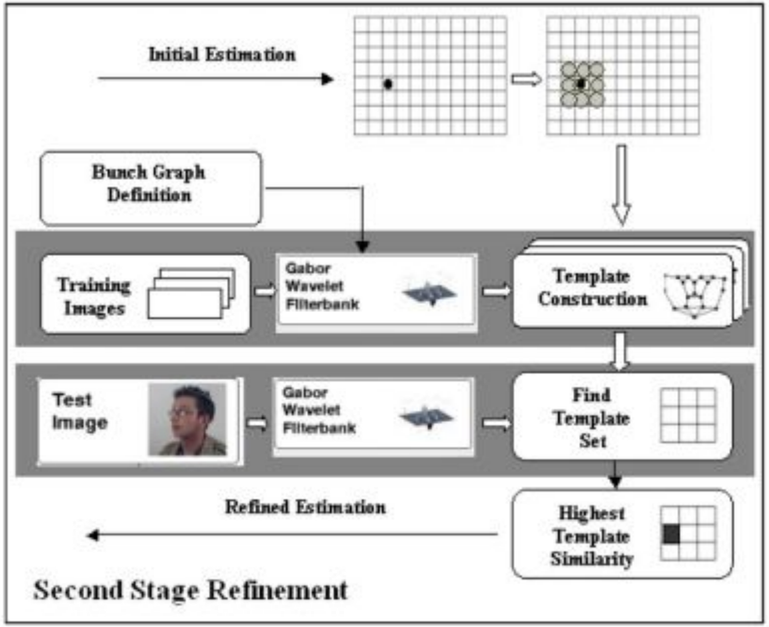
- Junwen Wu and Mohan M. Trivedi have proposed a two-stage approach for head pose estimation by combining subspace analysis together with the topography method.
- The first stage is based on the subspace analysis of Gabor wavelets responses. The pose estimation is taken by nearest prototype matching with Euclidean distance[40].



78

Subspace analysis with the topography method ctd.

- In the second stage, the pose estimate obtained from first stage is refined by analyzing finer geometrical structure details captured by bunch graphs.



Hybrid approaches for head pose estimation

- Yuxiao Hu et al. proposed a robust pose estimation approach combining facial appearance asymmetry and 3D geometry.
- Junwen Wu et al. presented a two-level approach for estimating face pose from a single static image. In the first level, an estimate of the pose is derived within some uncertainty. In the second level, the output from the first level is systematically processed to minimize its uncertainty by analyzing finer structural details captured by the bunch graphs.

Hybrid approaches for head pose estimation ctd.

- Youding Zhu and Kikuo Fujimura came up with an adaptive head pose estimation method for driver attention monitoring. PCA and 3D motion estimation were used for head pose estimation from an image sequence.
- Louis-Philippe Morency et al. presented a method for estimating the absolute pose of a rigid object based on intensity and depth view-based eigenspaces which are built across multiple views of example objects of the same class.

81

Existing APIs & SDKs

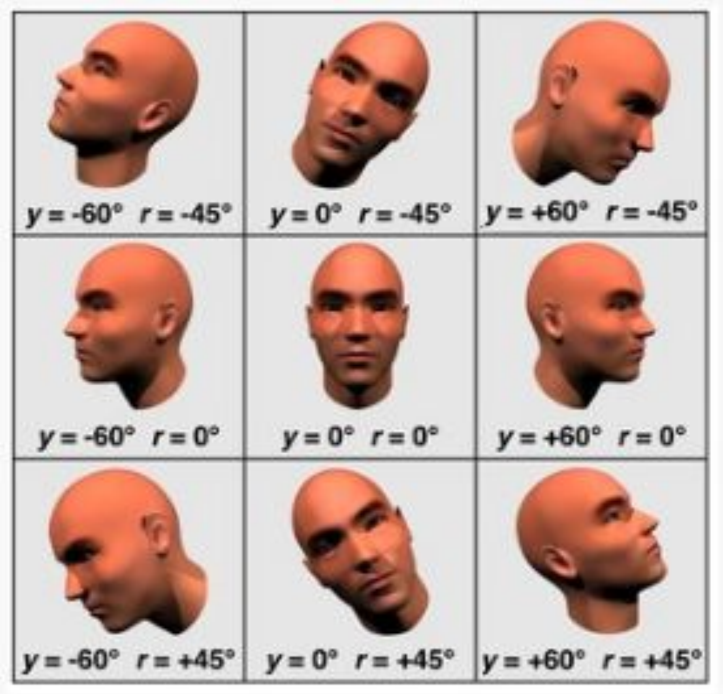
82

Google Mobile Vision API

- **Google Mobile Vision API** was developed as a way to facilitate finding objects in photos and video, using real-time on-device technology.
- Its face API has the capability to find human faces in photos, videos or live streams.
- It also has the ability to track facial landmarks such as the eyes, nose and the mouth
- This Vision API can be integrated to mobile applications running on Android 2.3 (Gingerbread) or higher

83

Google Mobile Vision API Ctd.

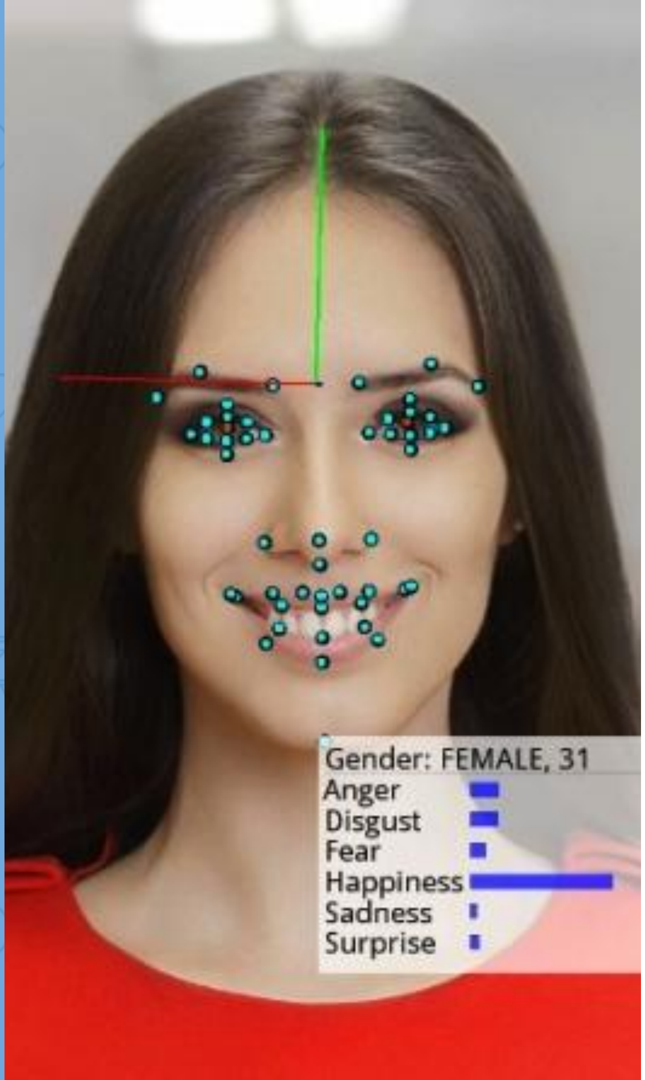


Pose angles determined by Google Mobile Vision API where y = yaw and r = roll

Example facial landmarks identified and tracked by Google Mobile Vision API

8

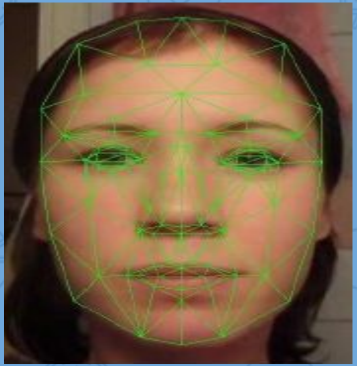
Visage Technologies visage|SDK™



- **visage|SDK™** is a software package that could be utilized to develop applications that involve interactive animation, avatar control, sophisticated user interaction and other advanced features in games, arts and other applications.
- The software provides accurate tracking of head pose, gaze direction, facial actions and eye closure.

85

Visage Technologies visage|SDK™



- visage|SDK™ has several components such as,
  - **FaceTrack:** FaceTrack package is a very powerful, fully configurable face tracking toolkit.
  - **HeadTrack:** HeadTrack engine can be used for applications that require comparatively high performance and robust 3D head pose tracking in a video sequence.
  - **FaceDetect:** FaceDetect engine performs face detection by identifying facial features in images which have one or more human faces.
  - **FaceAnalysis:** FaceAnalysis package utilizes different machine learning algorithms to provide gender estimation as to male or female, probability of various emotions and approximate age estimation.

86

OpenCV (Open Source Computer Vision Library)

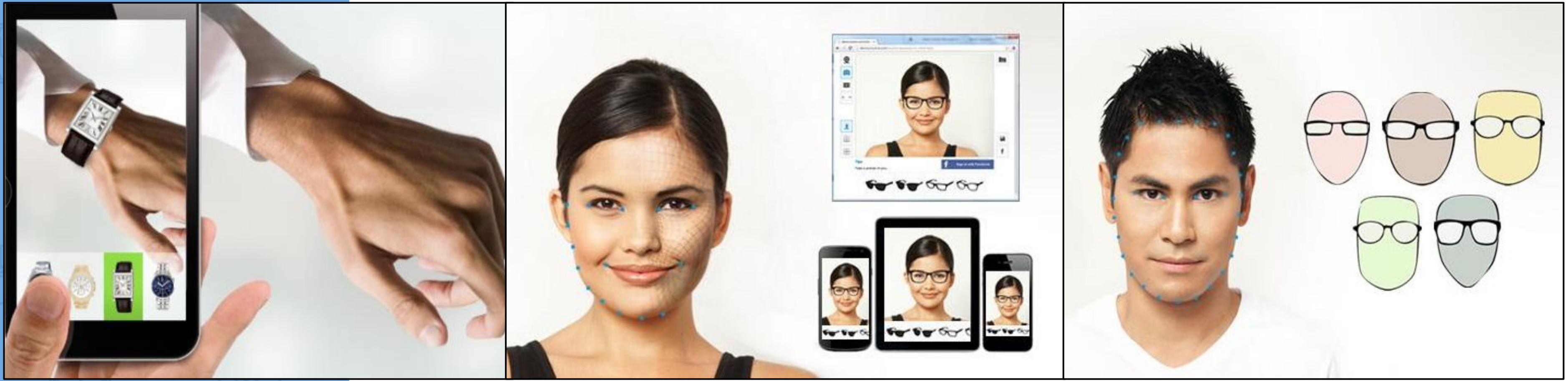
- OpenCV is an open source computer vision and machine learning software library which provides interfaces for **C++**, **C**, **Python**, **Java** and **MATLAB**. It supports **Windows**, **Linux**, **Android** and **Mac OS**.
- OpenCV library can be used for many applications such as detecting and recognizing faces, identifying objects, tracking camera movements, tracking moving objects and many more.

Existing applications

88

TryLive

- Provides a platform for retailers and e-commerce merchants to take their showrooms to the customer's home.
- Provides intuitive try-on experience for a range of products such as eyewear, apparel, watches and furniture.



89

Masquerade

- The showcase mobile app for Masquerade face tracking and 3D face placement SDK
- Allows recording video selfie animations and taking selfie photographs by changing the way you look.
- Has a library of different types of masks that can be tried out.
- Has a proprietary self learning face tracking algorithm and a framework for creating special effects.
- Can be incorporated into videoconferencing platforms.

90

Masquerade

Different masks of Masquerade

91

Snapchat

snapchat

- An image messaging application software product that consists of a feature to create special graphical effects to user's face.
- **"Geofilters"**
  - Allows special graphical overlays to be available if the user is within a certain geographical location, such as a city, event, or destination.
- **"Selfie lens"**
  - Allows users to add real-time effects using face detection into their snaps activated by long-pressing on a face within the viewfinder.

92

Snapchat ctd.

Snapchat daily lenses

SpeculAR

94

Requirement Specification

95

Functional Requirements

- Should be able to view the augmented model of the spectacles or accessory on oneself on different perspectives.
- Should be able to select a desired model.
- Should be able to update the models.
- Should be able to make an order.
- Should be able to take a snapshot or record a video and share it on social networks.
- Should be able to switch back to backcam when necessary.

96

Usability Requirements

- **Easy Navigation:** The users should be able to easily find the essential features quickly. While they are using the app, they should not have trouble in finding exactly what they are looking.
- **Great Aesthetics:** The app should be aesthetically pleasing and it should have a good look and feel.
- **Platform Usability:** The application should work in different mobile platforms.
- **Training time:** Training time for a normal user should be less than 5 minutes and a power user should be able to become productive at using a certain operation within less than 1 minute.

97

Reliability Requirements

&

Supportability Requirements

- SpeculAR requires high state of reliability.
- Should operate without failure for any number of uses at any time.
- Need to exceed customer expectations. It should be more reliable than its competition.

- Should support its functionality in different kind of environments like indoor, outdoor etc.
- Should work fine in environments under different lightning conditions.

98

Performance Requirements

&

Design Constraints

- **Response time**: Quick response time is very important for a real time application. The head pose has to be calculated real time and the speed of interaction should be high.
- **Resource utilization**: Memory utilization by the application should be low due to the memory limitations in mobile devices.

- Low processing power and memory limitations in mobile devices.
- Sometimes the quality of the video stream captured by the front facing camera of mobile phones may be low.
- Different illumination conditions.

99 Personas

“Alice is a 31 year old woman who has been working as an Assistant Manager in a Bank. She has a Bachelor of Business Administration in Finance degree from University of Colombo. Also she has followed a certificate level computer course. She has a…“

“Grace Kalansooriya is a 65 years old mother of five children and a grandmother of seven. Her husband who had been a bank officer has passed away 5 years ago from a heart attack. She had been a teacher in Southlands College, Galle teaching English for students under Grade 5 for a period of 35 years ...“

Personas ctd.

“Nelika Erandi is doing a degree in Fashion Design and Product Development at University of Moratuwa. Her mother being a handcraft designer encouraged her to undertake this course and it is no wonder that she inherits her mother’s skills as a fortune to undertake this course very well. She is planning to …”

“Anuradha is a 25 year old married woman who works at a well reputed software development company. Her husband is a Lecturer at a well known university and drops her to work by his vehicle every morning on the way to his university. At work, Anuradha has a bunch of friends who have very much the same likes …”

Scenarios

“Alice bought a new office suit and she hope to wear it to the office next time. Alice likes to wear spectacles when going to work. She thinks that wearing spectacles will make her look more professional. Also she prefers to wear spectacles when working with the computer as she wants to protect her eyes. She search for eyewear in several e-commerce...”

“Grace has gone to the temple and forgotten where she has kept her pair of glasses. She searches for them everywhere but could not find them. Feeling miserable that it is the fourth time she lost her spectacles, she comes back home and relate the situation to her eldest son Kamal who is a Software Engineer…”

102 Scenarios ctd.

"Nelika has accidently sat on her reading glasses and has got them broken. One of the arms has broken off right at the hinge. Now Nelika wants to buy a new pair of reading glasses. She prefers to buy a fashionable one. Nelika knows that although some spectacles look very pretty, they do not match up to her face. She really wants to try them beforehand ..."

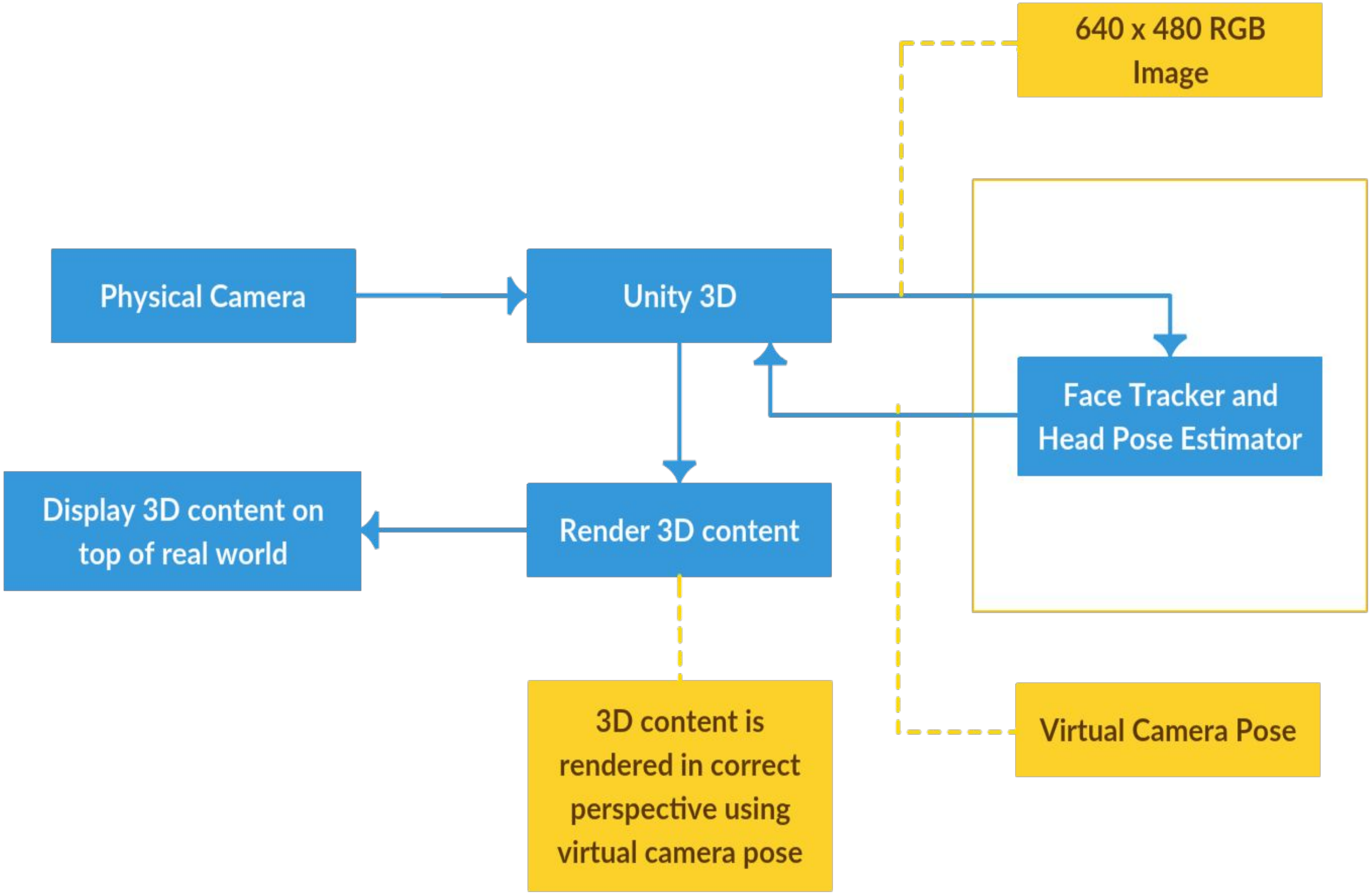
"Anuradha is at her office working and is thinking of whether or not she should go shopping that day. A while before she was supposed to be off, her husband calls in and tells her that he will be 2 hours late due to an urgent staff meeting that came up suddenly to discuss on some matter about the students at the university ..."

103

System Design

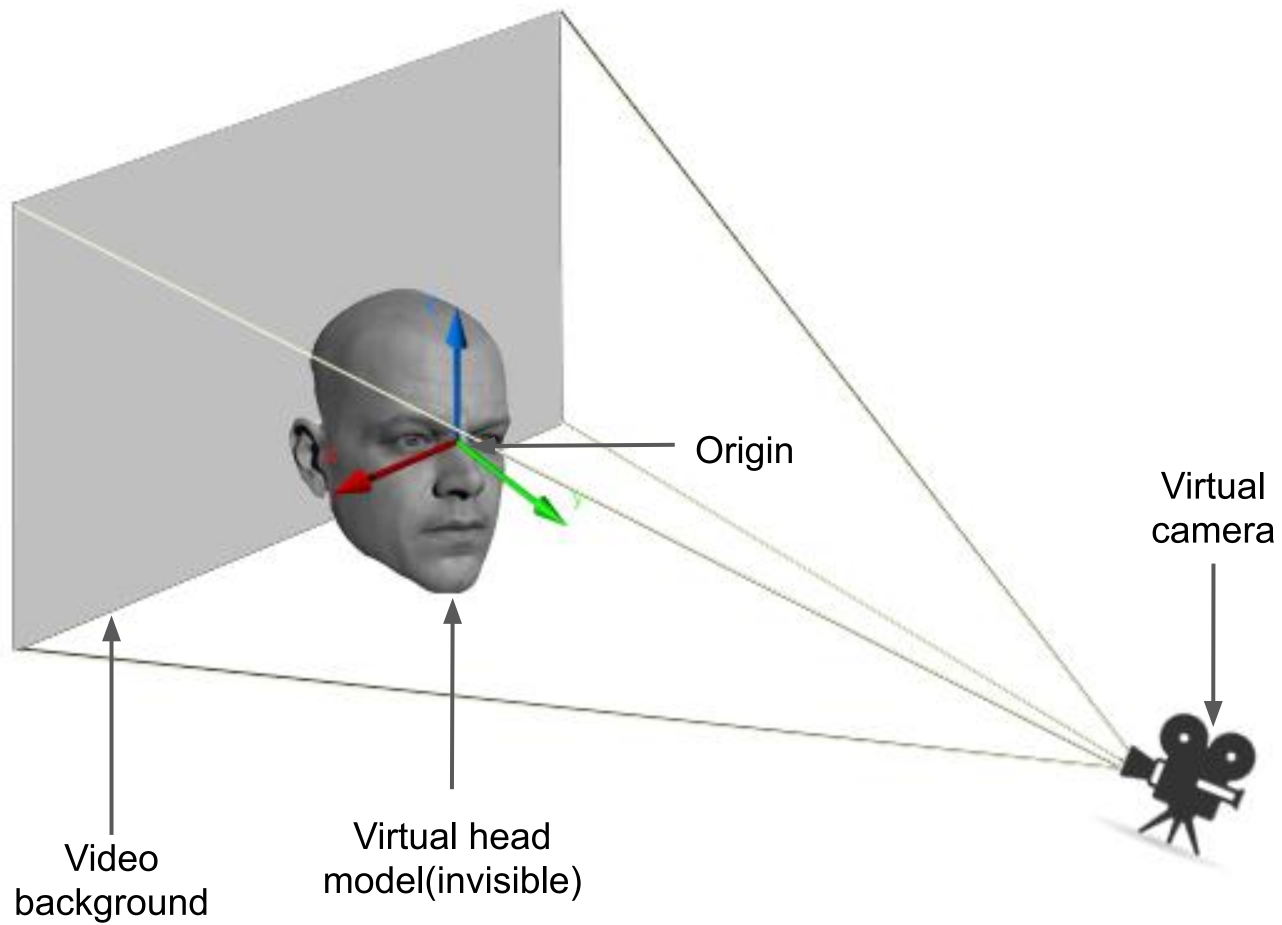
104

Architectural Representation



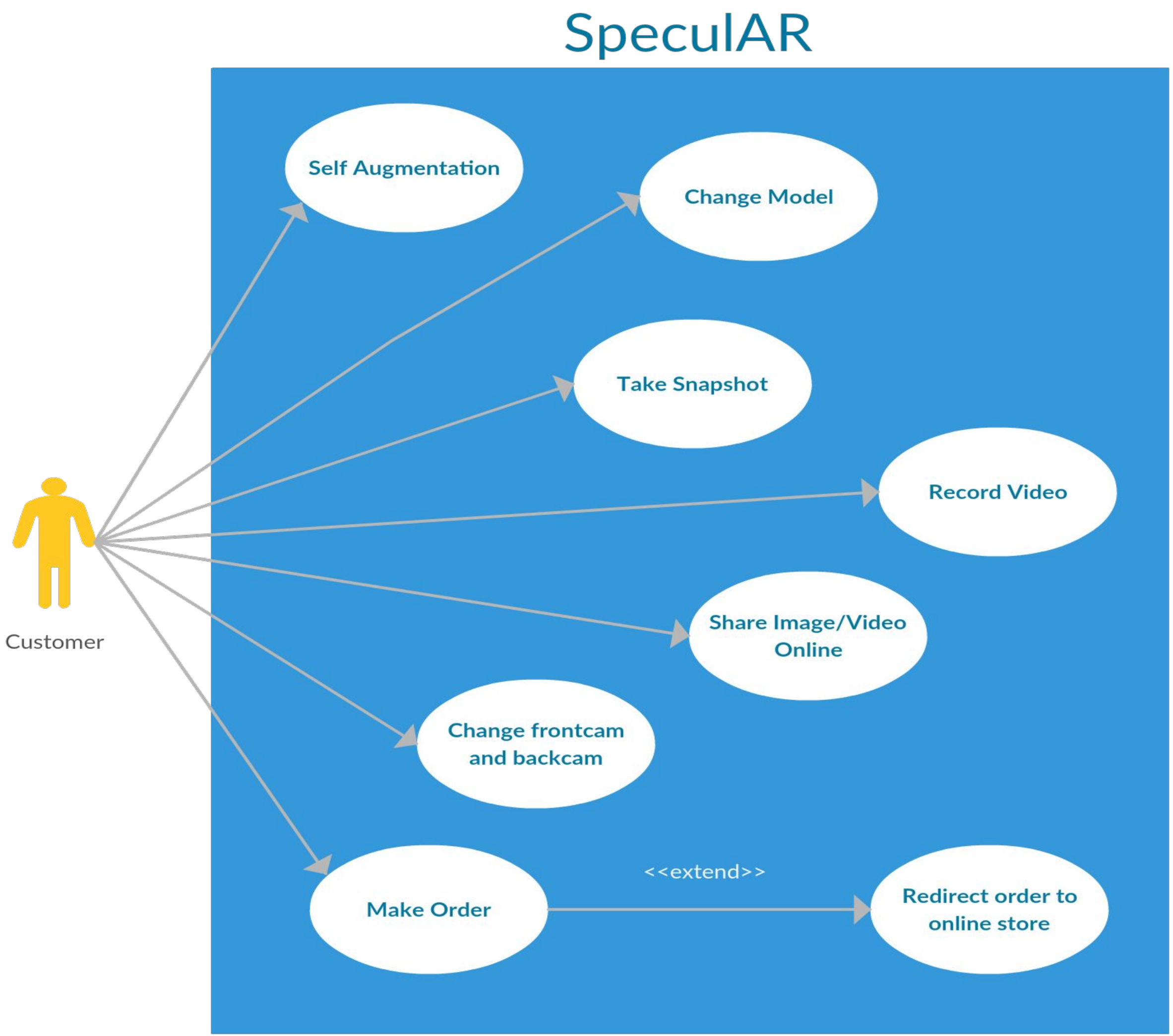
105

Implementation



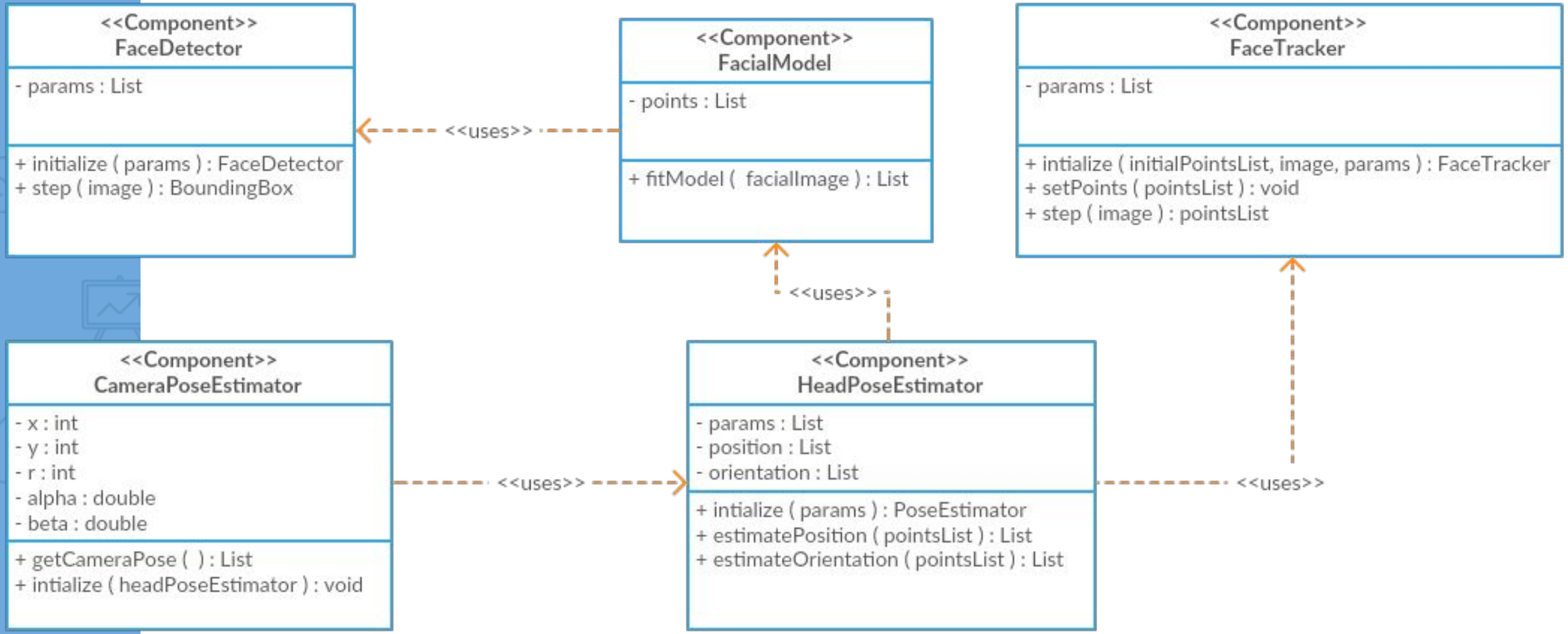
106

Use Case Realization



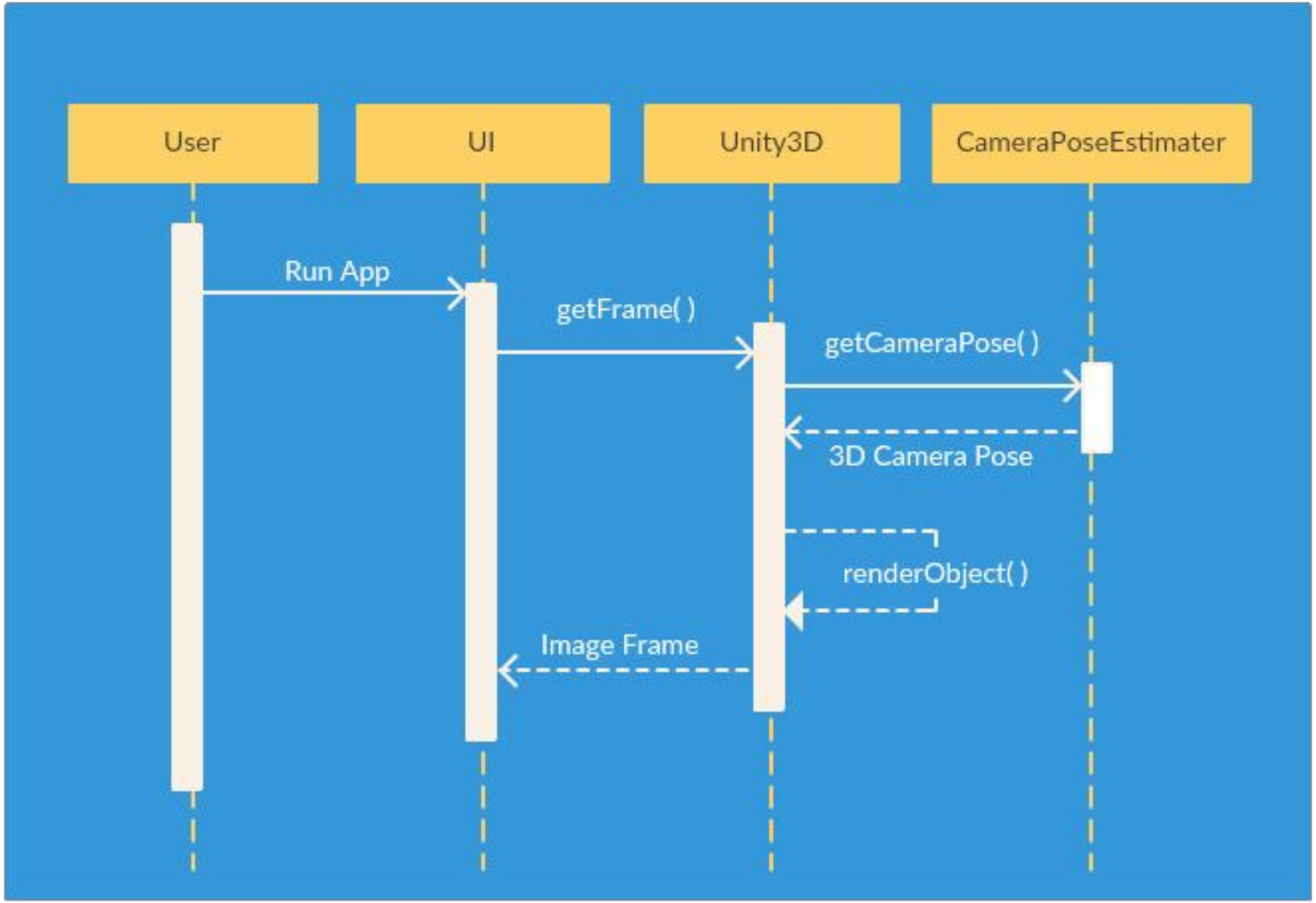
107

Class Diagram



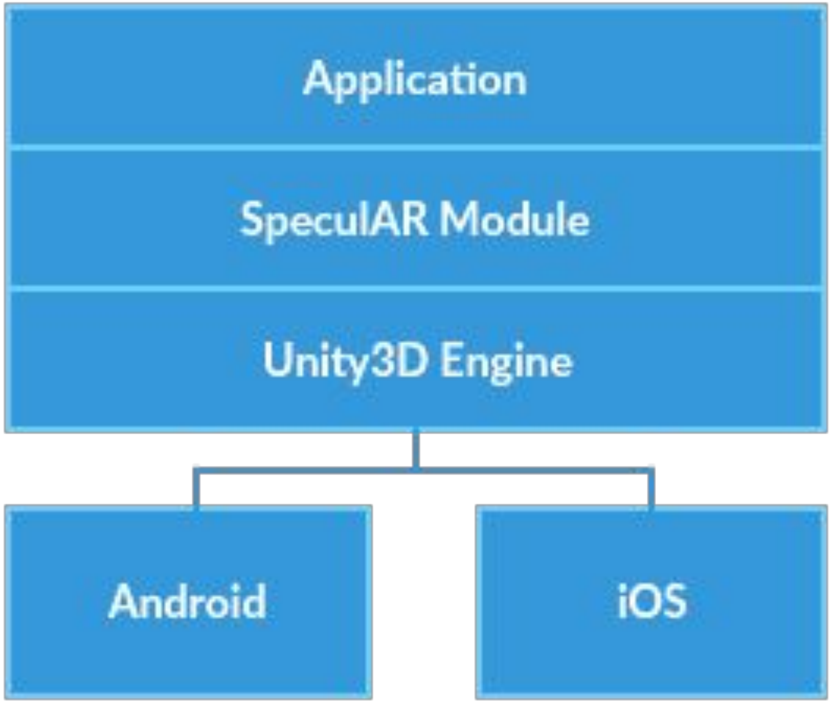
108

Basic Sequence Diagram



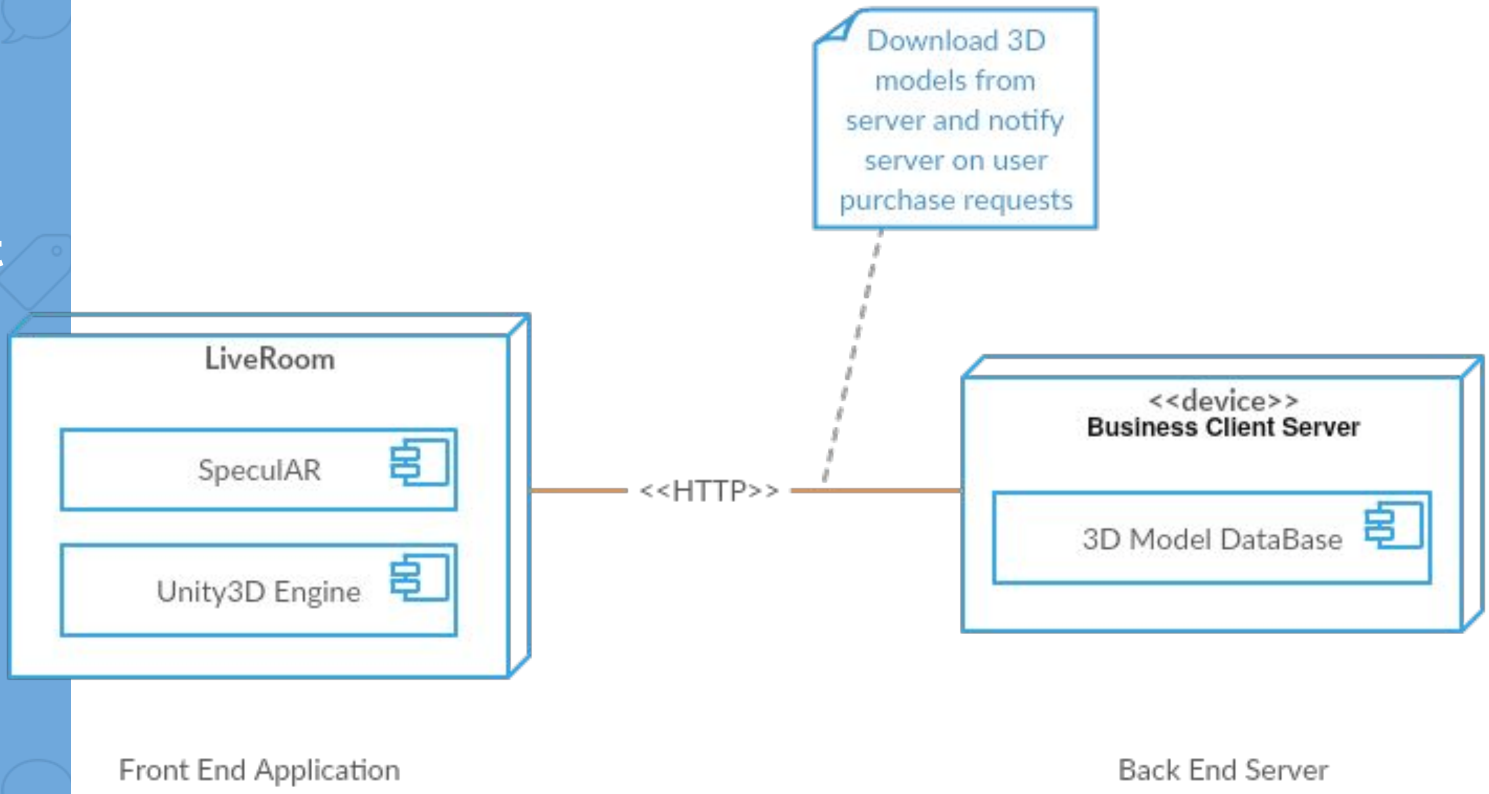
109

Layered View



110

Deployment Diagram



III

Proposed Solution

112

Proposed Solution

Stages	Percentage Completed (%)
1. Performing a comprehensive research on the work done so far in the field of face tracking and head pose estimation.	100%
2. Going through all the possible approaches in several iterations and narrowing down the list to a reasonable amount of approaches considering the extent they agree with our requirements.	100%
3. Implementing the selected approaches to be run first on a PC environment.	100%

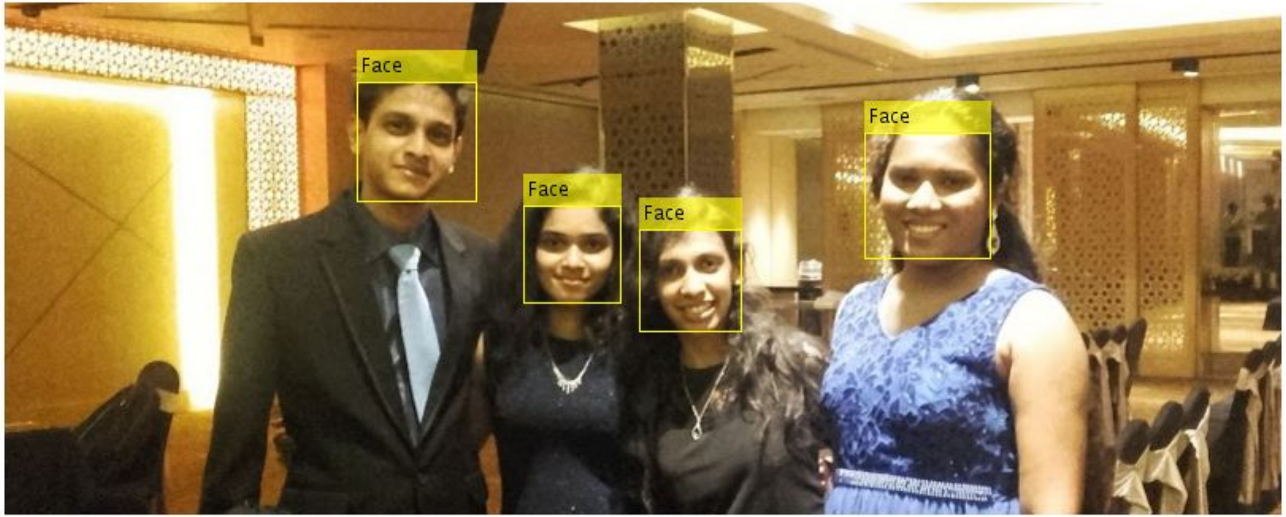
113

Implementations

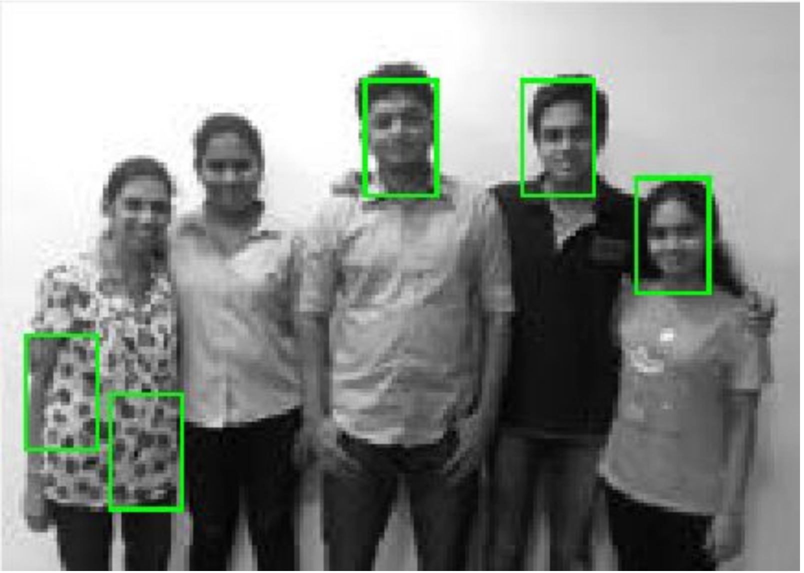
Selected approach implementations on a PC platform.

- **Viola-Jones Face Detection**
  - Tested Platform: PC Platform
  - Elapsed time: 0.3025 sec.
  - Results:

Detected faces



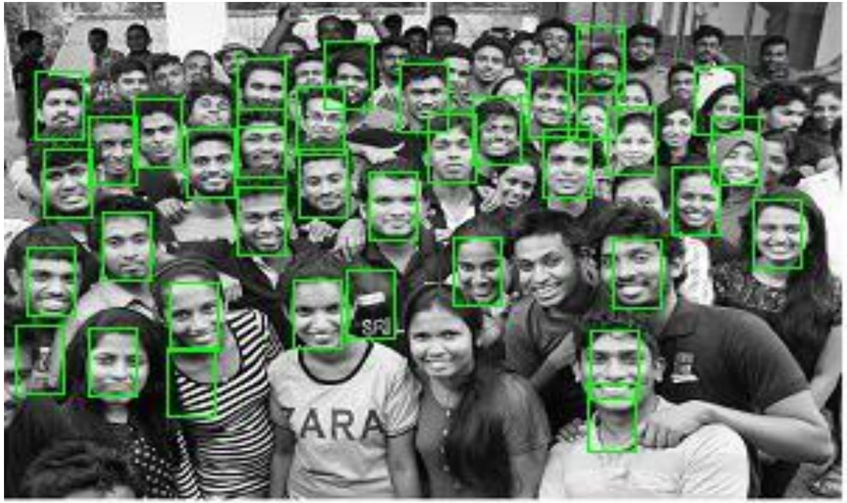
- **Neural Network based face detection**
  - Tested Platform: PC Platform
  - Elapsed time: 10.5820 sec.
  - Results:



116

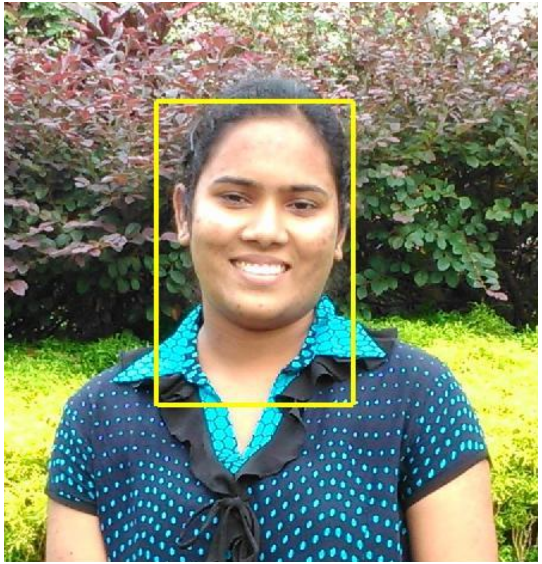
Face Detection Approaches

- **SVM based face detection**
  - Tested Platform: PC Platform
  - Elapsed time: 202.0102 sec.
  - Results:



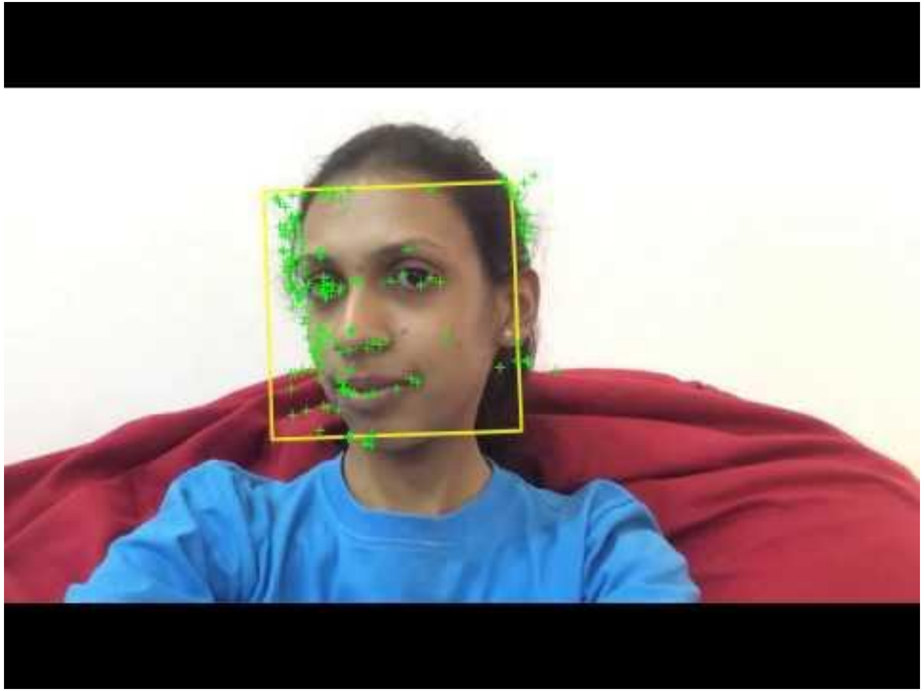
Face Detection Approaches

- **Color Based Face Detection**
  - Tested Platform: PC Platform
  - Elapsed time: 3.0996 sec.
  - Results:



Face Tracking Approaches

- **KLT Face Tracking**
  - Tested Platform: PC Platform
  - Tracking speed: 27.4562 frames/sec.

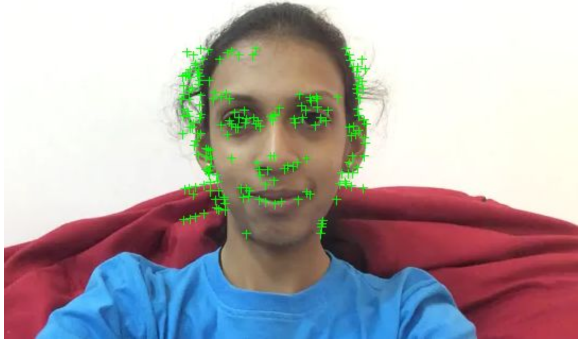


119

Face Tracking Approaches

RANSAC

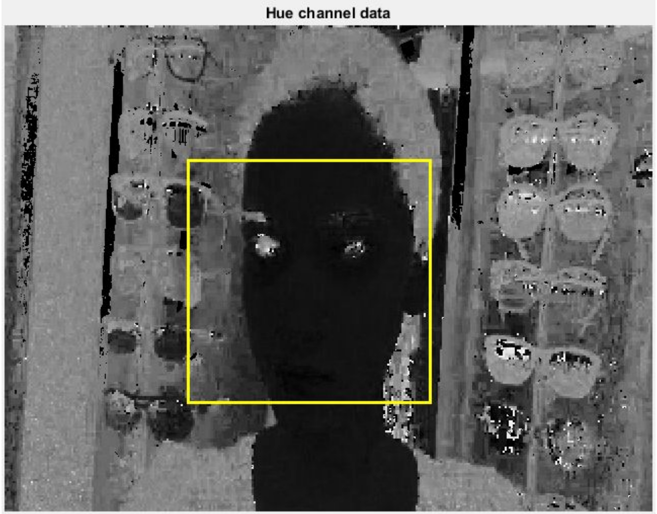
- Tested Platform: PC Platform
- Tracking speed: 5.0121 frames/sec.



Detected features

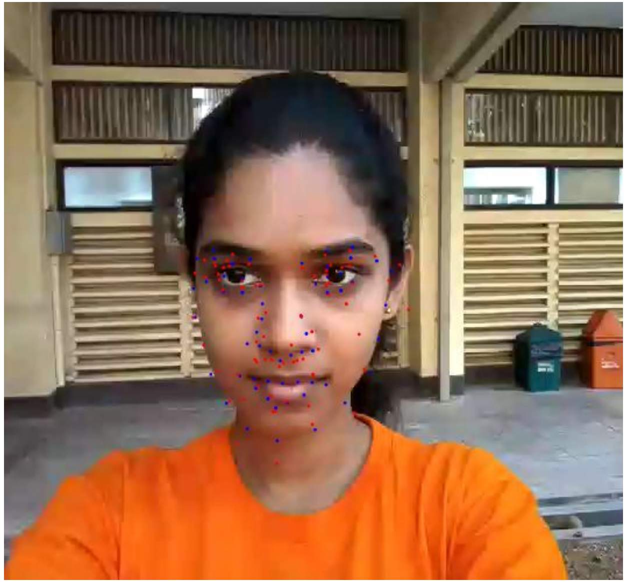
Face Tracking Approaches

- **Camshift Tracking**
  - Tested Platform: PC Platform
  - Tracking speed: 26.4563 frames/sec.



- **Active Shape Model (ASM)**
  - Trained on LFPW Database using 50 facial images.
  - Elapsed time: 31.6328 sec. (Poor initialization)
    - Using 55 landmark points with 4 contour points interpolated in between, 8 pixel search length and 40 search iterations.
  - Image fitting result:

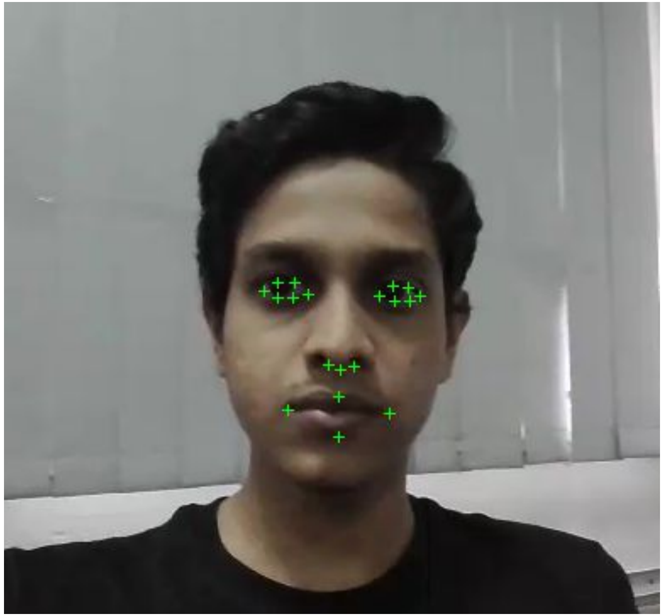
- **Active Shape Model (ASM)**
  - Trained on frames showing significant pose variation.
  - Elapsed time: 3.5309 sec.
    - Using 55 landmark points with 4 contour points interpolated in between, 20 pixel search length and 15 search iterations
  - Image fitting result:



123

Facial Fitting Approaches

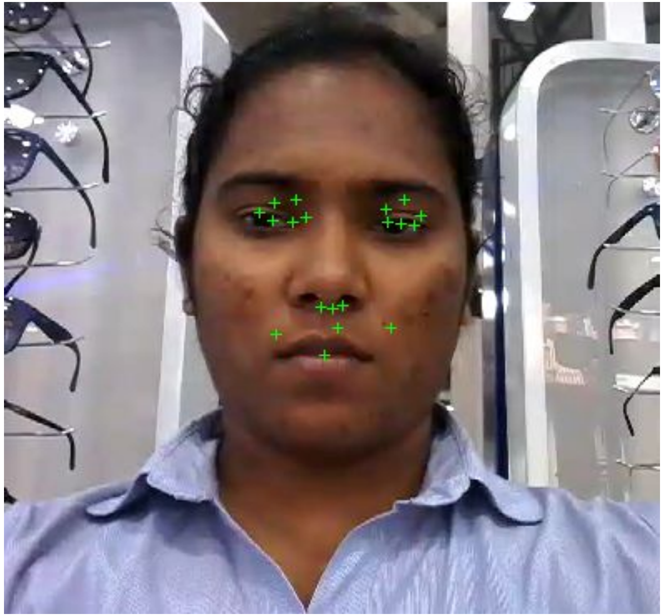
- **Active Appearance Model (AAM)**
  - Trained on LFPW database.
  - Elapsed time: 0.1383 sec.
    - With 19 points and 10 search iterations
  - Image fitting result:



124

Facial Fitting Approaches

- **Constrained Local Model (CLM)**
  - Trained on LFPW database.
  - Elapsed time: 1.3970 sec.
    - With 19 points, 16x16 pixel search area and 10 search iterations
  - Image fitting result:



Head Pose Estimation Approaches

- **Geometric Pose Estimation**
  - Elapsed time: 2.7095×10^{-4} sec.
    - To derive the pitch, yaw and roll angles from the 5 facial features detected in a single frame.
- **POSIT algorithm**
  - Elapsed time: 0.0087 sec.
    - To derive rotation and translation by matching 19 2D model points into the 3D model

126

Viola-Jones Face & Feature Detection + Kanade– Lucas–Tomasi Feature Tracker + Geometric Head Pose Estimation

20.8966 frames/sec.

127

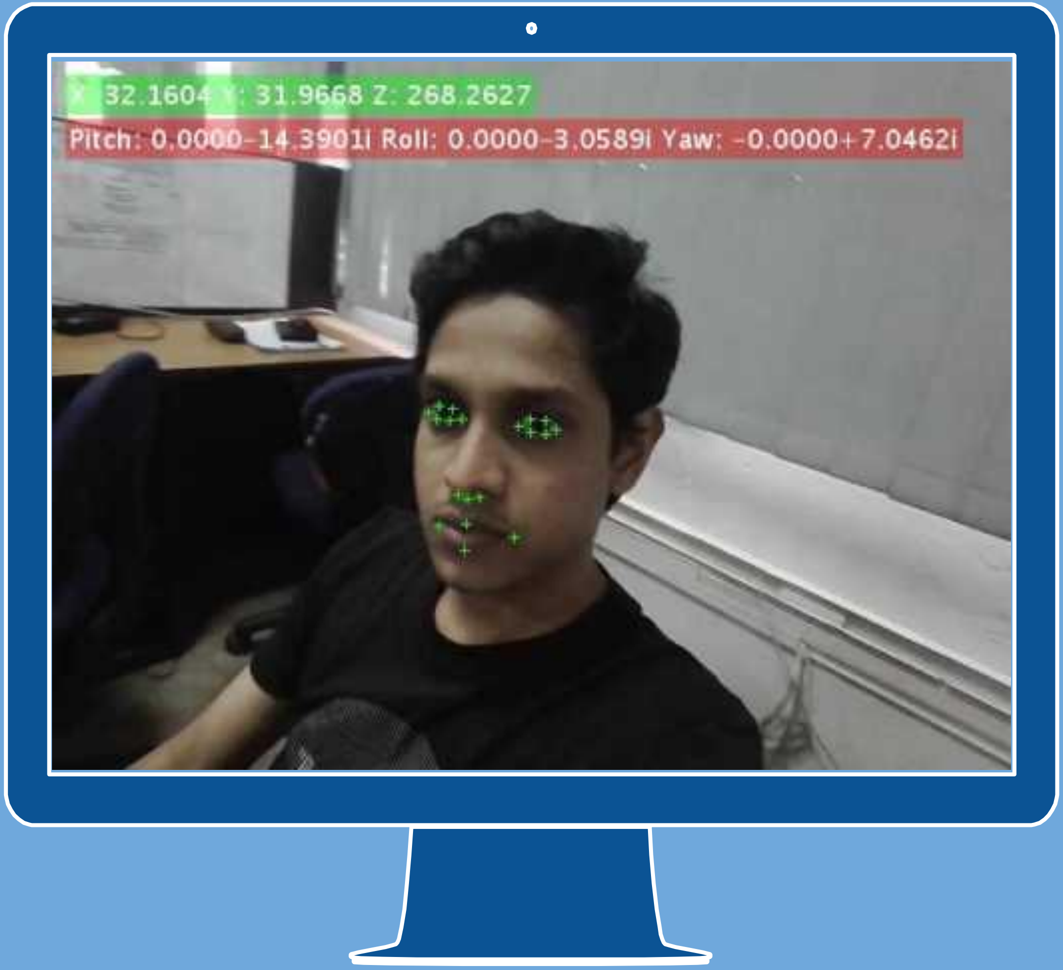
Viola-Jones Face Detection + AAM Facial Fitting + POSIT Head Pose Estimation

7.2326 frames/sec.

128

Viola-Jones Face Detection + Lucas-Kanade Initialization + AAM Facial Fitting + POSIT Head Pose Estimation

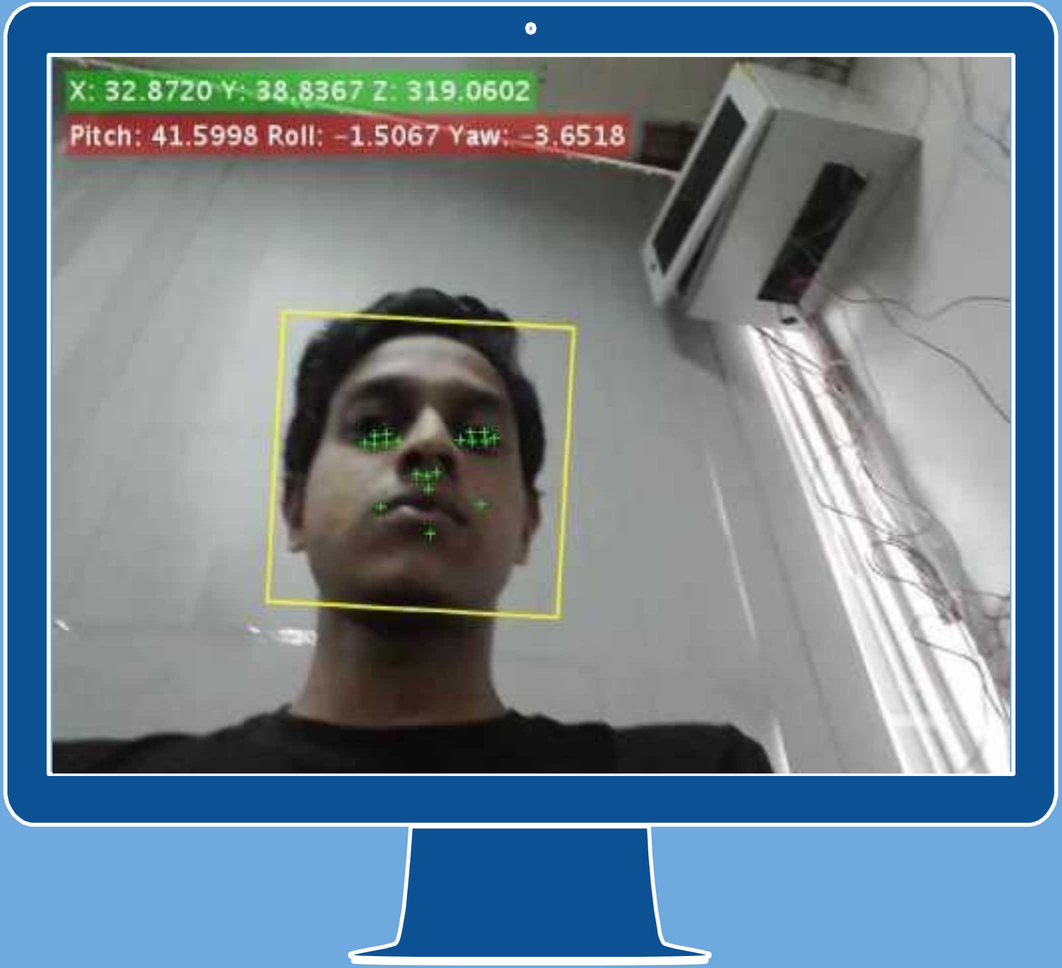
4.5672 frames/sec.



129

Viola-Jones Face Detection + Lucas-Kanade Initialization + AAM Facial Fitting + Divergence Check + POSIT Head Pose Estimation

4.5654 frames/sec.



130

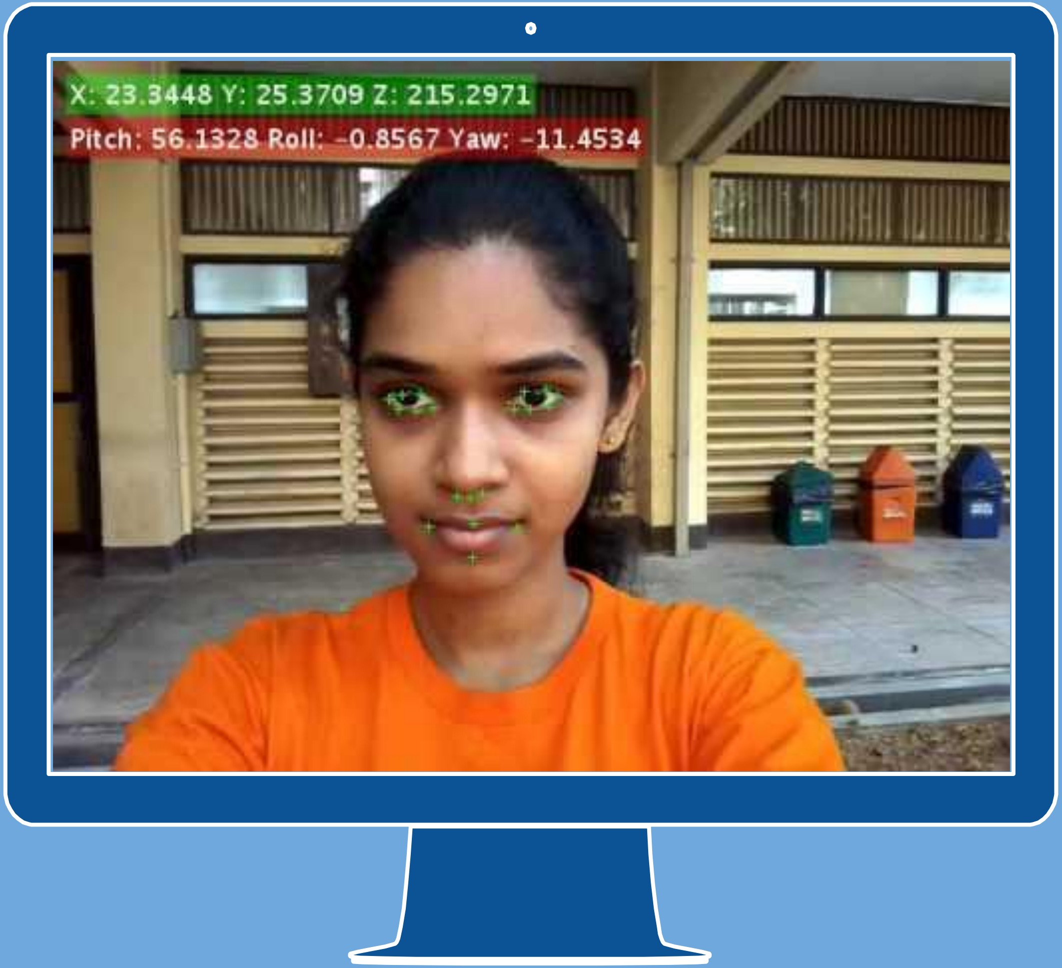
Viola-Jones Face Detection + CLM Facial Fitting + POSIT Head Pose Estimation

0.7158 frames/sec.

131

Viola-Jones Face Detection + Lucas-Kanade Initialization + CLM Facial Fitting + POSIT Head Pose Estimation

0.4563 frames/sec.



132

Viola-Jones Face Detection + ASM Facial Fitting + POSIT Head Pose Estimation

Trained on frames with significant pose variation of test video.

0.2832 frames/sec.

Proposed Solution ctd.

Stages	Percentage Completed (%)
4. Performing a comparison of the performance of each of the approaches based on different criteria like speed, accuracy and precision and further narrowing down the possible set of approaches to proceed with.	60%

134

Test Video Suite

Test Videos under different environmental conditions, devices and camera specifications.

135

Test Video Suite

Environment: Indoor environment

Lighting condition: Diffused light

Device Specifications: Samsung Galaxy Grand Prime

Camera Specifications:

5 MP, f/2.2, 1080p

Environment: Outdoor environment

Lighting condition: Diffused light

Device Specifications:

Samsung Galaxy Core Prime

Camera Specifications:

2 MP

136

Test Video Suite ctd.

Environment: Living room

Lighting condition: Diffused light

Device Specifications: iPhone 6

Camera Specifications: 1.2 MP. f/2.2, 720p video

Environment: Opticians Showroom

Lighting condition: Shop lights

Device Specifications: Samsung Galaxy Grand Prime

Camera Specifications:

5 MP, f/2.2, 1080p

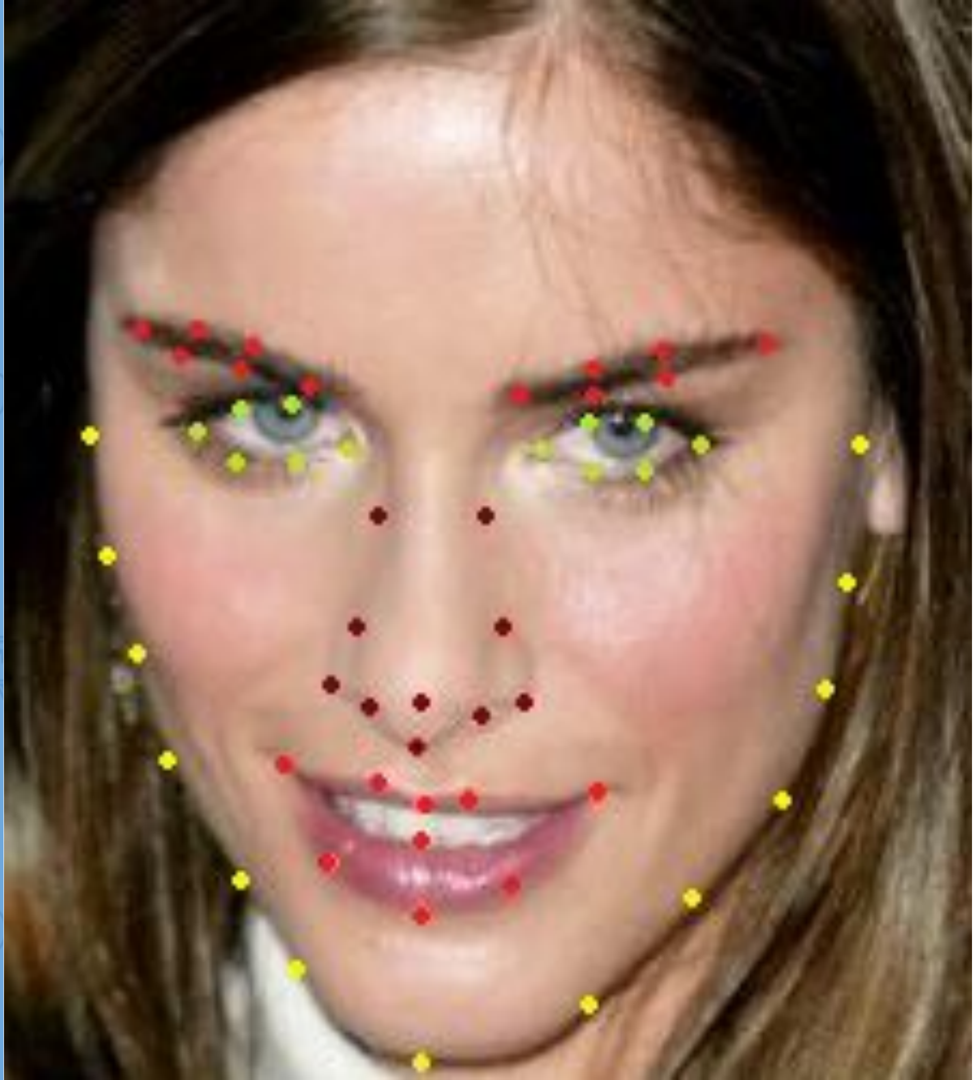
137

Facial Annotations

Annotated facial feature points in each frame of the test videos.

138

Facial Annotations



58 points annotated

Video 1 - 553 frames

Video 2 - 715 frames

Video 3 - 289 frames

Video 4 - 414 frames

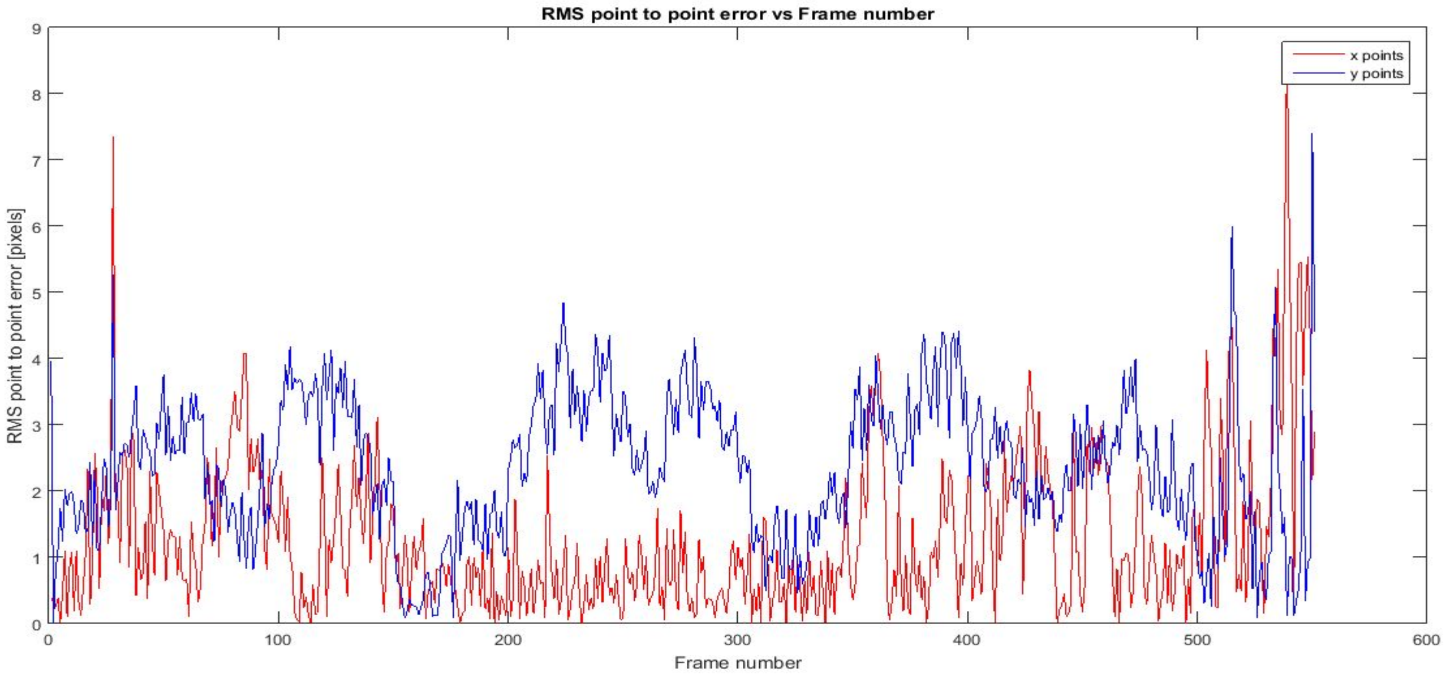
Evaluation Results

RMS point to point errors and FPS.

140

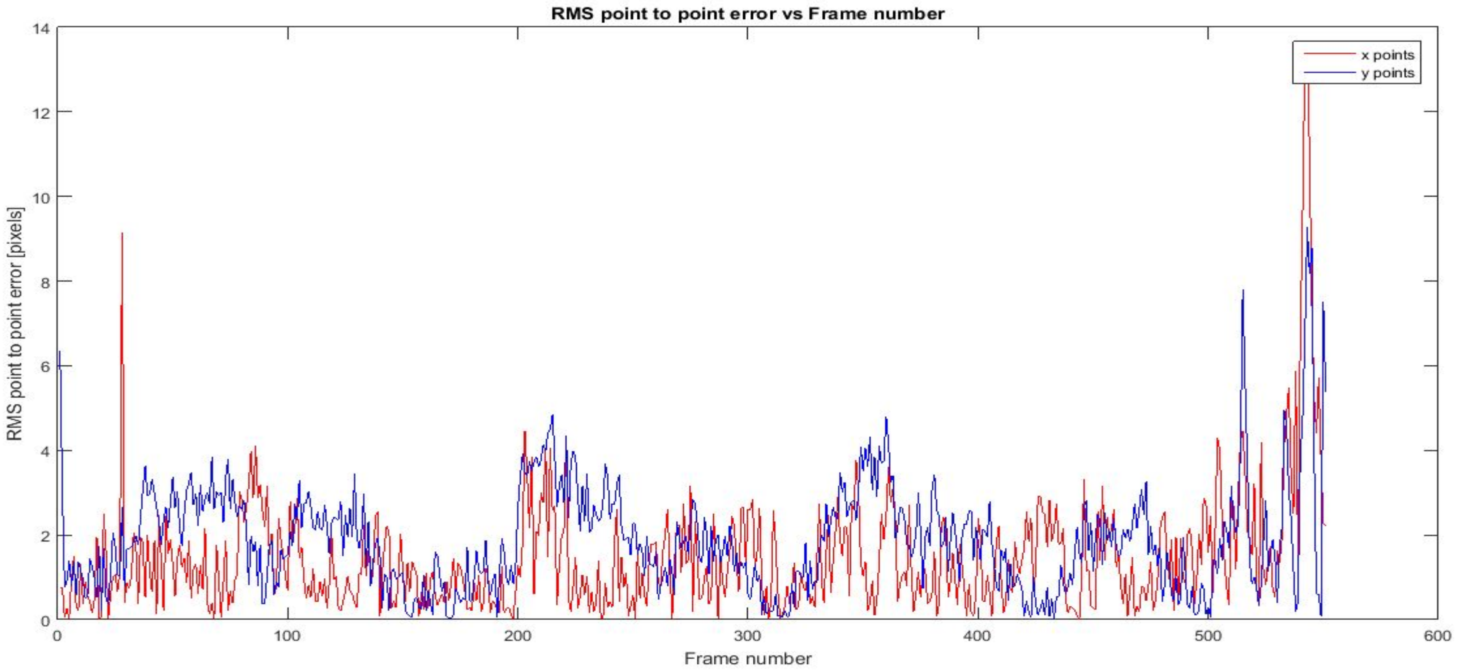
Evaluation Results

- **Viola-Jones Face Detection + AAM Facial Fitting + POSIT Head Pose Estimation**
  - Platform: PC Platform, Points: **19**, Iterations: **10**
  - FPS: **6.0386**



Evaluation Results

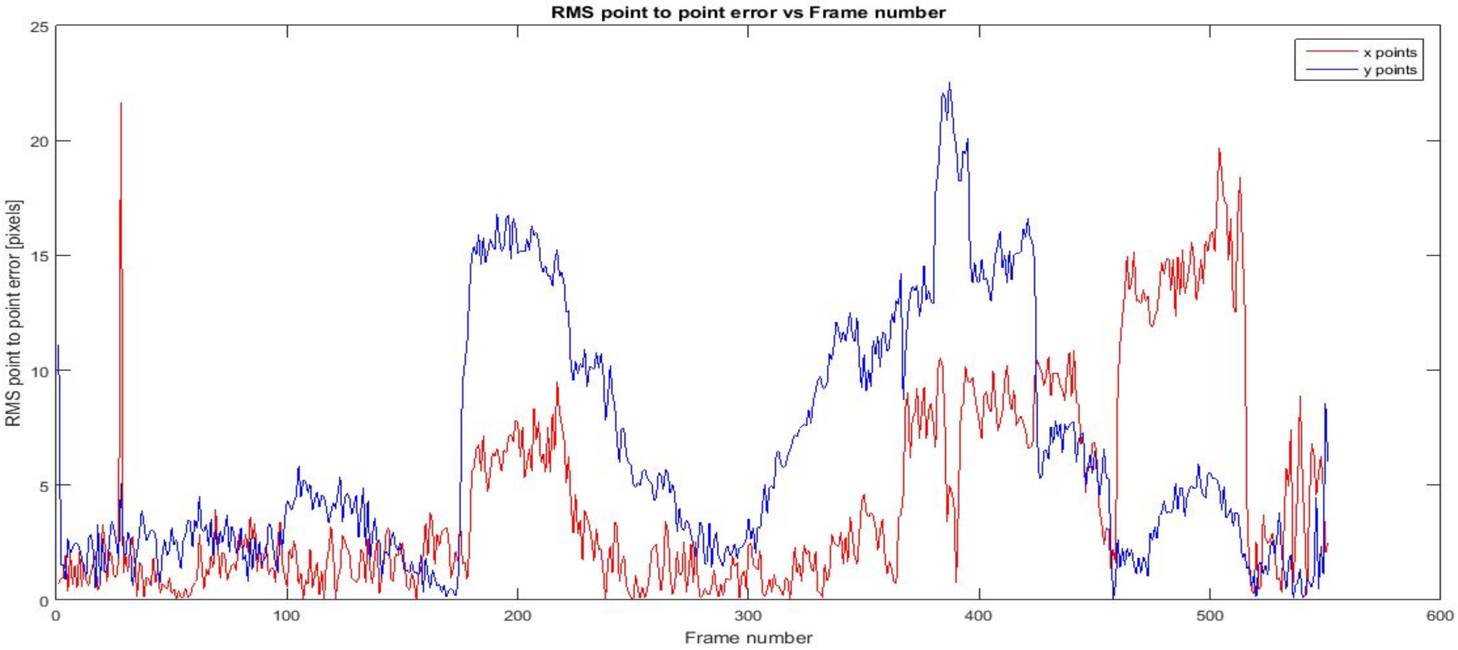
- **Viola-Jones Face Detection + AAM Facial Fitting + POSIT Head Pose Estimation**
  - Platform: PC Platform, Points: **13**, Iterations: **10**
  - FPS: **9.6696**



142

Evaluation Results

- **Viola-Jones Face Detection + AAM Facial Fitting + POSIT Head Pose Estimation**
  - Platform: PC Platform, Points: **5**, Iterations: **10**
  - FPS: **14.1231**



Proposed Solution ctd.

Stages	Percentage Completed (%)
5. Implementing the final set of candidate algorithms on a mobile platform to compare their performance and decide on the best candidate for the purpose.	0%
6. Optimizing the algorithm to achieve better results on a mobile platform.	0%
7. Developing the algorithm as a plugin for the Unity game engine.	0%
8. ***If time permits***, extending the algorithm to take into account the ears as well which would allow to extract more useful information from the video feed like the position of ears.	0%

References

References ctd.

[1] "Augmented reality", Wikipedia, 2016
[2] “LiveRoom”, liveroom.xyz, 2016
[3] "Six degrees of freedom", Wikipedia, 2016
[4] "Total Immersion (augmented reality)", Wikipedia, 2016
[5] Viola and M Jones (2001): Rapid object detection using a boosted cascade of simple features In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001. Proceedings of 2001, vol 1, 1-511. IEEE
[6] M Turk and A Pentland (1991): Eigenfaces for Recognition In: Journal of Cognitive Neuroscience, 1991. Vol. 3, no. 1, 71-86.
[7] J Hannuksela, Diploma thesis, Department of Electrical and Information Engineering, University of Oulu, Finland, 2003.
[8] HA Rowley, S Baluja and T Kanade (1998): Neural network-based face detection In: IEEE Transactions on pattern analysis and machine intelligence, 1998. 20(1), 1-511. IEEE.
[9] E Osuna, R Freund and F Girosit (1997): Training support vector machines: an application to face detection In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1997. Proceedings of 1997, 130-136. IEEE.

References ctd.

[10] T Kawaguchi, M Rizon, Iris detection using intensity and edge information. Pattern Recognition. 36, 549-562 (2003). doi:10.1016/s0031-3203(02)00066-3
[11] M Soriano, B Martinkauppi, S Huovinen, M Laaksonen, Adaptive skin color modeling using the skin locus for selecting training pixels. Pattern Recognition, 36, 681-690 (2003). doi:10.1016/s0031-3203(02)00089-4
[12] RL Hsu, MA Mottaleb, AK Jain (2001) Face detection in color images In: IEEE Transactions on Pattern Analysis and Machine Intelligence, May 2002.
[13] E Hjelmås,B Low, Face Detection: A Survey. Computer Vision And Image Understanding. 83, 236-274 (2001). doi:10.1006/cviu.2001.0921
[14] Marr, David, E Hildreth,Theory of edge detection. Proceedings of the Royal Society of London B: Biological Sciences 207, no. 1167, 187-217 (1980).
[15] K Yow, R Cipolla, Feature-based human face detection. Image And Vision Computing. 15, 713-735 (1997). doi:10.1016/s0262-8856(97)00003-6
[16] K Sobottka, I Pitas, A novel method for automatic face segmentation, facial feature extraction and tracking. Signal Processing: Image Communication. 12, 263-281 (1998). doi:10.1016/s0923-5965(97)00042-8

147

References ctd.

[17] A Sirohey, A Rosenfeld, Eye detection in a face image using linear and nonlinear filters. Pattern Recognition. 34, 1367-1391 (2001).
[18] R Brunelli, T Poggio, Face recognition: Features versus templates. IEEE transactions on pattern analysis and machine intelligence.15, 1042-1052 (1993).
[19] G Chow, X Li, Towards a system for automatic facial feature detection. Pattern Recognition. 26, 1739-1755 (1993). doi:10.1016/0031-3203(93)90173-t
[20] DW Wagener, B Herbst, Face Tracking: An Implementation of the Kanade-Lucas-Tomasi Tracking Algorithm. In: PRASA, South Africa. (2001).
[21] TF Cootes, CJ Taylor, H Cooper and J Graham (1995): Active shape models-their training and application In: Computer vision and image understanding, 61(1), 38-59.
[22] S Milborrow and F Nicolls (2001): Locating facial features with an extended active shape model In: European conference on computer vision, October, 2008. 504-513. Springer Berlin Heidelberg.
[23] TF Cootes, GJ Edwards and CJ Taylor (2001): Active appearance models In: IEEE Transactions on pattern analysis and machine intelligence,23(6), 681-685. IEEE.

References ctd.

[24] J Xiao, T Moriyama, T Kanade, J Cohn (2003): Robust full-motion recovery of head by dynamic templates and re-registration techniques In: International Journal of Imaging Systems and Technology. 13(1), 85-94.

[25] DF DeMenthon and LS Davis (1992): Model-based object pose in 25 lines of code In: European conference on computer vision, 1992. 335-343. Springer Berlin Heidelberg.

[26] Z Gui and C Zhang (2006): 3D head pose estimation using non-rigid structure-from-motion and point correspondence In: TENCON 2006-2006 IEEE Region 10 Conference, 2006. 1-3. IEEE.

[27] TF Cootes, GV Wheeler, KN Walker and CJ Taylor (2002): View-based active appearance models In: Image and vision computing, 2002. 20(9), 657-664.

[28] J Aghajanian and S Prince (2009): Face Pose Estimation in Uncontrolled Environments In: BMVC, 2009. vol 1, No. 2, 3.

[29] P Jiménez, J Nuevo, LM Bergasa and MA Sotelo (2009): Face tracking and pose estimation with automatic three-dimensional model construction In: IET Computer Vision, 2009. 3(2), 93-102.

149

References ctd.

[30] J Ng and S Gong (1999): Multi-view face detection and pose estimation using a composite support vector machine across the view sphere In: Proc. IEEE International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, 1999. IEEE.

[31] J Sherrah, S Gong and EJ Ong (1999): Understanding Pose Discrimination in Similarity Space In: BMVC, September, 1999. 1-10.

[32] Y Li, S Gong and H Liddell (2000): Support vector regression and classification based multi-view face detection and recognition In: Fourth IEEE International Conference on Automatic Face and Gesture Recognition, 2000. Proceedings of 2000, 300-305. IEEE.

[33] E Murphy-Chutorian, A Doshi and MM Trivedi (2007): Head pose estimation for driver assistance systems: A robust algorithm and experimental evaluation In: IEEE Intelligent Transportation Systems Conference, 2007. 709-714. IEEE.

[34] Y Ma, Y Konishi, K Kinoshita, S Lao and M Kawade (2016): Sparse bayesian regression for head pose estimation In: 18th International Conference on Pattern Recognition (ICPR'06), 2006. vol 3, 507-510. IEEE.

[35] A Gee and R Cipolla (1994): Determining the gaze of faces in images In: Image and Vision Computing, 12(10), 639-647.

References ctd.

[36] L Zhao, G Pingali and I Carlborn (2002): Real-time head orientation estimation using neural networks In: International Conference on Image Processing, 2002. Proceedings of 2002, vol 1, 1-297. IEEE.

[37] R Rae and HJ Ritter (1998): Recognition of human head orientation based on artificial neural networks In: IEEE Transactions on neural networks, 1998. 9(2), 257-265.

[38] V Krüger and G Sommer (2002): Gabor wavelet networks for efficient head pose estimation In: Image and vision computing, 2002. 20(9), 665-672.

[39] M Osadchy, Le Cun and ML Miller (2006): Synergistic face detection and pose estimation with energy-based models In: Toward Category-Level Object Recognition, 2006. 196-206. Springer Berlin Heidelberg.

[40] J Wu, M Trivedi (2008) A two-stage head pose estimation framework and evaluation In: Pattern Recognition, vol. 41, no. 3, pp. 1138-1158

THANKS!