US
UNIVERSITY OF SUSSEX
Sussex Research

Learning based automatic face annotation for arbitrary poses and expressions from frontal images only

Akshay Asthana, Roland Goecke, Novi Quadrianto, Tom Gedeon

Publication date
18-08-2009

Licence
This work is made available under the Copyright not evaluated licence and should only be used in accordance with that licence. For more information on the specific terms, consult the repository record for this item.

Document Version
Accepted version

Citation for this work (American Psychological Association 7th edition)
Asthana, A., Goecke, R., Quadrianto, N., & Gedeon, T. (2009). *Learning based automatic face annotation for arbitrary poses and expressions from frontal images only* (Version 1). University of Sussex. https://hdl.handle.net/10779/uos.23401040.v1

Published in
Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition; Florida, USA; 20-25 June 2009

Link to external publisher version
https://doi.org/10.1109/CVPR.2009.5206766

Copyright and reuse:
This work was downloaded from Sussex Research Open (SRO). This document is made available in line with publisher policy and may differ from the published version. Please cite the published version where possible. Copyright and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners unless otherwise stated. For more information on this work, SRO or to report an issue, you can contact the repository administrators at sro@sussex.ac.uk. Discover more of the University's research at https://sussex.figshare.com/

Learning Based Automatic Face Annotation for Arbitrary Poses and Expressions from Frontal Images Only

Akshay Asthana[1] Roland Goecke[1,3] Novi Quadrianto[1,4] Tom Gedeon[2]
[1]RSISE and [2]DCS, Australian National University, Australia
[3]Faculty of Information Sciences and Engineering, University of Canberra, Australia
[4]SML Group, NICTA Canberra Research Laboratory, Canberra, Australia
aasthana@rsise.anu.edu.au , roland.goecke@ieee.org , novi.quad@gmail.com , tom.gedeon@anu.edu.au

Abstract

Statistical approaches for building non-rigid deformable models, such as the Active Appearance Model (AAM), have enjoyed great popularity in recent years, but typically require tedious manual annotation of training images. In this paper, a learning based approach for the automatic annotation of visually deformable objects from a single annotated frontal image is presented and demonstrated on the example of automatically annotating face images that can be used for building AAMs for fitting and tracking. This approach employs the idea of initially learning the correspondences between landmarks in a frontal image and a set of training images with a face in arbitrary poses. Using this learner, virtual images of unseen faces at any arbitrary pose for which the learner was trained can be reconstructed by predicting the new landmark locations and warping the texture from the frontal image. View-based AAMs are then built from the virtual images and used for automatically annotating unseen images, including images of different facial expressions, at any random pose within the maximum range spanned by the virtually reconstructed images. The approach is experimentally validated by automatically annotating face images from three different databases.

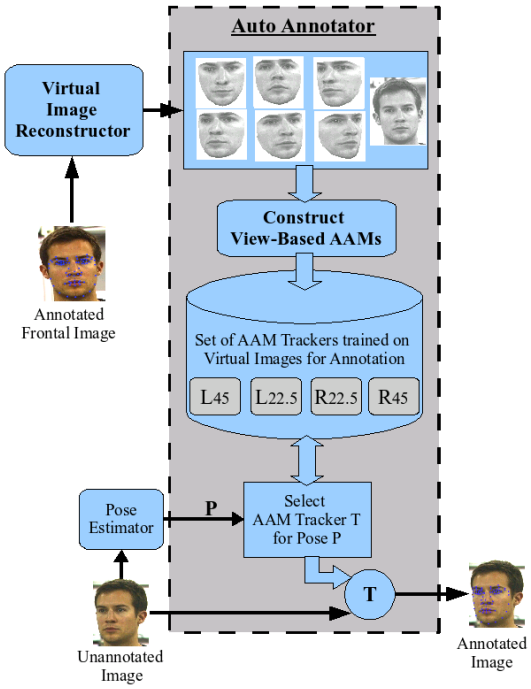


Figure 1: Overall architecture of the *Auto Annotator*

1. Introduction

In recent years, statistical approaches have been widely and successfully used for building non-rigid deformable models. Prominent members of this family of approaches include the Active Shape Model (ASM) [7], Active Appearance Model (AAM) [11] and 3D Morphable Model (3DMM) [4]. Their power lies in the combination of a compact parametric representation and an efficient alignment method. However, one major drawback of these approaches is that they require a labelled dataset of training images, which typically equates to tedious manual annotation.

Taking the AAM as an example, pseudo-dense annotations are required for every training image to build the statistical models of shape and texture. With the object annotation in each training image often requiring the labelling of dozens or hundreds of corresponding points, manually annotating large image databases is tedious and error prone. In addition, objects that only present a small number of distinct landmark points, e.g. a face containing mostly edge structures, make it difficult to consistently annotate the same landmarks (for example, on the chin line).

The quest is thus to develop an automatic annotation method, which in turn would allow for an automatic model

building process. While a number of methods for finding automatic correspondences have been proposed [2, 5, 6, 14, 15, 21, 22, 25], these suffer from various drawbacks such as requiring sufficient salient features in the visual object to build good appearance models, ignoring the global image structure, requiring a large number of parameters for the warping function that results in slow optimisation, or not taking into account the sequential nature of image sequences where available.

In this paper, an approach for automatic annotation of face images is presented that only requires annotated frontal images, thus drastically simplifying the model building process. We firstly propose a data-driven approach to learn the correspondence between manually annotated landmark points of frontal and varying viewpoint images of a face. Secondly, we propose a framework to reconstruct virtual images, of any arbitrary pose and expression, from frontal view images and use these to create *View-Based AAMs* [8] that can then be used to locate the facial features in an image and, hence, annotate it automatically. Finally, we show the utility of our proposed framework by automatically annotating the images from the CMU PIE [23], Face Pointing [18] and FERET [19] databases, and verifying the results by comparing them with a ground truth obtained from manual annotations.

Fig. 1 shows a schematic of the proposed method which only requires a single annotated frontal face image (per expression) as input. The *Virtual Image Reconstructor* (Sec. 4) reconstructs the virtual face images at different poses. Then, the *View-Based AAMs* (Sec. 5) are constructed from these virtual images and stored in the repository for future use. Now, given a new image to be annotated, the system first estimates the pose via a *Pose Estimator* (Sec. 5.2) and selects the appropriate AAM from the repository of View-Based AAMs, which is then used to locate the features in the given image and, hence, annotate it automatically.

The remainder of this paper is structured as follows. Sec. 2 provides a brief overview of related work in automatic annotation and model building. Sec. 3 gives details of regression based learning as further background information. In Sec. 4, the process of reconstructing virtual images from a single annotated frontal image is described. The process of automatically building view-based AAMs is presented in Sec. 5. The approach is experimentally validated in Sec. 6.

2. Related Work

The issue of automatically annotating face images and building AAMs has received considerable attention in the research literature in recent years. More generally, the issue is one of finding pseudo-dense correspondences across images of the same object class. Approaches can be broadly categorised into either image or feature based approaches.

In image based methods, dense image correspondences are found through a nonlinear warping function that minimises some error measure between the pixel intensities. [15] models images as 'bags' of pixel values, enabling the computation of correspondences simultaneously for all images. In [6], a groupwise registration using a set of nonlinear diffeomorphic warps is proposed, providing dense correspondences between all images, avoiding the need for manual annotation of training images. In [2], the problem of automatic annotation and model building is re-posed as an energy-minimising image coding problem. Image based methods have the advantage that the global image structure is taken into account, thus better mimicking the AAM for which the correspondences are used later. The main disadvantage is that the warping function will generally need to be parameterized using a large number of parameters (as a set of landmark points), which results in a very large optimisation problem that is slow to optimise and prone to terminating in local minima.

In feature based methods, correspondences are found between salient image features through examining the local structure of the features. In [14], a sparse polygonal representation of one shape's boundary is matched onto a second shape's boundary via a greedy optimisation of a cost function. In [5], the point matching algorithm uses an iterative joint clustering and matching strategy, which reduces the computational complexity while maintaining accuracy. [21] and [25] make use of images sequences to automatically build models, with only the first frame needing to be annotated manually, thus exploiting the fact that it is easier to track correspondences than finding them between two arbitrary images. In [22], this line of work is extended to take the scene geometry into account through the epipolar constraints in stereo images. The advantage of feature based methods is that the feature comparisons and calculations are comparatively cheap. Their disadvantages are that they require a sufficient number of salient features in the object and that the global image structure is ignored, as the feature comparisons generally only consider local image structure, which can lead to suboptimal fitting results.

3. Regression Based Learning

Given m observed data points $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{m}$, where $y_i \in \mathcal{Y}$ (the set of outputs/targets) and $x_i \in \mathcal{X}$ (the set of inputs), the goal of learning is to infer a function $f : \mathcal{X} \rightarrow \mathcal{Y}$. In a regression setting, $y_i \in \mathbb{R}$. In the following, we briefly review the regression techniques used for the experiments in this paper.

3.1. Support Vector Regression (SVR)

SVR is one of the most ubiquitous regression methods used for single output data. SVR minimises the regularised risk functional, $\left\{ \frac{1}{2} \left\| w \right\|^2 + \frac{C}{m} \sum_{i=1}^{m} \left| y_i - f(x_i) \right|_\epsilon \right\}$, where

C is the (inverse) regularisation parameter and the ϵ-insensitive loss function is defined as

$$|y_i - f(x_i)|_\epsilon := \begin{cases} 0, & \text{if } |y_i - f(x_i)| \leq \epsilon \\ |y_i - f(x_i)| - \epsilon, & \text{otherwise.} \end{cases}$$

The prediction at a new input x^* is given by $f(x^*) = \sum_{i=1}^{m}(a_i - \hat{a}_i)k(x^*, x_i) + b$, where a_i and $\hat{a}_i$ are the Lagrange multipliers of the dual objective formulation of the regularised risk functional, $k(x^*, x_i)$ is the kernel function and b is the bias parameter. By exploiting the *Karush-Kuhn-Tucker (KKT)* conditions, only data points with nonvanishing coefficients ($a_i \neq 0$ or $\hat{a}_i \neq 0$) will affect the prediction and are called support vectors [3].

3.2. Boosted Support Vector Regression

In order to improve the predictive capacity of standard SVR, an Adaboost style algorithm [12] can be used. The main idea for boosting SVR is to iteratively train a sequence of SVR models on the weighted data while increasing the weight on the samples that obtained a large error from the previous SVR model at every iteration. Refer to [10] for a detailed discussion of the boosting procedure used for the experiments presented in this paper.

3.3. Gaussian Process Regression (GPR)

GPR has gained increased popularity in statistical machine learning as it offers a principled nonparametric Bayesian framework for inference, model fitting and model selection [3]. In this framework, we observe a noisy output $y_i = f(x_i) + \epsilon_i$ at input location x_i and the noise term is assumed to be independent and normally distributed, $\epsilon_i \sim \mathcal{N}(0, \sigma_{\text{n}}^2)$. Placing a Gaussian process prior over functions will lead us to a Gaussian predictive distribution

$$y^*|x^*, \mathcal{D} \sim \mathcal{N}(\mu, \sigma^2), \text{ with} \quad (1)$$

$$\mu = k^*[K + \sigma_{\text{n}}^2 I]^{-1} y \quad (2)$$

$$\sigma^2 = k(x^*, x^*) + \sigma_{\text{n}}^2 - k^*[K + \sigma_{\text{n}}^2 I]^{-1} k^{*T} \quad (3)$$

for a noisy query point x^*. In these equations, we have $K \in \mathbb{R}^{m \times m}$, $K_{ij} = k(x_i, x_j)$ and $k^* \in \mathbb{R}^{1 \times m}$, $k_i^* = k(x^*, x_i)$. Here, k denotes a covariance function which encodes our assumptions about the function we wish to learn. In this work, we employ a squared exponential covariance function, $k(x_p, x_q) = \sigma_f^2 \exp(-\frac{1}{2\sigma_l^2} \|x_p - x_q\|_2^2)$ which has the characteristic length scale l and the signal variance σ_f^2 as free parameters (hyperparameters). From this Gaussian predictive distribution, we are interested to make a point prediction y_{guess} by minimising the expected loss or risk as

$$y_{\text{optimal}}|x^* = \underset{y_{\text{guess}}}{\text{argmin}} \int \mathcal{L}(y^*, y_{\text{guess}}) p(y^*|x^*, \mathcal{D}) dy^* \quad (4)$$

where $\mathcal{L}(.,.)$ is the loss function. For squared loss functions (or any other symmetric loss function), the optimal point prediction at query point x^* is

$$y_{\text{optimal}}|x^* = \mathbf{E}_{(y^*) \sim p(y^*|x^*, \mathcal{D})}[y^*] = \mu \quad . \quad (5)$$

4. Virtual Image Reconstructor

In this section, we investigate the process of virtual image reconstruction of a face at any arbitrary pose from a single frontal view annotated image. The basic idea here is to learn the correspondence between the landmark points of the gallery (frontal view images) and the probe images (images at any arbitrary pose) exhibiting arbitrary expressions. Once this learner has been trained, it can be used to predict the spatial arrangement of the landmark points for any other (unseen) face at any arbitrary pose and expression.

4.1. Learning the Reconstruction Parameters

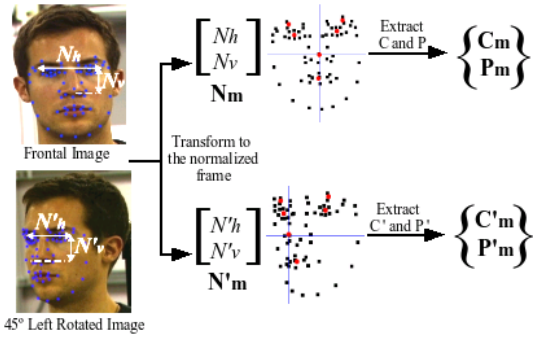


Figure 2: Extracting Normalised, Centroid and Point Vectors

We wish to learn the correspondence between the n landmark points of m gallery and probe images. Given the manual annotations for the gallery and probe images, we extract 3 vectors: *Normalisation Vector*, *Centroid Vector*, and *Point Vector* from each of the gallery and probe images (Fig. 2).

- **Normalisation Vector** ($\mathbf{N}$ from gallery images and $\mathbf{N}'$ from probe images) is a 1D vector containing information about the normalisation distances used to normalise the feature vectors with respect to the varying shape and size of faces of different people in the database. Horizontal normalisation distance (N_h) is the horizontal distance between the eye corners. Vertical normalisation distance (N_v) is the vertical distance between the eye corners and the nose tip (also the reference point in the normalised frame)

$$\mathbf{N} = [N_h; N_v]^T \qquad \mathbf{N}' = [N_h'; N_v']^T \quad . \quad (6)$$

- **Centroid Vector** ($\mathbf{C}$ from gallery images and $\mathbf{C}'$ from probe images) is a 1D vector containing the location of the centroids of six individual facial features (left and right eyebrows, left and right eyes, nose and mouth) in the normalised frame. For this, we create a dictionary of landmark points providing us the information about which of the six facial features each of the n landmark points represents. Hence, if the number of landmark

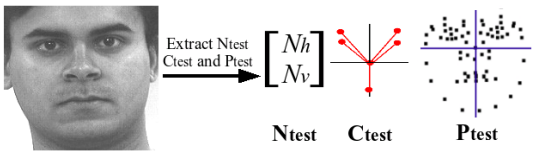


Figure 3: Extracting Normalised, Centroid and Point Vectors from an annotated frontal image

points c represents a facial feature, say the mouth, then the centroid $(\vec{x}, \vec{y})$ of this facial feature is computed as $\vec{x} = \left(\frac{\sum_{i=1}^{c} x_i}{c}\right)$ and $\vec{y} = \left(\frac{\sum_{i=1}^{c} y_i}{c}\right)$, where (x_i, y_i) is the location of each landmark point representing this facial feature in the normalised frame. $\mathbf{C}$ and $\mathbf{C}'$ are then represented as

$$\mathbf{C} = [\vec{x_1}; \vec{y_1}; \ldots; \vec{x_6}; \vec{y_6}]^T \quad \mathbf{C}' = [\vec{x_1'}; \vec{y_1'}; \ldots; \vec{x_6'}; \vec{y_6'}]^T \quad (7)$$

- **Point Vector** ($\mathbf{P}$ from gallery images and $\mathbf{P}'$ from probe images) is a 1D vector containing information about the location of each of the n landmark points in the normalised frame and is represented as

$$\mathbf{P} = [x_1; y_1; \ldots; x_n; y_n]^T \quad \mathbf{P}' = [x_1'; y_1'; \ldots; x_n'; y_n']^T \quad (8)$$

At this stage, we have $m \times 3$ pairs of normalisation, centroid and point vectors, respectively, with each pair representing a gallery image and its corresponding probe image. Next, we construct 3 different training sets

$$\mathcal{T}_{\mathcal{N}} = \left\{(\mathbf{N_i}, \mathbf{N_i'}) \mid i \in (1, 2, \ldots, m)\right\} \quad (9)$$

$$\mathcal{T}_{\mathcal{C}} = \left\{(\mathbf{C_i}, \mathbf{C_i'}) \mid i \in (1, 2, \ldots, m)\right\} \quad (10)$$

$$\mathcal{T}_{\mathcal{P}} = \left\{(\mathbf{P_i}, \mathbf{P_i'}) \mid i \in (1, 2, \ldots, m)\right\} \quad (11)$$

and train a learner via the regression methods described in Sec. 3 to learn 3 different sets of regression models $\mathcal{R}_{\mathcal{N}}$, $\mathcal{R}_{\mathcal{C}}$, $\mathcal{R}_{\mathcal{P}}$ for predicting the normalisation, centroid and point vector respectively, where

$$\mathcal{R}_{\mathcal{N}} = \left\{R_{N_h}, R_{N_v}\right\} \quad (12)$$

$$\mathcal{R}_{\mathcal{C}} = \left\{R_{C_i} \mid i \in (1, 2, \ldots, 12)\right\} \quad (13)$$

$$\mathcal{R}_{\mathcal{P}} = \left\{R_{P_i} \mid i \in (1, 2, \ldots, 2n)\right\} \quad . \quad (14)$$

Here, R represents the regression model learnt over a particular training set.

4.2. Reconstruction of Virtual Images

Given an annotated frontal face image, a virtual image with the same pose as the probe images is reconstructed by predicting the new landmark locations and warping the texture from the frontal image via Piecewise Affine Warping (PAW). PAW, commonly used in AAM methods [11, 17],

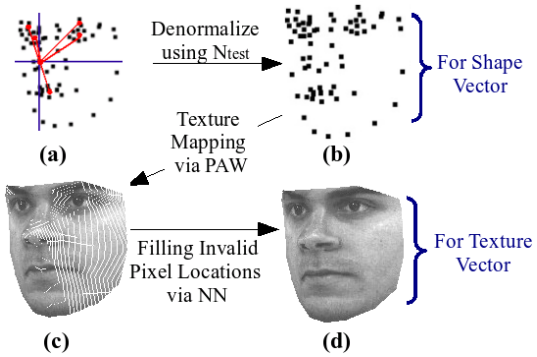


Figure 4: Step by step reconstruction of virtual image

is a spatial transformation function for which the reference frame is divided into a set of non-overlapping regions such that all locations within each region are warped using the same affine transformation. These regions are generally defined by some type of triangulation of a point set, such as the Delaunay triangulation [9], defined in the reference frame. The result is that locations in the reference frame are warped to locations in the destination frame with the same barycentric coordinates, with respect to its encompassing triangle. Once the texture has been warped, all the invalid pixel locations are filled by their nearest-neighbours to complete the virtual image reconstruction procedure. The step by step procedure is given in Algorithm 1.

5. Creating the AAM from Virtual Images

Once the virtual images have been reconstructed, we use them to train AAMs and use an existing AAM fitting method to locate the facial features in the original images and, hence, annotate them automatically. However, it should be noted that the reconstructed virtual images have no information about the background or the area outside the convex hull of the canonical shape. This lack of information has the potential to compound the effect of already existing problems associated with AAMs, such as adaptability to a changing background, problems with poor initialisation and ill-defined borders.

To deal with the problem of adaptability to the changing background, we use the Simultaneous Inverse Compositional method (SIC) [1], a generative fitting method, where the update model is generated directly from background free components (i.e. the mean appearance and their modes of variation) and has no specialisation to any particular background. However, poor initialisation is again an issue as far as generative fitting methods (SIC in this case) are concerned. When initialisation is far from the optimum, with a large proportion of the image under the current warp estimate consisting of background pixels, these approaches are prone to terminating in local minima. Also, if the border

Algorithm 1 Reconstruction of Virtual Images

Require: Annotated Frontal Face Images Img.

1: Extract $\mathbf{N}_{test}$, $\mathbf{C}_{test}$, $\mathbf{P}_{test}$ as shown in Figure 3.
2: Use $\mathcal{R}_{\mathcal{N}}$ (Eqn.12), $\mathcal{R}_{\mathcal{C}}$ (Eqn.13), $\mathcal{R}_{\mathcal{P}}$ (Eqn.14) to predict $\mathbf{N'}_{test}$, $\mathbf{C'}_{test}$ and $\mathbf{P'}_{test}$.
3: $\mathbf{P'}_{test}$ contains the new location of each of the n landmark points w.r.t. the origin in the normalised frame. Arrange the landmark points in the normalised frame accordingly.
4: $\mathbf{C'}_{test}$ contains the new location of six centroids, representing six individual facial features (Eqn.7), w.r.t. the origin in the normalised frame. Arrange these centroids in the normalised frame accordingly. Red dots, joined by red lines (Fig. 4(a)), represent the location of the new centroids in the normalised frame.
5: Transform the group of landmark points representing each of the six individual facial features (arranged in Step 3) to the new centroid locations drawn in Step 4. Black dots in Fig. 4(a) represent the location of the new landmark points in the normalised frame.
6: Reconstructed landmark points in the normalised frame are de-normalised (Fig. 4(b)) using the $\mathbf{N'}_{test}$.
For $i = 1$ to n
$x_i^{real} = x_i^{norm}.(N'_h)_{test}$
$y_i^{real} = y_i^{norm}.(N'_v)_{test}$
End For
7: Warp the texture from Img to the new locations (Fig. 4(b)) using PAW. Fig. 4(c) shows the texture warping result.
8: Fill the invalid pixel locations (shown white in Fig. 4(c)) with their nearest neighbours. Fig. 4(d) shows the reconstructed virtual face image.

of an object is ill-defined, i.e. there is little difference in the texture on either side of the border, for example, the outer boundary of the face especially when it is rotated sideways, AAMs tend to overfit these boundaries and, therefore, result in an overall poor fitting [24]. Hence, the lack of information about the area outside the convex hull of the canonical shape and the background in the virtual images makes the model vulnerable and can result in inaccurate automatic annotations of original images.

5.1. View-Based Active Appearance Model

To deal with the problem of *'the lack of information'*, a view-based AAM building approach is adopted. The motivation behind this *View-Based AAM* [8] approach is to reduce the search space for a single AAM by dividing it among several AAMs. This helps reducing the amount of shape and texture variation to be handled by a single AAM, thus, reducing the probability of the fitting procedure to converge to a local minimum in case of poor initialisation and also reducing the overfitting problem of ill-defined borders. This view-based approach for training and fitting is given in Algorithm 2.

Algorithm 2 Training and Fitting via View-Based AAMs

For Training -

Require: Annotated face images with pose from L° Left to R° Right.

1: Divide the pose range into l equal intervals.
2: Prepare l training sets, one for each interval of pose range.
3: Train l AAMs and save them in the repository of AAMs for future use.

For Fitting -

Input: Face image to be annotated.

1: Estimate the pose p (Sec. 5.2)
2: From the repository of l AAMs, choose the AAM that contains pose p in its range of pose variation.
3: Use this AAM to fit the input image and obtain the automatic annotation.

For the experiments presented in this paper, pose varied from 45° Left to 45° Right and was divided into 4 intervals of 0° to 22.5° Left, 22.5° to 45° Left, 0° to 22.5° Right and 22.5° to 45° Right for training the view-based AAMs.

5.2. SIFT Descriptor Based Pose-Estimator

Recently, an image retrieval based approach has been proposed for real-time 3D pose estimation, showing robustness to extreme out-of-plane rotation, background variation and facial expressions [13]. Since our aim here is to *estimate* the interval in which the pose p lies (0° to 22.5° Left, 22.5° Left to 45° Left, 0° to 22.5° Right or 22.5° Right to 45° Right in this paper) this approach suits the purpose well. As proposed in [13], we prepare a database of registered face images of different people with different poses (0°, 22.5° Left, 45° Left, 22.5° Right and 45° Right in this case) from the CMU PIE database [23]. We use SIFT descriptors [16] for matching images. The input image, for which we are trying to compute the pose, is used as a query image and our database of registered face images is searched for the most similar image. We use the *pose accumulation simple voting scheme* [13] to compute the best matching score and, hence, the closest pose.

6. Experiments

We first conducted experiments on the CMU PIE database [23]. Overall, 427 images across 31 persons (26 males and 5 females), with a face cropped area of approximately 140×150 pixels, were manually annotated with 69 landmarks each. There were 62 images from each of the

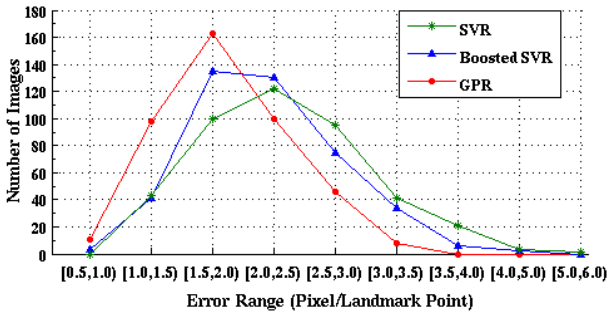


Figure 5: Annotation error distribution for CMU PIE Db

Cameras C27 (Frontal Pose), C07 (30° Down[1]), C09 (30° Up), C29 (22.5° Left), C11 (45° Left), C37 (45° Right) and 117 images from Camera C05 (22.5° Right) exhibiting various facial expressions. Six different learners were trained (Sec. 4.1), one each for predicting the new landmark locations for poses 22.5° Left, 45° Left, 22.5° Right, 45° Right, 30° Up and 30° Down. Given annotated frontal images of an unseen person with arbitrary facial expressions, virtual images were reconstructed (Sec. 4.2) and 4 View-Based AAMs (Sec. 5) were created ($L22.5$, $L45$, $R22.5$ and $R45$). These View-Based AAMs were then used to automatically annotate unseen images of the person, having any random facial expression and random pose within the maximum range spanned by the virtually reconstructed images ($\pm 45°$ Left-Right and $\pm 30°$ Up-Down, in this case).

A leave-one-out cross-validation scheme was adopted throughout the experiments, i.e. data from 1 person was used for validation and data from the remaining 30 persons for training. The learners were trained using 3 different regression techniques (Sec. 3) and the complete set of 427 images from the CMU PIE Db were annotated automatically using the proposed framework. In order to evaluate the performance of the proposed framework, the pixel error per landmark point was computed between the manual and automatic annotations for every image. It should be noted here that consistently manually annotating the outer boundary of the face is highly error prone due to the lack of distinct features. Therefore, we computed the pixel error per landmark point by excluding the 13 landmark points that represent the outer boundary for each face. Fig. 8 shows this error for the entire dataset obtained from each of the regression techniques.

From these results, it is clear that boosting improves the predictive capacity of the standard SVR to an extent, however, GPR outperforms both SVR and Boosted SVR convincingly. SVR and Boosted SVR have one hyperparameter, C, that needs to be tuned during the learning phase.

[1]Represents the pose captured by Camera C09 and C07 in CMU PIE Database. Since the exact poses have not been provided, we assume them to be approximately 30° Up and Down respectively, throughout this paper.

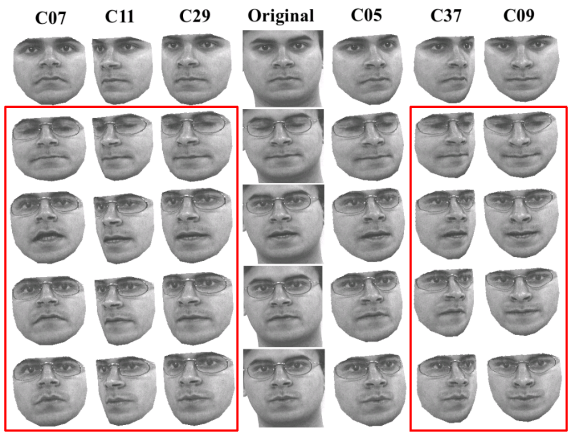


Figure 6: Virtual image reconstruction results for one subject from the CMU PIE database using GPR. Horizontal: different poses. Vertical: different facial expressions.

Denoting the set cardinality of possible hyperparameter values as $|C|$, this means that the parameter search space is of size $|C|^{|\mathcal{R}_\mathcal{N}|+|\mathcal{R}_\mathcal{C}|+|\mathcal{R}_\mathcal{P}|}$. As this is impractical (if not impossible), we restrict ourselves to a search space of $|C|^3$ by assuming all the learners within $\mathcal{R}_\mathcal{N}$, $\mathcal{R}_\mathcal{C}$, and $\mathcal{R}_\mathcal{P}$ share common hyperparameters for SVR and Boosted SVR. GPR offers a principled way of model fitting by maximising the log marginal likelihood, $\log p(y|x)$,

$$-\frac{1}{2}y^T(K+\sigma_n^2 I)^{-1}y - \frac{1}{2}\log\left|K+\sigma_n^2 I\right| - \frac{n}{2}\log 2\pi \quad (15)$$

with respect to the hyperparameters, $\Theta = \{l, \sigma_f^2, \sigma_n^2\}$. This model fitting procedure which respects Occam's razor principle (choosing the simplest model which best explains the observed data) allows us to have different hyperparameters within $\mathcal{R}_\mathcal{N}$, $\mathcal{R}_\mathcal{C}$, and $\mathcal{R}_\mathcal{P}$. Therefore, GPR is more suited for this learning task than SVR and Boosted SVR. Fig. 5 clearly shows the domination of GPR over SVR and Boosted SVR in the form of an annotation error distribution obtained from the CMU PIE database. Assuming the face cropped area to be 16×18 cm in the real world, we computed the real world errors[2] (mm per landmark point) for SVR, Boosted SVR and GPR to be 2.84 ± 0.85 mm, 2.66 ± 0.74mm and 2.26 ± 0.65mm, respectively.

Fig. 6 shows the reconstructed virtual images for a sample subject obtained by GPR. It should be noted here that for the CMU PIE database, the subjects were asked to provide a neutral expression, to smile, to blink and to talk [23]. For talking, a 2s video containing 60 frames was recorded. However, this video is only provided for cameras C27, C22 (3/4 profile camera - *not used for experiments in this paper*) and C05. Thus, the learners for cameras C07, C09, C29, C11 and C37 in our work were trained only on images

[2]http://users.rsise.anu.edu.au/~aasthana/CVPR09/CVPR09Supp.pdf

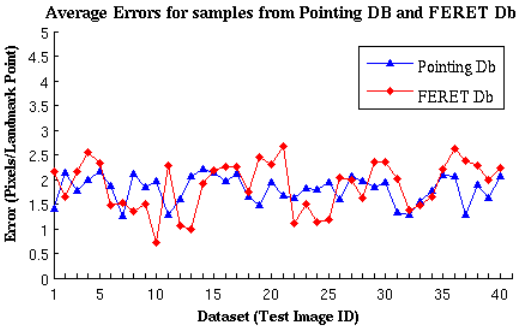


Figure 7: Average error for samples from Pointing Db and FERET Db using GPR

exhibiting neutral and smiling expressions, as they lacked the training data for the other expressions and arbitrary lip movements. In Fig. 6, the virtual images, with arbitrary lip movement, reconstructed from these learners have been marked by red boxes to highlight this fact.

The fitting procedure via view-based AAM trained on virtual images failed to converge for only a single test image (C29-Test Image 54) in the database (out of 427 test images) due to irregularities in the texture of this image caused by excessive reflection on the glasses of the subject. However, the fitting procedure converged accurately for 13 other images of the same speaker in the database that had lesser texture irregularities.

To evaluate the generalisability of the framework, we tested the learners, trained by GPR entirely on the images from the CMU PIE database, for automatically annotating images from the Face Pointing [18] and FERET databases [19]. To this end, we manually annotated 40 random images across 10 speakers from the Face Pointing database and 40 random images across 13 speakers from the FERET database with the pose varying from 45° Left to 45° Right and 30° Up to 30° Down, including different facial expressions. All these images were automatically annotated using the proposed framework and the pixel error per landmark point was computed between the manual and automatic annotations for every image. Fig. 7 shows this error for the image sets from the Face Pointing and FERET databases. Overall, our proposed framework was able to accurately annotate images from CMU PIE Db, Face Pointing Db and FERET Db with a similar average pixel error of $\approx$2mm.

7. Conclusions and Future Work

A regression based learning based approach for the automatic annotation of face images for any arbitrary pose and expression from annotated frontal images only has been presented, which dramatically simplifies the AAM building process. The framework exhibits excellent generalisability, as shown by accurately annotating images from the CMU PIE, Face Pointing and FERET databases[3]. The experiments showed that Gaussian Process Regression gave the best results and, hence, is better suited for this learning task. In future, we plan to extend the approach presented here to also provide a solution for automatically annotating the frontal images, thus making the entire process completely automatic. We also intend to use the reconstructed virtual images directly for face recognition and, hence, to provide a plausible solution to the problem of face recognition from a single image per person.

8. Acknowledgements

The authors would like to thank Jason Saragih for the use of the DeMoLib software [20]. The work presented in this paper was in part supported by the ARC grant TS0669874.

References

[1] S. Baker, R. Gross, and I. Matthews. Lucas-Kanade 20 years on: A unifying framework: Part 3. Technical report, RI, Carnegie Mellon University, USA, 2003.

[2] S. Baker, I. Matthews, and J. Schneider. Automatic Construction of Active Appearance Models as an Image Coding Problem. *IEEE PAMI*, 26(10):1380–1384, 2004.

[3] C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[4] V. Blanz and T. Vetter. Face Recognition Based on Fitting a 3D Morphable Model. *IEEE PAMI*, 25:1063–1074, 2003.

[5] H. Chui, L. Win, R. Schultz, J. Duncan, and A. Rangarajan. A Unified Non-rigid Feature Registration Method for Brain Mapping. *Medical Image Analysis*, 7(2):113–130, 2003.

[6] T. Cootes, S. Marsland, C. Twining, K. Smith, and C. Taylor. Groupwise Diffeomorphic Non-rigid Registration for Automatic Model Building. In *Proc. ECCV 2004*, volume 3024 of *LNCS*, pages 316–327, 2004.

[7] T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active shape models - their training and applications. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.

[8] T. Cootes, K. Walker, and C.J.Taylor. View-Based Active Appearance Models. In *Proc. IEEE FG'00*, pages 227–232, 2000.

[9] M. de Berg, M. van Kreveld, M. Overmars, and O. Schwarzkopf. Computational Geometry (2nd Edition). In *Springer-Verlag*, 2000.

[10] H. Drucker. Improving Regressors using Boosting Techniques. In *Proc. ICML'97*, pages 107–115, 1997.

[11] G. Edwards, C. Taylor, and T. Cootes. Interpreting Face Images Using Active Appearance Models. In *Proc. IEEE FG'98*, pages 300–305, 1998.

[12] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *Journal of Computer and System Sciences, 55(1)*, pages 119–139, 1997.

[3] http://users.rsise.anu.edu.au/~aasthana/CVPR09/CVPR09DivX.avi

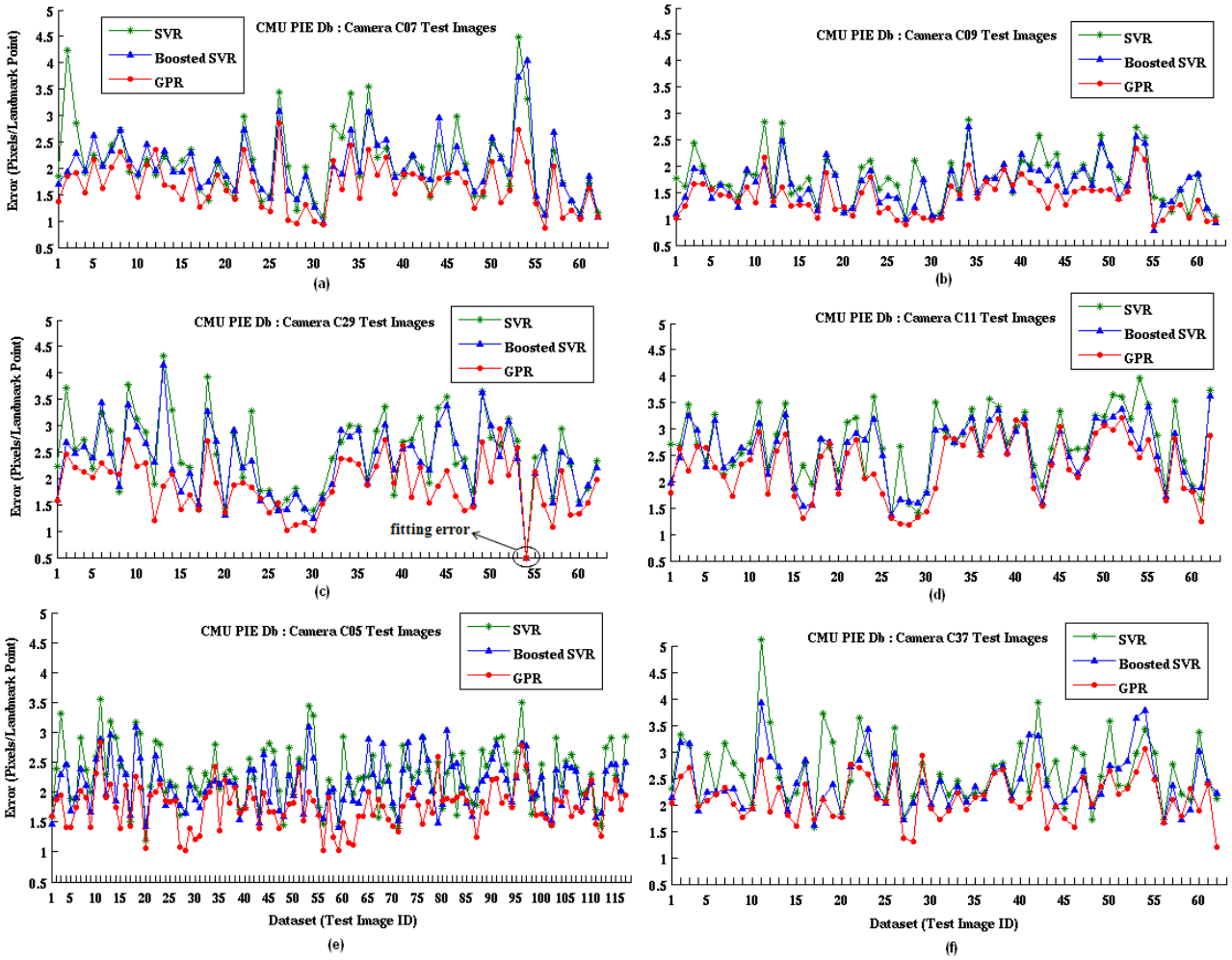


Figure 8: Automatic annotation results for CMU PIE Db

[13] N. Grujic, S. Ilic, V. Lepetit, and P. Fua. 3D Facial Pose Estimation by Image Retrieval. In *Proc. IEEE FG'08*, 2008.

[14] A. Hill and C. Taylor. A Method of Non-rigid Correspondence for Automatic Landmark Identification. In *Proc. BMVC 1996*, volume 2, pages 323–332, 1996.

[15] T. Jebara. Images as Bags of Pixels. In *Proc. IEEE ICCV 2003*, volume 1, pages 265–272, 2003.

[16] D. Lowe. Object recognition from local scale-invariant features. In *Proc. IEEE ICCV*, pages 1150–1157, 1999.

[17] I. Matthews and S. Baker. Active Appearance Models Revisited. *IJCV*, 60(2):135–164, 2004.

[18] J. L. C. N. Gourier, D. Hall. Estimating Face Orientation from Robust Detection of Salient Facial Features. In *Proc. of Pointing 2004, ICPR, Int. Workshop on Visual Observation of Deictic Gestures, Cambridge, UK*, 2004.

[19] P. Phillips, H. Moon, S. Rizvi, and P. Rauss. The FERET Evaluation Methodology for Face Recognition Algorithms. In *IEEE PAMI*, pages 1090–1104, 2000.

[20] J. Saragih. *The Generative Learning and Discriminative Fitting of Linear Deformable Models*. PhD thesis, Australian National University, Canberra, Australia, Oct. 2008.

[21] J. Saragih and R. Goecke. Learning Active Appearance Models from Image Sequences. In *Proc. VisHCI 2006*, volume 56 of *CRPIT*, pages 51–60, 2006.

[22] J. Saragih and R. Goecke. Monocular and Stereo Methods for AAM Learning from Video. In *Proc. IEEE CVPR*, 2007. DOI: 10.1109/ICCV.2007.4409106.

[23] T. Sim, S. Baker, and M. Bsat. The CMU Pose, Illumination, and Expression Database. In *IEEE PAMI*, pages 1615–1618, 2003.

[24] M. Stegmann. *Active Appearance Models: Theory, Extensions & Cases*. PhD thesis, Technical University of Denmark, DTU, 2000.

[25] K. Walker, T. Cootes, and C. Taylor. Automatically Building Appearance Models from Image Sequences Using Salient Features. *Image and Vision Computing*, 20(5–6):435–440, Apr. 2002.