Face Alignment with Part-Based Modeling

Vahid Kazemi
vahidk@nada.kth.se
Josephine Sullivan
sullivan@nada.kth.se

CVAP
KTH Institute of Technology
Stockholm, Sweden

Abstract

We propose a new method for face alignment with part-based modeling. This method is competitive in terms of precision with existing methods such as Active Appearance Models, but is more robust and has a superior generalization ability due to its part-based nature. A variation of the Histogram of Oriented Gradients descriptor is used to model the appearance of each part and the shape information is represented with a set of landmark points around the major facial features. Multiple linear regression models are learnt to estimate the position of the landmarks from the appearance of each part. We verify our algorithm with a set of experiments on human faces and these show the competitive performance of our method compared to existing methods.

1 Introduction

This paper presents a new method for accurate face alignment with part-based modeling. Although we focus on the class of human faces, this method could potentially be applied to any type of deformable object. We aim to learn a regression function mapping a feature representation of the appearance of the face to its shape represented by a set of connected landmarks forming contours around the major facial features. Ideally, we want to use linear regression functions to describe this mapping as they need less training data, have a lower chance of over-fitting, and are fast to compute. However, the relation between the global appearance of an object and its shape is highly nonlinear. Previously, piece-wise linear regression has been applied [12, 15] to deal with this non-linearity. Instead, in this work, the strategy is to learn regression functions for individual parts, see figure 1. The parts are chosen to ensure the linearity of the regression mapping. The main advantage of this approach is that it requires less training data, although it necessitates good part-detection.

Part-based methods, mostly used in recognition tasks, train different classifiers for each part of the model. Assuming each part is a rigid structure, deformation is limited to the relative transformation of each part. The optimal solution for such a multi-part based matching problem can be found efficiently by simplifying the dependency of parts to form a tree structure as is presented by Felzenszwalb *et al.* [8]. This method cannot be directly applied to solve dense matching problems because it requires creating individual classifiers to detect each landmark which is not only computationally expensive but impractical, since most of the landmarks do not have a distinctive local appearance.

A large group of methods have been developed which rely on the global appearance of a deformable object to tackle the alignment problem, these include Active Shape Mod-

BMVC 2011 http://dx.doi.org/10.5244/C.25.27

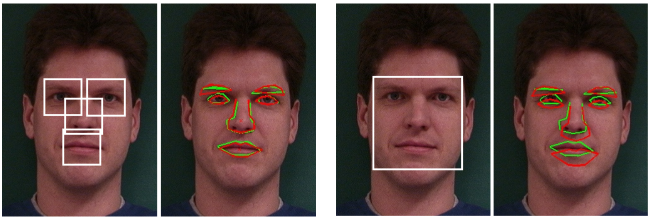

(a) with parts (b) without parts

Figure 1: This figure shows the benefit of using parts in the performance of a regression function to accurately predict the location of landmark points. In both cases a linear regression model is learnt to map the appearance descriptors inside the patches to the location of the landmarks associated with the patch. Green lines represent the ground truth shape and the red lines are the prediction of the regression function. As can be seen greater accuracy is achieved when (a) using a separate regression function for each localized part as opposed to (b) one regression function from the global face patch.

els(ASM) [3], and Active Appearance Model(AAM) [4]. Some attempts have been made to improve the robustness, and accuracy of these methods [1, 5, 9, 10, 13], but the main problem which remains unsolved in these methods is that they need a good initial estimate and are not able to adapt the model to fit a subject when the initial error is high. As an example in face tracking applications, AAM usually fails to converge when there is a sudden large deformation or motion of the subject.

One main advantage of our method compared to the global methods just described is that the regression functions directly estimate the landmark positions by-passing the need to perform iterative non-linear optimization on a complicated cost function. As a result, our method can be used on image sequences with fast motion, and it does not need any initialization. Furthermore the part-based nature of our method, and also the use of HOG descriptors [6] (in contrast to intensity vectors as in AAM), enhance the robustness and generalization ability of our algorithm. On the minus side the performance of our method is highly dependent on a proper part detection and this requires strong and distinctive features in the appearance of the object. In the case of a human face which is the focus of our work, a sensible selection of parts is a division into the nose, mouth and the eyes. For the more general case we would need an automatic part selection method which is a an interesting and challenging topic investigated for future work.

2 Landmark localization via regression

In this paper the localization of landmark points on the face is viewed as a regression problem. Assume we have N training image patches of the same size and that the appearance of each patch is described by a feature vector $\mathbf{f}_i \in \mathbb{R}^K$. Each patch contains L landmark points whose coordinates are defined relative to an origin at the centre of the patch and are then stacked into the vector $\mathbf{X}_i$. This training data is used to learn a regression function

$$q : \mathbb{R}^K \longrightarrow \mathbb{R}^{2L} \quad \text{s.t.} \quad q(\mathbf{f}) = \mathbf{X} \tag{1}$$

mapping a feature vector $\mathbf{f}$ to the coordinates of the landmark points $\mathbf{X}$. There are many options to approximate $q(\cdot)$ such as linear regression, nearest neighbour regression and relevance vector machines. We focus on modelling q as a linear function

$$q(\mathbf{f}) = W\mathbf{f} + \mathbf{w} \tag{2}$$

where $W \in \mathbb{R}^{2L \times K}$ and $\mathbf{w} \in \mathbb{R}^{2L}$. The question then remains which approach should be used to estimate $(W, \mathbf{w})$. The performance of several standard methods - ordinary least squares regression, ridge regression, principal component regression - are investigated and the results are reported upon and compared to the baseline of a nearest neighbour regression function in the experimental section. However, to summarize ridge regression, which is fast in both training and testing phases, was found to perform well. Ridge regression minimizes a least squares loss function combined with a regularization term:

$$W_{\text{ridge}}, \mathbf{w}_{\text{ridge}} = \arg\min_{W,\mathbf{w}} \left(\sum_{i=1}^{N} \|\mathbf{X}_i - W\mathbf{f}_i - \mathbf{w}\|^2 + \lambda(\text{trace}(WW^t) + \mathbf{w}^t\mathbf{w}) \right) \tag{3}$$

where λ is the non-negative regularization factor and defines the complexity of our model.

The approximated regression functions will be applied to arbitrary image patches $I_{\mathbf{b}}$. Such patches are defined by a bounding box $\mathbf{b} = (\mathbf{x}, sw, sh)$, where $\mathbf{x}$ is the centre of the patch, w and h are the width and height of the training patches and s is the ratio between the width of the test patch and the training patches. We assume test patches have the same aspect ratio as the training data. Then the coordinates of the landmark points, estimated by our learnt regression functions, have to be rescaled and translated:

$$g(I_{\mathbf{b}}) = s\,q(\mathbf{f}_{I_{\mathbf{b}}}) + \mathbf{x} \tag{4}$$

where $\mathbf{f}_{I_{\mathbf{b}}}$ is the feature description of the image patch $I_{\mathbf{b}}$.

2.1 Feature description of an image patch

In this subsection we introduce the feature descriptor $\mathbf{f}_{I_{\mathbf{b}}}$ which is used to describe the appearance of an image patch. We use a variant of the PHOG [2] descriptor. Figure 2 shows a schematic representation of this descriptor. At the first level a histogram of gradient orientations of the whole patch is computed. While at the second level the patch is divided into 8 sub-regions. Each of these regions can once again be recursively divided into 8 more sub-regions until the required level of detail is obtained. However, our experiments show that it is more effective to just recursively subdivide the square sub-regions at each level of the pyramid into 8 more sub-regions. At the end, all the histograms are concatenated to form the final descriptor. This descriptor allows us to capture the appearance of an image patch at different scales (same as PHOG) as well as the joint appearance of adjacent regions both horizontally and vertically. In this way we can better represent shape information while maintaining a degree of spatial invariance.

3 Part based regression

As stated we want to use linear regression to model the mapping from the appearance of a face to its shape. However, the relationship between the global appearance of the face and the position of all its landmark points is non-linear. We need, therefore, to model simpler relationships which can be better approximated with a linear function. Our strategy is to

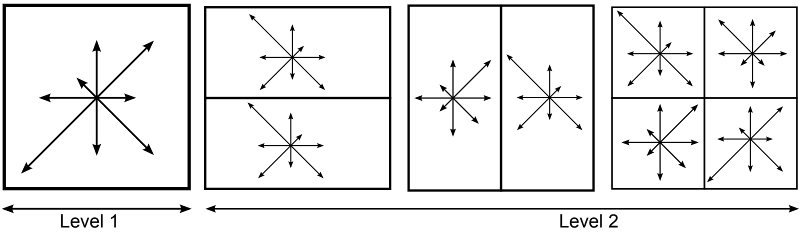


Figure 2: This figure demonstrates how an appearance descriptor is calculated from an image patch. The main patch is divided into 8 sub-regions. For each sub-region the histogram of gradient orientations is calculated. These histograms are then concatenated to form the final descriptor.

split the face into P parts, see figure 4, and learn separate regression functions, mapping each part's appearance to its associated landmark positions. This approach, of course, introduces a new set of problems. The first of these is - *How do we define and set the number of parts?*

Solving this problem automatically is non-trivial and is not tackled in this paper. The parts and their number are set by hand, see figure 4. The parts are defined as a partition of the landmark points and the choice of this partition was guided by an effort to ensure

- the part landmark points can be well aligned across the training images,
- the part can be detected and accurately located in novel images and
- linear regression accurately models the mapping from each part's image feature space to the coordinates of its landmark points.

More formally we partition the set of L landmark points into P subsets with $L_1, \ldots, L_P$ points respectively and where each set of landmark points are spatially grouped. For instance in the face, the landmark points associated with the nose form one subset. The P regression functions we now learn are denoted by

$$q_p : \mathbb{R}^K \longrightarrow \mathbb{R}^{2L_p} \quad \text{s.t.} \quad q_p(\mathbf{f}) = \mathbf{X}_p \quad \text{for } p = 1, \ldots, P \tag{5}$$

where the learnt mapping is now from an image patch surrounding the particular subset of landmark points to their coordinates defined relative to the centre of the patch.

This means less training data is needed to learn each $q_p(\cdot)$ but they can cover a larger amount of variation for all the parts over learning one global regression function. Using parts, also, ensures that linear regression models are more likely to be a better approximation to the true mapping. As before, if w_p represents the width of the training image patches for part p, then the mapping from an image patch extracted from the bounding defined by $\mathbf{b} = (\mathbf{x}, s w_p, s h_p)$ is:

$$g_p(I_{\mathbf{b}}) = s\, q_p(\mathbf{f}_{I_{\mathbf{b}}}) + \mathbf{x}. \tag{6}$$

While the second problem introduced by this part based strategy is - *Given a novel image, I, how do we find the location and size of each part within I?* Fortunately, for us there are several standard approaches for achieving this and the one we will adopt is based on training part detectors and applying spatial constraints between parts in the manner of pictorial structures [8] which is described next.

3.1 Part detection

We model the appearance of an individual part by fitting a multivariate Gaussian distribution to the part's appearance descriptor. This is a very simple model and forms the basis for the match score, more sophisticated classification scores could be used, but it is sufficient for the face data-sets we examine. The spatial constraints between parts of the face are represented in the form of a tree. In our case, the nose is defined as the root of the tree, and the mouth and eyes are the leaves. The relative location of each pair of connected nodes are then modeled with a 2D Gaussian model. To detect the location of parts in a novel image, a window of a fixed size is slid across the image, and the Mahalanobis distance of the appearance descriptors from the mean model is calculated for each window to create a distance map. The method presented by Felzenszwalb *et al.* [8] is then used to find an optimal solution for the matching problem.

The details are as follows let $\mathbf{b}_p = (\mathbf{x}_p, sw_p, sh_p)$ denote the bounding box of the pth part. The scale factor s is the same for each part and is found by finding the global scale of the face via a frontal face detector such as the one defined by the Viola and Jones face detector. Then the unknown parameters of the bounding boxes $\mathbf{x}_p$ are found by solving:

$$\min_{\mathbf{x}_1,\ldots,\mathbf{x}_P} \left(\sum_{p=1}^{P} m_p(\mathbf{x}_p, I) + \sum_{(i,j)\in E} d_{ij}(\mathbf{x}_i, \mathbf{x}_j) \right) \tag{7}$$

where each $m_p(\mathbf{x}_p, I)$ computes a score of how well the appearance of image patch $I_{\mathbf{b}_p}$ matches the model of part p's appearance

$$m_p(\mathbf{x}_p, I) = (\mathbf{f}_{I_{\mathbf{b}_p}} - \mu_p)^t \, \Sigma_p^{-1} \, (\mathbf{f}_{I_{\mathbf{b}_p}} - \mu_p) \tag{8}$$

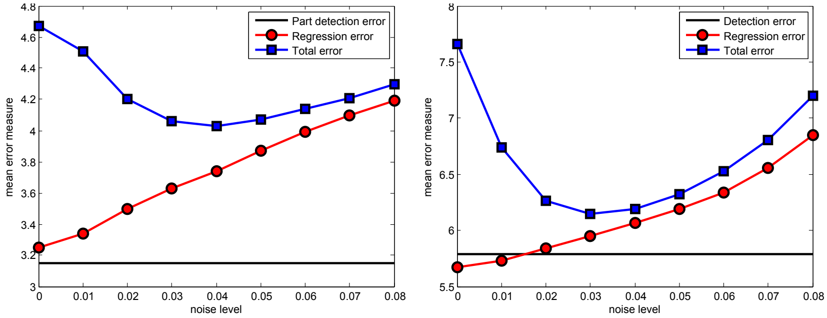
and the appearance is described with the same type of descriptor, $\mathbf{f}_{I_{\mathbf{b}_p}}$, as used in the regression functions. Each $d_{ij}(\mathbf{x}_i, \mathbf{x}_j)$ measures the likelihood of the relative layout of the ith and jth part:

$$d_{ij}(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j - \mu_{ij})^t \, \Sigma_{ij}^{-1} \, (\mathbf{x}_i - \mathbf{x}_j - \mu_{ij}) \tag{9}$$

and E is the list of edges in the tree structure defining the face. As the spatial deformation scores are modelled with a Mahalanobis distance the optimization problem defined in equation (7) can be efficiently solved using distance transforms and dynamic programming as described in [8].

3.2 Learning the part regression functions

For test images there will, in general, be small inaccuracies in the localization of each part and estimation of its scale. Therefore during the training of the individual regression functions we compensate for this fact by augmenting the training data. To do this a zero mean Gaussian noise was added to the location of patches and we created a variety of patches with slightly different scales and positions for every image. The regression model then can learn the mapping between shifted appearance descriptors to the correct shape data. Figure 3 shows a benchmark of algorithm with different levels of noise. As can be seen adding about 3 to 4% noise to the position of parts in the training set, improves the final results.



(a) with parts

(b) without parts

Figure 3: This figure shows the effect of adding noise to the location of the patches used in training on the performance of the regression functions. The *Total error* curve displays the average error in the prediction of the landmark positions when the regression functions use the output from the automatic part detection and the *Regression error* curve the error when instead the ground truth location of the part is used. Observe that perturbing the training data up to a certain noise level increases the regression functions robustness to inaccuracies in the part detection. Note the use of parts (a) significantly increases the accuracy of the algorithm.

4 Experiments and results

In our experiments we have used the IMM face database [14] which contains 240 still images of 40 different human faces with the resolution of 640 × 480 pixels with the average head size 240 × 320. The database includes a variety of facial expressions and head orientations, and contains both male and female subjects. Each image comes with 58 annotated landmarks, outlining the major facial features including eyebrows, eyes, nose, mouth, and jaw. Although in our experiments we only use 44 landmarks (excluding the landmarks around the jaw).

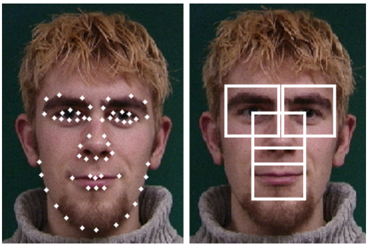


Figure 4: An annotated image from the IMM dataset. Each image comes with 58 annotated landmarks outlining the major facial features. We calculate the bounding boxes for each part based on the location of these landmarks.

We performed 40-fold cross-validation on the IMM dataset to benchmark the performance of our method. 40 different models were trained, for each model we excluded all of the images from one of the test subjects. Furthermore, we have tested our algorithm on novel image sequences and the qualitative results are presented in figure 7.

Four individual part detectors are trained for detecting the left and right eyes, nose, and mouth. We have used 90 × 90 pixel patches to model each part which we found to be optimal in our dataset. After locating the parts, we extract the appearance parameters around the detected point and compute the shape parameters based on the regression model.

The optimal configuration for the appearance descriptor which we found by experiment is $L = 3$, and $b = 5$, where L is the number pyramid levels, and b is the number of histogram bins, which will give a feature descriptor with length equal to $b(1+\sum_{l=1}^{L} 4^{2l}) = 845$. For part detection also we have used similar parameters except the pyramid level which was set to $L = 2$.

There are many options for finding the parameters of the linear regression. Figure 5 shows a comparison of the results from nearest neighbor, ordinary regression, and principal component regression with different dimension reduction rates. The results show that regularization significantly improves the performance of the regression model. The regularization factor is set to $\lambda = 200$ by inspecting its performance on a few validation images. Using PCA with ridge regression gives the top performing results with enough components but the results are not significantly better than ridge regression, although it can potentially improve the speed because of the reduction of the feature dimensionality.

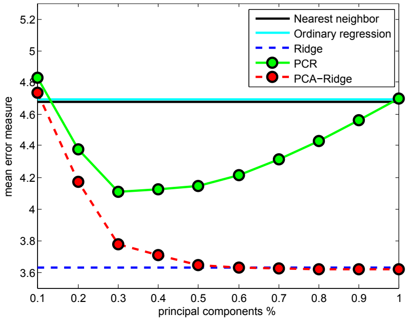


Figure 5: Comparison of performance of different regression methods. The results show that proper regularization significantly improves the results.

Table 1 shows a comparison of performance of our method with several existing methods on the same dataset. The error measure used in benchmarking is the mean of the Euclidean distance between detected landmarks and their ground truth location. The resolution of the images used in these experiments is 640 × 480 pixels, and size of the head is approximately 240 × 320 pixels. The average mean error as a result of cross-validation on the whole dataset in our method is about 4 pixels. This is better than the results of various implementations of AAM and ASM that we found. Figure 6 shows the detected and ground truth location of the landmarks on a different set of test images. We have compared the results with standard AAM implementation by Stegmann *et al.* [14], and CCA-AAM by Donner *et al.* [7], in addition to the results from [11] (ASM, DFM). The results show the superior precision of our algorithm despite the fact that the other methods need initialization with ±10% error.

We have also tried our method on a variety of novel videos, and image sequences and an example is demonstrated in figure 7.

5 Conclusion

In this paper we have introduced a new method for the alignment of faces which has proven to be effective in real world applications. Given 240 images of human faces, we were able

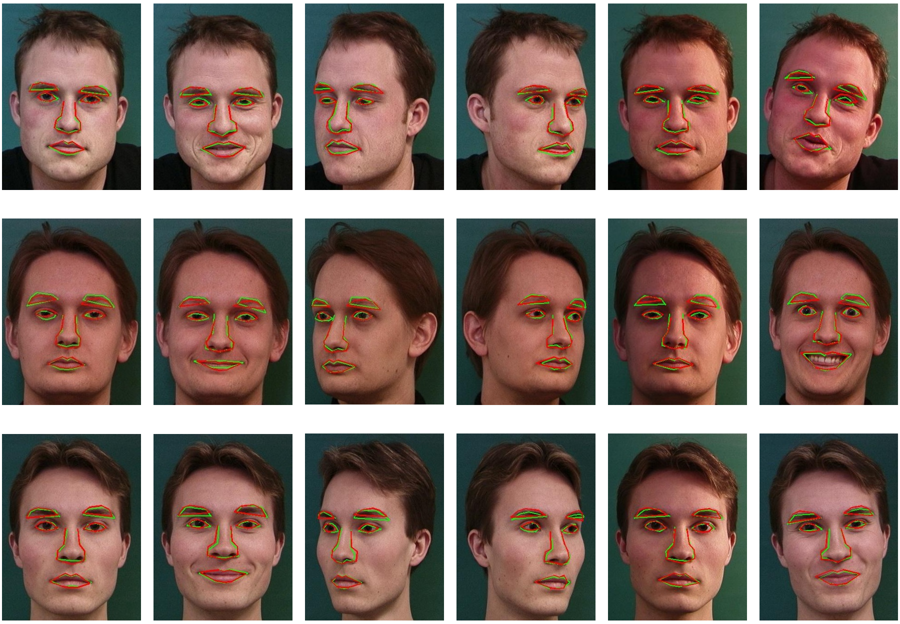


Figure 6: The result of our method on a test set with ground truth information. In this figure the green lines show the ground truth landmarks and the red lines are the predictions of our method.

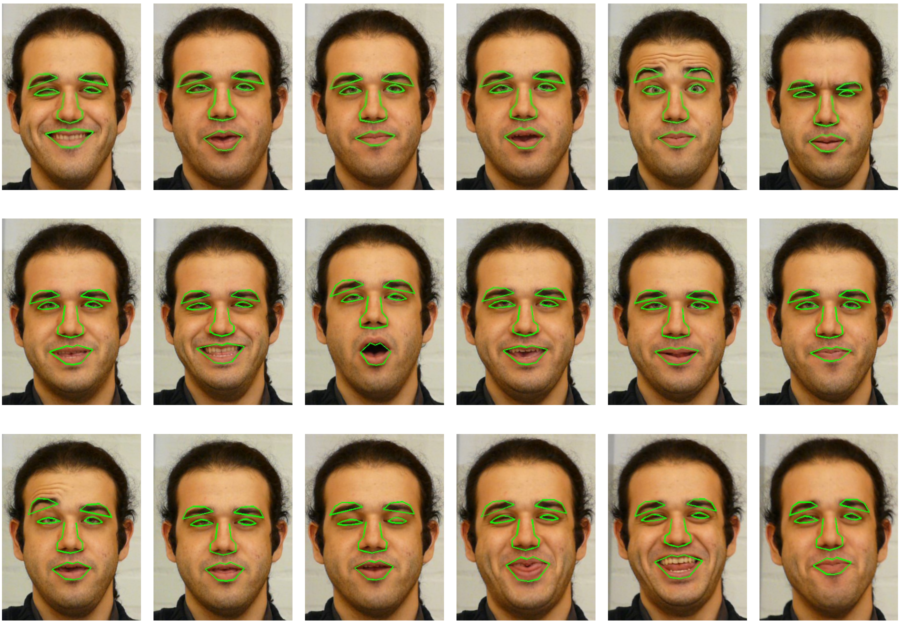


Figure 7: The result of our method, trained using IMM dataset, on an unknown subject, with unknown expressions. The results show that our model covers unseen faces with novel expressions and lighting conditions.

Method	Input Type	Mean error
Our method	Gray	**4.03**
DFM	Gray	4.80
CCA-AAM	Gray	5.70
AAM	Color	5.92
AAM	Gray	6.03
ASM	Gray	6.20

Table 1: Mean point to point error of detected landmarks in full size images. The error measure is the mean of the Euclidean distance between the detected landmarks and their ground truth location in pixels.

to create a model that can be used to align arbitrary faces with acceptable precision. The experiments prove that the method has a good generalization ability and is competitive in terms of precision with global methods such as Active Appearance Models.

As mentioned before the choice of parts is a critical factor in the performance of the algorithm. Further improvement could be achieved by designing an automated method to find the best part configuration. The parts need to be reliably detected and should be evenly spread across the face to give a complete representation of the face's appearance. Furthermore the algorithm can be extended to improve the precision by doing a local search around the detected landmarks to find the optimal match. Alternatively the detected landmarks could be used as an initial estimate for an iterative matching process such as AAM; this can potentially lead to better results since by providing a good initialization to AAM the chances of converging to the global optima increases.

References

[1] B. Amberg, A. Blake, and T. Vetter. On compositional image alignment, with an application to active appearance models. CVPR, 2009.

[2] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. CIVR, 2007.

[3] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models, their training and application. *Computer Vision and Image Understanding*, 61:38–59, 1995.

[4] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. ECCV, 1998.

[5] D. Cristinacce and T. F. Cootes. Boosted regression active shape models. BMVC, 2007.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. CVPR, 2005.

[7] R. Donner, M. Reiter, G. Langs, P. Peloschek, and H. Bischof. Fast active appearance model search using canonical correlation analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(10):1690–1694, 2006.

[8] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79, 2005.

[9] X. Hou, S. Z. Li, H. Zhang, and Q. Cheng. Direct appearance models. CVPR, 2001.

[10] Y. Huang, Q. Liu, and D. N. Metaxas. A component-based framework for generalized face alignment. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 41(1):287 –298, 2011.

[11] A. Kuhl, T. Tan, and S. Venkatesh. Automatic fitting of a deformable face mask using a single image. MIRAGE, pages 69–81. Springer-Verlag, 2009.

[12] R. Okada and S. Soatto. Relevant feature selection for human pose estimation and localization in cluttered images. ECCV, 2008.

[13] Julien Peyras, Adrien Bartoli, Hugo Mercier, and Patrice Dalle. Segmented aams improve person-indepedent face fitting. BMVC, 2007.

[14] M. B. Stegmann, B. K. Ersbøll, and R. Larsen. Fame - a flexible appearance modelling environment. *IEEE Transactions on Medical Imaging*, 22(10):1319–1331, 2003.

[15] A. Thayananthan, R. Navaratnam, B. Stenger, Ph. H. S. Torr, and R. Cipolla. Multivariate relevance vector machines for tracking. ECCV, 2005.