Image and Vision Computing 36 (2015) 40–50

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Editor's Choice Article

Facial landmarking for in-the-wild images with local inference based on global appearance ☆,☆☆

Brais Martinez [a,*], Maja Pantic [a,b]

[a] Computing Department, Imperial College London, UK
[b] Department of Computer Science, Twente University, The Netherlands

ARTICLE INFO

Article history:
Received 27 February 2014
Received in revised form 25 November 2014
Accepted 6 January 2015
Available online 21 February 2015

Keywords:
Facial landmarking
Regression
Part-based models
Gaussian processes

ABSTRACT

We present a novel method that tackles the problem of facial landmarking in unconstrained conditions within the part-based framework. Part-based methods alternate the evaluation of local appearance models to produce a per-point response map and a shape fitting step which finds a valid face shape that maximises the sum of the per-point responses. Our approach focuses on obtaining better appearance models for the creation of the response maps, and it can be used in combination with any shape fitting strategy. Local appearance models need to tackle very heterogeneous data when dealing with in-the-wild imagery due to factors as varying head poses, facial expressions, identity, lighting conditions, or image quality among others. Pose-wise experts are typically used in this scenario so that each expert deals with more homogeneous data. However, the computation cost at test time is significantly increased. Furthermore, choosing the right expert is not straightforward, which can lead to gross errors. We propose to dynamically select at test time the training examples used for inference. We use a global similarity measure to select the most adequate training examples for inference, and create a single test sample-specific expert using a localised inference technique. To illustrate the validity of these ideas, we capitalise on the recently proposed use of regression to generate local appearance models. In particular, we use Gaussian processes, as their non-parametric nature easily allows for localised regression. This novel way of constructing the response maps is combined with two state-of-the-art standard shape fitting algorithms, the popular Constrained Local Models framework and the Consensus of Exemplars method. We validate our method on two publicly available datasets as well as on a cross-dataset experiment, showing a considerable performance improvement of the proposed approach.

1. Introduction

1.1. Literature review

Most of the algorithms for accurate facial landmarking rely on separated models capturing the image appearance and the face shape information. The shape model encodes the possible constellations of facial landmark locations (face shape) and restricts the estimation to be anthropomorphically consistent. Facial landmark detection boils down to finding the valid face shape that maximises an alignment score, which is computed using the appearance model.

It is common to divide facial landmarking methods depending on how the face appearance information is modelled, leading to a distinction between holistic methods and part-based methods. Holistic methods include Active Appearance Models (AAMs) [1,2], and are typically generative methods that try to fully reconstruct the whole face appearance. This is done by using parametric models for the face shape and for the face appearance. Efficient gradient descent can be used to find the optimal parameters, from which facial landmarking results [3]. Instead, part-based methods train a discriminative model per landmark using the appearance of local patches centred at them. Examples of part-based methods include Active Shape Models (ASM) [4] and Constrained Local Models[1] (CLM) [5] and the Consensus of Exemplars [6]. In this work we present a part-based facial landmarking algorithm and, therefore, in the following we focus on the works within this category.

In the most common setting for part-based models, a classifier is trained per facial landmark so that it yields a high score when evaluated at the true target location and low score otherwise. At test time, a region of interest centred at the current landmark estimate is considered, and a response map is built per landmark by evaluating the classifier at every pixel location within the region of interest. This is followed by a shape fitting step, which finds the shape parameters that optimise the sum of individual responses. In most cases, fitting results from alternating both steps. However, some exceptions exist within the part-based facial landmark detection framework that follows variants of this procedure.

☆ This paper has been recommended for acceptance by Qiang Ji.
☆☆ Editor's Choice Articles are invited and handled by a select rotating 12 member Editorial Board committee. This paper has been recommended for acceptance by Dr. Qiang Ji.
* Corresponding author at: Intelligent Behaviour Understanding Group, Department of Computing, Imperial College London, 180 Queens' Gate, London SW7 2AZ, UK.
E-mail address: brais.martinez@imperial.ac.uk (B. Martinez).

[1] In [5] CLMs are presented as a generalisation of ASM, a criterion that we maintain here.

http://dx.doi.org/10.1016/j.imavis.2015.01.004
0262-8856/

For example, [7] showed that it is possible to effectively use a tree-structured shape model so that the global minimum is reached in a single step. Due to this computationally efficient shape fitting procedure, they are able to provide joint face detection, a rough head pose estimation, and facial landmarking. However, the precision of the detection is sometimes hindered by the excessive flexibility of the tree-based model. Furthermore, [8,9] showed that it is possible to directly estimate the shape parameters through regression. To this end, they do not use per-point individual appearance models, and instead obtain a sequence of shape increments by concatenating the per-point appearances and performing regression.

Despite these exceptions, most part-based landmarking works typically aim at improving either the per-point appearance models, or the shape fitting step. For example, the main methodological contribution proposed by Saragih et al. in [5] with respect to previous ASM models lies on the way the shape is fitted to the response maps. Previous ASM methods used a parametric distribution to approximate each response map, using for example a Gaussian or a GMM. Under this approximation, the shape fitting becomes easily tractable and computationally cheap. The CLM uses instead a non-parametric approximation of the response map, and then uses a mean shift algorithm to perform the shape fitting. In practise, the use of a non-parametric representation of the response map allows for capturing complex response maps as those arising in subject-independent unconstrained scenarios. Another improvement to the shape fitting step was proposed in [10], where the authors noted that the mean shift shape fitting process of [5] is prone to converging to local maxima. They proposed to improve the shape fitting step by using a discriminative search of global maxima, where regression models are trained to directly infer the shape parameters by using the responses to the appearance model as input. [6] proposed instead to avoid using a gradient ascent/mean shift shape fitting and use instead a model similar to RANSAC for the shape fitting. To this end, randomly selected shapes are aligned with a subset of the response map modes, computing then the score for the resulting alignment. The output results from combining the n shapes with the highest scores. Other methods focus instead on improving the appearance models so that better response maps can be obtained. Among them, regression algorithms have been recently proposed as an alternative to classifiers to construct the response maps [11–13]. To this end, a regressor is trained per point so that, given an image patch, it predicts its relative location with respect to the true target location. The response map results from considering a probabilistic output of the prediction, for example by using a Gaussian distribution of fixed covariance [13] or a non-parametric pdf [12], and accumulating these predictions in an additive manner. In particular, [12] recently proposed to use random forest within the CLM framework, and directly compared the performance when using regression against the performance when using a classification method, showing the superior performance of the former.

In this work, we aim at improving the quality of the response maps of part-based models. Therefore, our approach can be combined with any response map-based shape fitting algorithm. Because of their recent success, we use a regression methodology to construct the response maps. We use two popular shape fitting algorithms such as [5,6]. It is not in our scope to improve the shape fitting model and, in consequence, any work aiming at improving this step (e.g. [10]) should be seen as complementary to ours.

1.2. Method overview

In our work we tackle the problem of in-the-wild facial landmarking, i.e., facial landmarking of faces captured under uncontrolled conditions. This includes variation factors as illumination conditions, head pose, subject identity or facial expression. A successful in-the-wild facial landmarking method should therefore train its models with examples representative of such variability. However, it is well-known that the increase of the intra-class variability degrades the performance of machine learning algorithms. Therefore, variation factors other than the one targeted (in this case facial landmark locations) are nuisance factors that complicate the inference of the target one.

When it comes to facial landmark detection, it is possible to alleviate this problem by breaking down the learning of the appearance models, and employ a set of pose-wise experts. To this end, the total range of head pose variation is divided into a set of pose ranges. An appearance model is trained using the corresponding subset of the training data, so that each model is trained with less heterogeneous patterns. At test time, this approach requires the evaluation of all of the experts. Then, a criterion to select the best-performing model has to be applied, which typically consists of selecting the expert yielding the highest combined response for the whole set of points. For example, [7] trained head-pose specific models covering a range of -90 to 90° of yaw rotation, each expert corresponding to one of the head poses within the Multi-PIE dataset [14]. Similarly, works as [5,10] use 3 different experts to cover up to 45° of yaw rotation. However, evaluating different models at test time is costly, and an error when selecting the best-performing expert can result in gross landmark detection errors. Furthermore, other nuisance factors, such as facial expression or subject identity, are not accounted for. These factors can still have a significant impact when considering face images captured under unconstrained conditions.

The ideal inference would be achieved if we had an expert (and only one) tailored to the test example at hand, so that all of the examples used for inference present properties similar to the test face. Such similar properties would not only include head pose, but also factors as facial expressions and identity produce important variations on the face aspect. An obvious limitation is that these labels are unknown, both for the training and for the test images. Therefore, an unsupervised method for selecting the examples used for inference is required. Such an alternative would offer a significant potential gain, as unique and more specialised experts can be used. However, the unsupervised nature of the example selection would lead to a suboptimal selection of the examples inference relies on.

In order to accommodate for such idea, we propose to use a local inference method. Local inference methods select, at test time, a set of training examples close to the test example, measured within the (kernelised) feature space, and perform inference based only on the selected examples. The idea is that the decision boundaries are then specialised to the local topology instead of to the full feature space, reducing the impact of intra-class variability [15]. However, as local patches convey little information about head pose or facial expression, we do not seek for a direct application of a local inference method. In order to reason about these factors, it is necessary to consider broader face regions or even the full face appearance. Therefore, we propose to use a local inference technique in which the examples used for inference are local to the test sample in terms of the global properties of the face they belong to.

The advantages of this approach include considering factors of intra-class variability other than head pose variations. Furthermore, the inference cost does not grow as running several experts at test time is avoided. In fact, inference cost is sensibly reduced. As only a small set of the training examples are used at test time, the inference models are sensibly less complex. Finally, there is no risk of selecting the output of the wrong expert. The counterparts are the need of an oracle for selecting the most similar images, that performance depends on the quality of the retrieved training examples, and the need to store the full set of training examples. However, we experimentally show that very simple oracles (even constructed in an unsupervised manner) can effectively pick adequate training examples and boost the performance of the inference attained by the appearance models.

In the work presented here we propose to use Gaussian process (GP) regression for inference, although other methods for local inference exist (e.g. [16]). This is justified by the success of regression-based facial landmarking methods, and because GP naturally allow localised

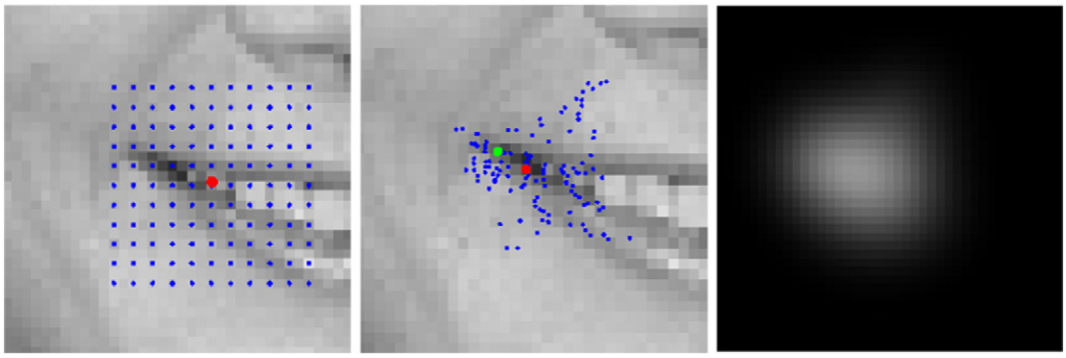

Fig. 1. Left: test locations for the left mouth corner (the red dot notes the current estimate). Centre: predicted locations (blue), maximum of the response map (red) and ground truth (green). Right: the response map constructed from the regression predictions displayed in the centre image.

inference. Their use for local inference is particularly simple as, due to their non-parametric nature, it is possible to modify the set of training examples used for inference without retraining the models. Works as [17,18] have already used such property to train local GP regression models. However, in their case the distance measure used for selecting the examples was the same as for performing inference. From a more general methodological perspective, we argue that the training examples with a feature representation distinct from the one of the test sample have already little influence on the output prediction. The most harmful examples are instead those that look alike in the feature space, but are associated to erroneous labels. We eliminate ambiguous examples (in the kernelised feature space) by using two different distance measures (using potentially different feature representations) that convey complementary information. While one distance measure is used to select the examples for inference, the other is used as the kernelised distance used to compute the predicted output.

The reminder of this paper is organised as follows. Section 2 is devoted to describe how to apply regressors in a part-based facial landmarking framework. Section 3 describes the proposed inference method, including a brief introduction of Gaussian processes and their localised version. Section 4 describes the setup of the experiments, which are later described in Section 5. Section 6 concludes the paper with the final remarks.

2. Regression-based facial landmark localization

Given a test location within a face image, the location of the target facial landmark can be directly estimated by means of two regressors. More specifically, given a test location l, a feature vector $f_l = f(l, I)$ is computed using a small patch from image I centred at l. f_l is then used as the input of two regressors, trained to estimate the real-valued variables Δx and Δy so that $t = l + (\Delta x, \Delta y)$, where t represents the true target location. In order to train such regressors, it is necessary to construct a training set by randomly sampling locations at a distance of up to a threshold (called sampling radius) to the true landmark locations, and extracting feature vectors from these locations. Selecting a small sampling radius yields precise predictions when the test location is within the training set, while it yields very poor predictions otherwise. When using a larger sampling radius, the effect is reversed, and precision is traded for robustness. In order to solve this, two tactics exist in the literature. Either a cascaded regression approach can be followed (e.g. [19,8]), or the regressors can be used to construct a response map, resulting in a part-based landmarking method (e.g. [13,12]). When following this last approach, the regressors are evaluated over a region of interest and their predictions are combined to construct the response map. In this paper we follow the second approach.

More specifically, in order to obtain a response map for a given facial landmark, a region of interest (ROI) is defined around the current landmark location estimate. Then, a set of test locations are defined over it, for example by using a grid layout, and an estimate is computed for each of the test locations. A response map can be built from the obtained set of estimates using Kernel Density Estimation (KDE). In case of a Bayesian regressor, such as Gaussian process or Relevance Vector Regression, the predicted variance can be used as the width of the kernels. A fixed kernel width can be used otherwise. Through this process, it is expected that only the correct estimates will correlate and produce a peak on the response map, while erroneous estimates will not correlate together. An illustration of this process is shown in Fig. 1.

Each iteration of the algorithm combines the construction of the response maps with a shape alignment step. This step can be considered independent of how the response maps are constructed. The shape alignment step consists of finding the valid shape (i.e., an anthropomorphically consistent shape) that maximises the combined individual responses. This is however a challenging maximisation as it is prone to converge to local maxima. As a consequence, many works focus on how to perform shape alignment effectively. In here we outline the shape fitting strategies of [5] and [6], which will be used in the experiments.

2.1. Constrained local models [5]

The CLM parameterises the space of valid face shapes using the Point Distribution Model (PDM). On it, any shape $\mathbf{s}$ can be parameterised as:

$$\mathbf{s} = f_\theta(\bar{\mathbf{s}} + \Phi\mathbf{p}) \tag{1}$$

where $\bar{\mathbf{s}}$ represents the training set mean shape, and Φ is a set of linear shape basis that capture flexible movements, both variables being computed at training time from the set of training examples. f_θ represents an affine transformation parameterised by θ, while $\mathbf{p}$ represents the coefficients of the test shape in the linear subspace spanned by Φ. Therefore, any test shape is fully described under this model by $(\theta, \mathbf{p})$, where θ encodes the rigid transformations of the face shape, and $\mathbf{p}$ encodes the flexible ones (that is to say, anything that cannot be removed through an affine transformation). The shape alignment step aims at finding the shape parameters $(\theta, \mathbf{p})$ that maximise the sum of the individual responses.

The optimal shape parameters are estimated by first finding individual increments on the landmark location estimate by applying a mean shift algorithm over the response map.[2] The individual location

[2] [20] noted that, when using a Gaussian kernel, applying the mean-shift algorithm is a mode-seeking algorithm for the KDE obtained from the point distribution. Another interesting equivalency was shown in [21], where the authors showed that Gaussian mean-shift is an EM algorithm.

increments are then translated into shape parameter increments by using the Jacobian of the shape parameters and solving a least squares problem. This iterative parameter update process is typically applied in a step-wise manner, alternating the estimation of θ and $\mathbf{p}$. This procedure is guaranteed to improve the prediction likelihood, although convergence to local maxima often occurs, probably because of the high parameter dimensionality. Empirical evidence of the convergence to local maxima can be directly derived from the success of approaches that do not follow a gradient ascent/mode seeking strategy to shape fitting, such as [6,10].

2.2. Consensus of Exemplars [6]

This approach relies on the use of a set of shape exemplars, for example the shapes within the training set. New shapes can however be produced by instantiating the shape model within Eq. (1). There is no explicit parameterisation of what a valid shape is. Instead, face alignment is attained at test time as follows. The best k modes of each response map are first computed. Then, the following process is repeated a fixed amount of times. First, a subset of the modes are randomly selected, with the restriction that each mode must relate to a different landmark. Then a random exemplar is aligned with these locations by finding the affine transformation minimising the average point-to-point distance. A score for the resulting alignment can be then computed in terms of the combined responses. Lastly, the final shape alignment is computed as the mean of the best n shapes. The validity of the shape output follows from the fact that it is computed as the average of shapes seen in the training set.

3. Proposed inference method

3.1. Gaussian process

Gaussian process (GP) [22] is a Bayesian method capable of performing regression using complex non-linear maps through the use of kernels (called covariances in the GP context). It is a non-parametric method, so that inference is performed by using the whole training set. This is as opposed to the case of other regression methods such as Support Vector Regression or Relevance Vector Machines, where given the support vectors computed during training, inference is independent of the rest of the training set. The only parameters to be estimated during the GP training are the covariance hyper-parameters, which can be optimised through an efficient gradient descent procedure. One advantage of GP is that they provide a probabilistic output, parameterised through a Gaussian distribution. That is to say, the output is a full pdf in the form of a normal distribution that models the probability for a given output value of being the true label. The equations for inference, derived and explained in detail in [22], take the form of:

$$\mu_i = k_*^T\left(K + \sigma_n^2\right)^{-1} y \tag{2}$$

$$\sigma_i = k(\mathbf{x}_*, \mathbf{x}_*) - k_*^T\left(K + \sigma_n^2\right)^{-1} k_* \tag{3}$$

where K represents the covariance between training points, k_* represents the covariance between the training examples and the test samples, and $k(\mathbf{x}_*, \mathbf{x}_*)$ is the covariance between test samples. It is possible to see from these equations that the full set of training examples has to be kept, and that its inversion is necessary. Therefore, GPs scale badly to the number of examples. Despite the efforts of many works that attempt to reduce the impact of the number of examples (e.g. [23]), this remains as a major drawback of GPs.

We refer to [22] for an in-depth discussion of GP, including how to find the optimal covariance parameters. However, some properties are particularly relevant to our approach, which we summarise in the following. In the first place, GP provide a measure of confidence on the prediction. Samples at test time unseen in the training set, as for example those corresponding to partial occlusions, yield a higher prediction variance. Secondly, inference is performed using only the covariance hyperparameters and the set of training examples. In the third place, GP scale badly in terms of the number of examples, so it is important to carefully select the examples used for inference and ensure that they are meaningful.

3.2. Local GP

Local GP (LGP) aims at performing inference using a subset among the training examples closest to the test sample, either in Euclidean or kernelised distance. The benefits of local inference are the use of locally optimal decision boundaries instead of the globally (on average) optimal ones, the use of less data, potentially reducing the computation time during inference, and the use of more meaningful examples for inference. GPs are particularly suited for localised learning due to its non-parametric nature. That is to say, as only the covariance hyperparameters and the training examples are necessary to perform inference, no re-training is necessary when the training set is altered.

More formally, if we note the GP regressor as r, the training set as $\{X_{tr}, Y_{tr}\} = \{(\mathbf{x}_{tr}^i, y_{tr}^i)\}_{i=1:N}$, and the covariance hyperparameters as θ, then we can write:

$$(\mu_*, \sigma_*) = r(\mathbf{x}_*|\theta, X_{tr}, Y_{tr}) \tag{4}$$

where $\mathbf{x}_*$ is the test feature vector.

We can compute the k nearest neighbours of $\mathbf{x}_*$ within X_{tr} using a (kernelised) distance

$$d_i = K\left(\mathbf{x}_*, \mathbf{x}_{tr}^i\right). \tag{5}$$

Then the vicinity of the test feature vector used for inference is:

$$(X_{tr}(\mathbf{x}_*), Y_{tr}(\mathbf{x}_*)) = \left\{\left(\mathbf{x}_{tr}^i, y_{tr}^i\right) \text{ s.t. } \quad i \in I_k(x_*)\right\} \tag{6}$$

where $I_k(x_*)$ represents the indexes corresponding to the k lowest values of $\{d_i\}_{i=1:N}$.

For LGP, both the training examples and the hyperparameters used for inference are conditioned to the test feature vector:

$$(\mu_*, \sigma_*) = r(\mathbf{x}_*|\theta(\mathbf{x}_*), X_{tr}(\mathbf{x}_*), Y_{tr}(\mathbf{x}_*)). \tag{7}$$

The local hyperparameters $\theta(\mathbf{x}_*)$ depend now on the test feature. The locally optimal hyperparameters can be approximated without resorting to re-training for each new test sample [18]. In practise, this training process is time consuming and we found a very small improvement with respect to using globally optimal hyperparameters. Therefore, we use globally optimal hyperparameters throughout our experiments.

3.3. Local GP based on global face appearance

Our method proposes to substitute the distance used to define the neighbourhood of the test example (Eq. (5)) for a distance measure that takes into account the whole face appearance. The region of the image used to extract the face appearance depends on the ground truth shape for the training examples, and on the current shape estimate for the test example. More specifically, given the current shape estimate $\hat{\mathbf{s}}_*$, we need to construct a similarity measure depending on it, noted $S(I_i, \mathbf{s}_i, I_*, \hat{\mathbf{s}}_*)$, capable of measuring the similarity of the test image with respect to the images within the training set. The similarity should be high whenever the properties of I and I_* are similar, but without requiring the estimation of any variable such as the head pose explicitly.

The process to obtain an appearance feature representation h_i for image i from a shape $\mathbf{s}_i$ is as follows. First we register the manually annotated face shape (or estimated shape) of image i to the mean shape of the training set using an affine transformation, yielding the registered shape $\mathbf{s}_i^{reg}$. The same affine transformation can be readily applied to the training (test) image, yielding the registered image I_i^{reg}. Then we define $P_i = P(I_i^{reg}, \mathbf{s}_i^{reg})$ as a subpatch of I_i^{reg} defined by the bounding box tightly containing the points $\mathbf{s}_i^{reg}$. Finally, a feature representation $\mathbf{h}_i = f(P_i)$ is extracted from it.

The similarity between images is then computed as a (kernelised) distance over such features:

$$S(I_i, \mathbf{s}_i, I_*, \hat{\mathbf{s}}_*) = K(\mathbf{h}_i, \mathbf{h}_*). \tag{8}$$

Since we want to consider variations on the head pose and facial expressions we compute, for each image, appearance descriptors for the whole face, the eye and eyebrows region, the mouth region and the nose region. In practise, we use the whole face similarity for all the points, while the appearance of the facial components is used only if the target landmark lies within them.

We then define the locality in terms of the global image similarity as:

$$X_{tr}(\mathbf{x}_*, \hat{\mathbf{s}}_*) = \left\{ \mathbf{x}_j \text{ s.t. } \mathbf{v}(j) \in I_k(\mathbf{h}_*) \right\} \tag{9}$$

where $\mathbf{v}(j)$ is a pointer vector that indicates the training image from which feature vector j was extracted. Then inference is performed according to Eq. (7).

A visual depiction of the inference process is shown in Fig. 2. This figure shows that a feature representation representing the global face appearance is extracted from the test image and the set of training images (note that the shape used to extract the feature representation of the test image is approximated). Then the nearest images in the feature space are selected (marked in red), and examples drawn from these images are used to perform inference.

Throughout our experiments, we use a simple LBP feature representation of the face, which has been applied successfully to problems as facial expression recognition [24] and face recognition [25]. We use the intersection kernel in Eq. (8), as it is adequate for histogram-based feature representations as LBP, and no hyperparameters have to be estimated. More complex learning-based algorithms could be applied instead or in combination with this descriptor. For example, supervised learning would likely boost the performance of head pose estimation. However, it is not easy to training head poses in in-the-wild conditions, as there is no reliable ground truth in these cases. We show however that a simple unsupervised measure is enough to boost the results.

4. Experimental setup

4.1. Initialisation

A Viola and Jones (V&J) face detector was used to initialise the algorithm. In case that the face was not correctly detected, a face bound was inferred by finding the closest shape among the training examples with successful face detection. Then, the translation and scaling best aligning both manually annotated shapes was computed and applied to transfer the face bound to the undetected face. Through this process we avoid introducing a bias in the results by excluding the most non-frontal and non-standard faces, and the initial fitting error compares to that of similar images. The face region is then resized to 100×100 pixels. It is important to note that normalising to a larger face size typically increases the accuracy of the method, but also increases the computation cost. Different methods report results using different normalisation sizes, and 100×100 is a small and conservative one. Finally, the mean shape was used to initialise the search algorithm. The mean shape was computed through generalised Procrustes Analysis from the manually annotated shapes on the training set.

4.2. Feature representation

We use HOG features [26] as the local appearance descriptor throughout our experiments. HOG features are widely used for facial landmark detection (e.g. [7,10]), so using them improves the significance of the comparisons. The patch sizes used are of 16×16 pixels, yielding a feature vector dimensionality of 81. Applying PCA over the descriptor and keeping around 30 dimensions yields a small improvement, and reduces the computational and storage cost at test time. However, we keep the full dimensionality of the HOG descriptors in our experiments so that the comparison with other state of the art experiments is more meaningful.

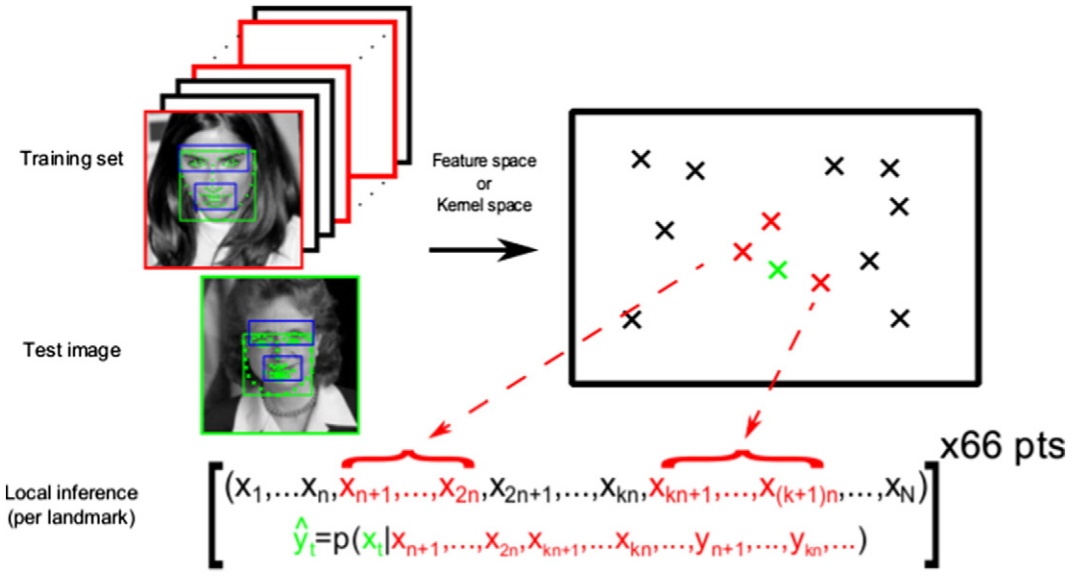

Fig. 2. Depiction of the inference process: the training set (red and black) and test (green) face images are mapped into the feature or kernel space. Nearest neighbours are performed (closest are marked in red). Only the training examples from the "nearest faces" are used for inference (there are 66 inference processes, one per facial point).

Table 1
Inference performance over the MultiPIE dataset, measured in average Euclidean distance of the predicted landmark location (mean/standard deviation). *Ideal expert* refers to inference only using examples with the same head pose and facial expression labels as those of the test image (test image labels need to be used at test time).

GP	Our method	Ideal expert
3.43/2.02	3.06/1.81	2.95/1.79

The global appearance is represented using a uniform-pattern LBP descriptor [27]. LBP descriptors are popular for face analysis, and they have been successfully applied to face identification [25] and facial expression recognition [28]. The LBP descriptor has interesting properties for our problem. An LBP descriptor results from histogramming over a region, so all of the spatial information within the region is discarded. This provides some in-built robustness to misalignment. In our algorithm, the extracted global appearance depends, at test time, on the estimated shape, which is only approximate. Therefore, invariance to misalignment is of particular importance. Furthermore, LBP descriptors are insensitive to global uniform illumination changes, and they are robust to general illumination changes. This is also important for our purpose, as we deal with imagery obtained with unconstrained illumination conditions. Specifically, we use a block-based representation where each patch is divided into sub-patches in a grid manner. An LBP descriptor is computed for each sub-patch, and the patterns are concatenated into a single feature vector. The use of a block-based representation significantly improved performance in some preliminary experiments.

4.3. Training set creation

We start by registering every face in the training set to the mean shape using a Procrustes transformation. That is to say, we register the training faces as much as possible using rigid transformations. The mean shape is computed in face-size-normalised coordinates so the resulting registered face is about 100×100 pixels. Then, for each landmark, we sample a number (15 in our case) of random locations around at up to 10 pixels displacement horizontally and vertically in the coordinate system of the registered faces. A HOG descriptor is then computed at each sampled location using a 16×16 patch. The training labels result from subtracting the sampled locations from the location of the ground truth.

4.4. Evaluation criterion

In order to normalise the error with respect to the differences in face sizes, we divide the point-to-point L_2 error by the inter-ocular distance (IOD), leading to the IOD-normalised error. This is the most standard way of presenting the fitting error [13,29,12] and although it penalises non-frontal head poses, it facilitates comparison. It is important to note that some works (e.g. [7]) use a different error-normalisation criterion. Some works (e.g. [13]) further distinguish between the facial landmarks lying on the contour of the face and those within the facial components. Contour landmarks are hard to define objectively for manual annotation, and they are also hard to distinguish during automated detection. Fittings that look correct to the naked eye might still present high errors for the facial landmarks within the face contour. We therefore report separate errors for both the whole set of landmarks (66 in our case) and for only the landmarks lying within the facial components (49 in our case).

Table 2
Quantitative results over the LFPW (Inlayers/all points).

Method	Mean	Std.	Median
Prop + ConsEx	0.058/0.072	0.022/0.031	0.052/0.064
GP + ConsEx	0.066/0.084	0.035/0.047	0.057/0.071
Prop + CLM	0.063/0.075	0.024/0.034	0.059/0.066
GP + CLM	0.073/0.086	0.034/0.044	0.063/0.072
CLM-ITW [5]	0.080/0.091	0.026/0.033	0.073/0.084
DRMF [10]	0.070/0.081	0.022/0.025	0.066/0.076
SDM [9]	0.051/0.069	0.019/0.027	0.047/0.063

5. Experimental results

In this section we include the following set of experiments. Firstly, we conduct an experiment on a dataset with controlled conditions in which labels as the facial expression and the head pose are provided. We use this setting to perform preliminary experiments showing how our inference method performs against a golden standard where information such as the head pose or the facial expressions are known in advance at test time. That is to say, we compare against an algorithm in which the example selection method performs perfectly. We then show the performance of our algorithm on imagery captured under uncontrolled conditions using two standard in-the-wild datasets. These experiments can be divided into dataset-dependent experiments and cross-dataset experiments. In the former, each dataset is divided into training and testing partitions as specified by the dataset protocol. The training partition is then used to train the models, and the performance is computed over the test partition. Instead, in the cross-dataset experiments we train on the training partition of one dataset and evaluate on the testing partition of the other. These experiments are designed to show the generalisation capabilities of the proposed algorithms. All experiments are subject-independent, so that the any

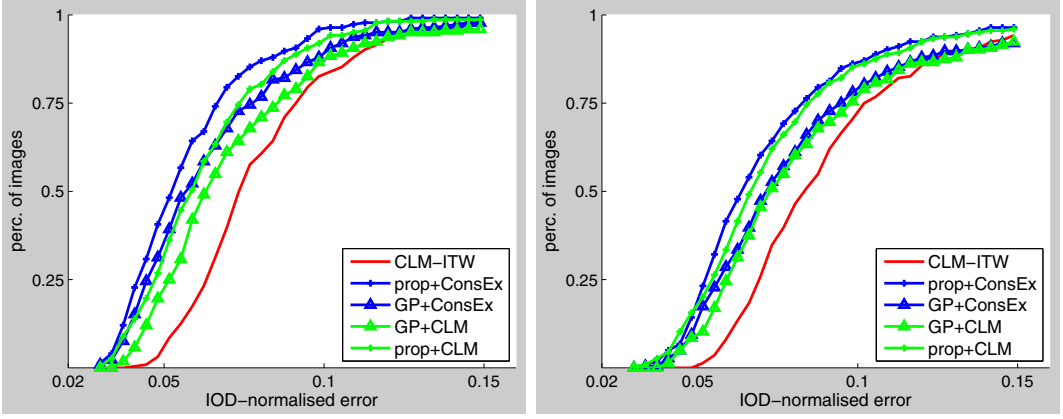

Fig. 3. Cumulative iod-normalised error distribution on the LFPW dataset. Left: error for inlaying points. Right: all 66 landmarks (including contour points). See text for the acronyms.

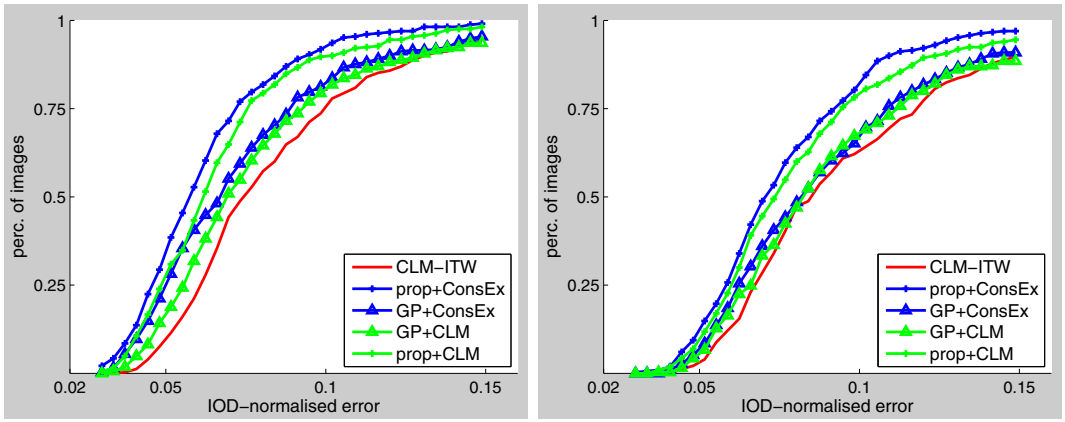

Fig. 4. Cumulative error distribution on the Helen dataset. Error for inlaying points (left) and for all 66 landmarks (right).

subject present in the testing set cannot be included on the training partition.

5.1. Experiments on data under controlled conditions

The proposed method relies on the performance gain obtained by the proposed localised inference method. Therefore, the first experiment consists on studying the performance gain of such approach with respect to standard inference methods. That is to say, with respect to methods where all the training set is used for training the regressors that are later applied irrespective of the test image properties. Furthermore, we want to evaluate the inference performance when the training examples used for inference are known to have the same head pose and facial expression as the ones in the test example. That represents a golden standard in which the right head pose and expression-specific expert is used. We therefore refer to it as an *ideal expert*. In this case, the labels for head pose and facial expression need to be known in advance for both the training set and the test example, which is obviously an unrealistic setting. We compare these two approaches to our approach, where the examples used for inference are chosen depending on the appearance of the test sample in a data-driven manner.

In order to carry out these experiments we need a dataset where the head pose and facial expression labels per image are given. To this end, we use the MultiPIE dataset [14]. It contains a large set of images captured under controlled conditions. Each image on the dataset has an associated label indicating the head pose, taken from a discrete set of possible head poses, and the facial expressions displayed, taken again from a discrete set of predefined posed facial expressions (e.g. smile, neutral, scream, etc.).

For this experiment, we used a large set of annotated images from the MultiPIE dataset with head poses ranging from −30 to 30° and of varied facial expression. Then we constructed a training set as previously explained (see Section 4). Subsequent training and testing partition are then constructed using a subject-independent cross-validation strategy. Therefore, a different set of regressors are constructed for each of the subjects considered in the experiment. The error in this case is measured as the average Euclidean distance between the predicted location and the true landmark location of the test examples. The results for this experiment are shown in Table 1. It is possible to see that our method attains very close performance compared to choosing the ideal expert, despite not making explicit use of the head pose and facial expression labels of the test sample.

5.2. Dataset-dependent experiments

In this section, we provide a quantitative evaluation of the proposed method. We consider that our baseline method is the one building the response maps through standard GP regression. This results in two methods when combined with the CLM and the Consensus of Exemplars shape alignment strategies, which we note here as GP + CLM and GP + ConsEx. These baseline methods are compared against the combination of the proposed response map construction strategy and the same shape alignment strategies, resulting in the methods referred as proposed + CLM and proposed + ConsEx. However, we also include another baseline in which the CLM shape alignment strategy is combined with classification-based response maps. In order to provide a fair comparison, these classifiers have been trained for in the wild conditions. This last baseline is in fact a replication of the experiments presented in [12], and serve as a confirmation of their findings; that regression-based response maps are more precise than classification-based ones. This method is noted here as CLM-ITW. Furthermore, we provide a performance comparison with two state of the art methods. We compare against the method proposed in [10], noted as DRMF, and the method presented in [9], noted as SDM. The former is included as it reports the best performance for response map-based methods. The latter is a direct method and as such it does not use response maps. It is however included here as it is the current overall state of the art. In order to obtain a fair comparison, we have trained this method with the same dataset and number of training samples as our method, and used the same face detector so that the initial shape is the same.

Our tests are performed on two datasets, the LFPW [6] and the Helen [29] datasets. The authors of the LFPW dataset specified a partitioning of the images into a training and a testing partition, a division that we follow in our experiments. However, only the URLs of the images are provided. Out of the total of 1100 training and 300 test images, we were only able to retrieve around 700 images for training and 224 images for testing. The images within this dataset have a large range of variation, including non-frontal head poses, lower resolution images, varying expressions (although a significant portion of them are polite smiles), ethnic background and illumination conditions. Fig. 3 shows the performance comparison with and without the proposed example selection strategies. It is possible to observe a significant improvement irrespective of which shape alignment strategy is used. Furthermore, it is clear that its combination with the Consensus of Exemplars provides the best performance overall. Finally, this figure highlights the large performance improvement when using regression-based response maps with respect to the use of classifier-based ones. Table 2 provides quantitative results in terms of the mean, standard deviation and median[3] errors, both for points lying inside the face and for all the points

[3] Large errors have a large impact on the mean error, while the median is more robust to gross errors. This is interesting as the error associated to gross misdetections can be somewhat arbitrary.

Table 3
Quantitative results over the Helen dataset of different methods for both the landmarks lying within the face and all the landmarks (including those lying on the face contour).

Method	Mean	Std. dev.	Median
Prop + ConsEx	0.062/0.078	0.024/0.029	0.058/0.070
GP + ConsEx	0.074/0.092	0.035/0.042	0.067/0.081
Prop + CLM	0.067/0.083	0.028/0.035	0.062/0.073
GP + CLM	0.079/0.095	0.035/0.044	0.069/0.082
DRMF [10]	0.067/0.081	0.026/0.034	0.062/0.074
CLM-ITW [5]	0.085/0.097	0.035/0.041	0.074/0.085
CompASM [29]	0.091	–	0.073
SDM [9]	0.058/0.075	0.021/0.026	0.053/0.067

(including landmarks that lay in the face contour). It is possible to observe from these numbers that the proposed method yields the best performance for response map-based methods, while it trails behind the SDM method.

The Helen dataset is pre-divided into 330 images for testing and 1158 images for training. It contains images of higher resolution than the LFPW dataset, but with a larger variation in the facial expressions and head poses. It is important to note that a shape model based on the PDM can have trouble fitting the examples in the Helen dataset, as for example some subjects are pulling faces to the camera or can present asymmetric facial expressions as winking only one eye. This comes as no surprise as this dataset was introduced in [29], where the authors proposed a more flexible shape model, and the Helen dataset was designed to show the benefits of their approach. Fig. 4 shows the performance comparison with and without the proposed example selection strategies, both for landmarks that lay within the face and for all landmarks. It is possible to observe that the relative improvement when using the proposed method is even larger than for the LFPW dataset. This might be due to the relatively larger range of facial expressions. The same observations can be made again in terms of the superior performance of the Consensus of Exemplars and regression-based response maps. Similar quantitative results are shown in Table 3. Again, the relative performances are very similar to those on the LFPW dataset. However, we include here the reported performance obtained by method presented in [29], noted as CompASM.

5.2.1. Per-landmark average error

Fig. 5 shows the per-landmark error for the Helen dataset. The error statistics on the left hand side figure are computed using the proposed method in combination with the Consensus of Exemplars. The radii of the circles are proportional to the landmark error. The face image depicted only serves illustrative purposes. It is possible to see that, unsurprisingly, the contour points and the outer part of the eyebrows are the ones resulting in the largest errors. The right hand side graph shows instead the relative improvement attained per landmark by the proposed method with respect to the GP + ConsEx baseline. The x axis indicates the landmark index, and labels are given as to relate each index to a specific facial component. It is possible to see that the largest improvement is attained for the nose and the mouth regions. The results for the LFPW dataset are not included in this document as they yield almost identical per-landmark error graphics.

5.2.2. Influence of the number of examples used

We now analyse the performance of the inference algorithm as a function of the number of examples used for inference. Specifically, the number of examples depends on two variables: the number of (globally similar) images selected at test time, and the number of training examples extracted from each of them. Our aim here is to show two properties. 1) When fixing the number of images selected, performance saturates at a relatively low number of training examples. This means that the matrix inversion at test time is still computationally cheap (see Section 5.5); and 2) selecting a small subset of images at test time (20 in our case) yields the best performance, highlighting the efficacy of the image selection strategy.

In order to show 1), we constructed a training set following the procedure specified in Section 4 using the dataset training partitions. The same procedure was followed on the test partitions to create a test set on which to evaluate the performance of the inference method. Therefore, we measured the inference error rather than the overall landmarking performance. Fig. 6 (left) shows the result of this experiment for both the Helen and LFPW datasets. Throughout the experiments presented in this work, we used a total of 300 training examples (20 images, 15 examples per image), as it offers a good trade-off when considering the computation cost.

Conversely, Fig. 6 (right) shows the overall performance over the LFPW dataset as a function of the number of images selected (leaving the total number of training examples used for inference fixed). This graph clearly shows how performance is minimised when using a small number of images, with performance increasing monotonically thereafter. When the number of training examples is the full training dataset, then the method is reduced to performing standard GP, which is our baseline.

5.2.3. Initial iterations

Inference, as proposed in here, depends on the current estimate of the facial landmarks. These estimates are typically poor for the first iteration of the algorithm, as they are initialised based on the Viola and Jones face detection. Thus, it is reasonable to wonder 1) whether performance is still better for the proposed inference method for early iterations, and 2) whether a poor first estimate is particularly damaging for our method.

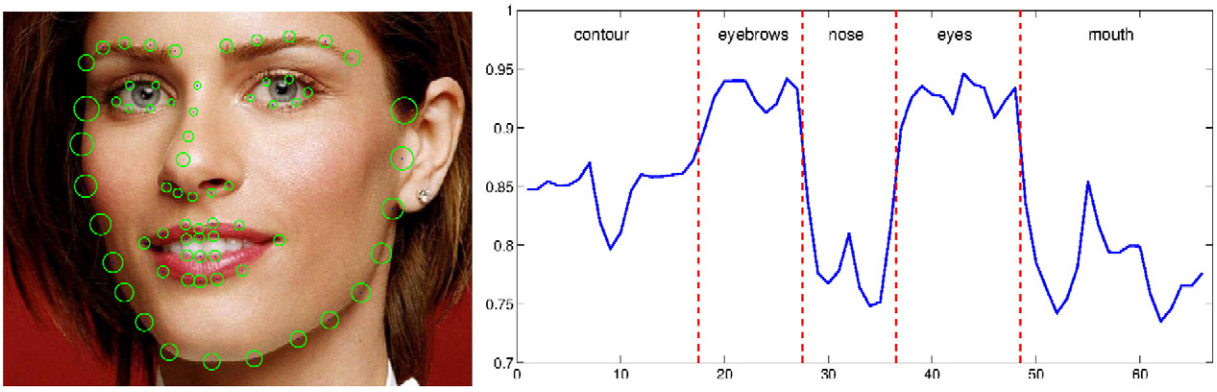

Fig. 5. Left: per-landmark error of the proposed method on the Helen dataset. The radius of each circle is proportional to the average error for the corresponding landmark. The image is used to illustrate the location of each point. Right: relative improvement, measured as the ratio of per-landmark error, between the proposed + ConsEx and GP + ConsEx, together with labels of the face component they belong to.

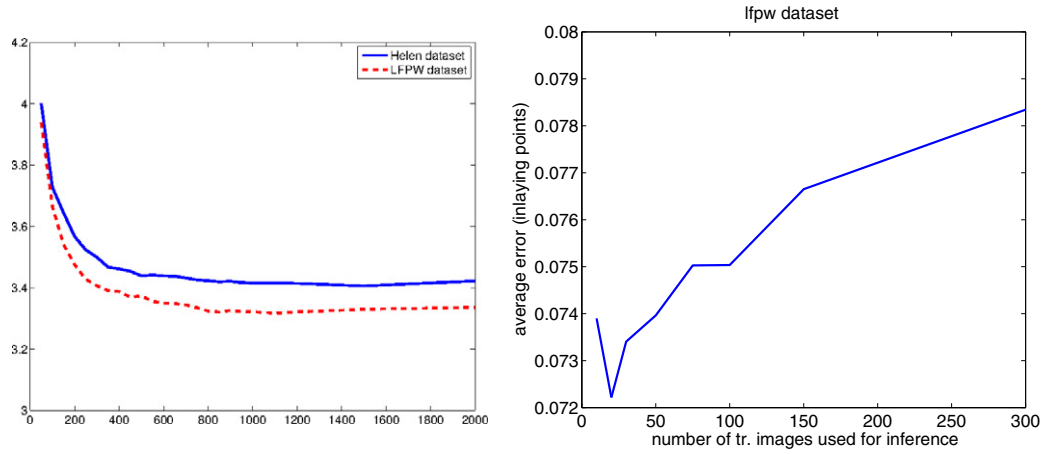

Fig. 6. Left: mean inference error (y axis) vs. number of training examples selected for inference (x axis) for both the LFPW and Helen datasets. Right: overall performance on the LFPW dataset with respect to the number of images selected at test time. This experiment was conducted for the selection of 10, 20, 30, 50, 75, 100, 150, and 300 images (out of a total of 717).

Even if 1) was the case, it would still be feasible to perform a small number of iteration using standard inference methods, and resort to the proposed inference method on the later stages. However, we have experimentally found that our method yields a consistently better estimate after the first iteration, although sometimes only marginally better. Specifically, relative performance after the first iteration shows a relative improvement when using the proposed inference algorithm within the range of 1% to 5%. This includes the LFPW, Helen and cross-dataset experiments, and is the case both for the subset of inlaying landmarks and for all landmarks.

Regarding question 2), Fig. 7 shows the initial error (x axis) vs. the final error (y axis) for every image in the LFPW dataset. The left-hand side graph shows performance for the proposed inference method, while the right hand side shows performance for the baseline inference method. Furthermore, linear regression has been performed as to highlight the underlying relation between initial and final error. The dashed line corresponds to the linear regression for standard inference for the ease of comparison. It is possible to see in this figure that our method does behave comparatively better for cases of poor initialisation.

5.3. Cross-dataset experiments

In order to assess the generality of the proposed method, we include a cross-dataset evaluation using the LFPW and the Helen datasets. In this experiment, we train on the training partition of one of the datasets, and we compute performance over the testing partition of the other dataset. Fig. 8 shows the performance of the cross-dataset experiment, and compares it with respect to the dataset-dependent one, i.e., when the training and test sets are constructed from the same dataset. We employ here the Consensus of Exemplars shape alignment strategy, as it is the one yielding the best overall performance. It is possible to see how training on the Helen dataset and testing on the LFPW dataset yields a comparable performance than training and testing on the LFPW dataset, while this is not true when inverting the roles of both datasets. This might be an indication that, as hypothesised in [30], training with cleaner data (as the one in Helen dataset) yields better-performing models than training with more noisy data, even when applied to a similarly noisy test image. The wider range of expressions displayed on the Helen dataset might be another factor explaining this performance difference, while it is also important to note that the number of training images is significantly larger on the Helen dataset (717 vs. 1147 in our case).

5.4. Qualitative results

Finally, we provide some qualitative fitting results in Fig. 9. This is useful for understanding the nature of the images included in the datasets, and for illustrating the visual meaning of the fitting errors. To avoid cherry picking, we show the two best and worst fits (leftmost and rightmost columns) and the tertiles of the error. That is to say, we show the images with lower error than 2/3 of the test images (centre left column) and error lower than 1/3 of the test images (centre right column). This is shown for the LFPW dataset (2 upper rows) and the Helen dataset (lower 2 rows) for results obtained with the proposed

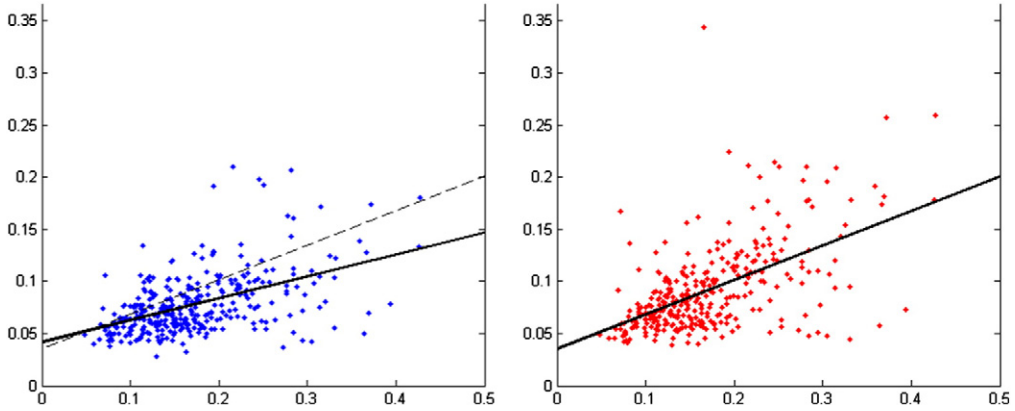

Fig. 7. Initial (x-axis) vs. final error (y-axis) on the LFPW when using the proposed inference approach (left) and standard GP (right). Linear regression was performed as to highlight the trend.

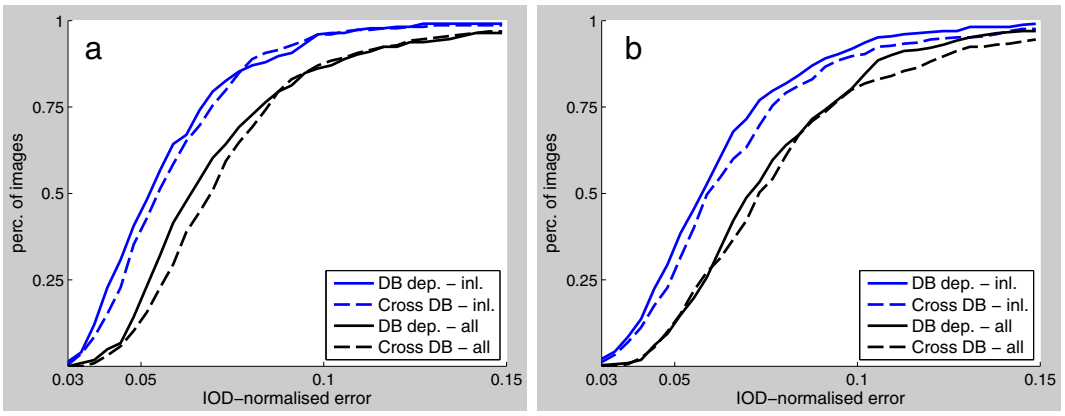

Fig. 8. Cumulative error distribution when (a) training on the Helen dataset and testing on the LFPW dataset (b) training on the LFPW dataset and testing on the Helen dataset. Separate curves for inlayers and all landmarks are included.

algorithm in combination with the Consensus of Exemplars (1st and 3rd rows), and for the GP + ConsEx method (2nd and 4th rows).

It is interesting to see where the algorithm produces the worst results, i.e., the images in the fourth column. In particular, failures might happen when the initial shape is far from the true shape (for example because of poor face detection, as in the top right image). Furthermore, partial occlusions are not directly handled in this algorithm, and can produce poor results (see the third column). It is worth noting that the failures shown for the GP + ConsEx method are mostly due to non-frontal head poses, while the failures for the proposed algorithm are related to either a poor initialisation or a partial occlusion. This might be due to the use of more pose-specific examples for inference on our algorithm.

5.5. Computational cost

A typical way of speeding up inference with GP is to pre-invert the covariance matrix. The cost of the matrix inversion is dominated by the Cholesky decomposition, which is $\mathcal{O}(n^3)$, where n is the dimensionality of the matrix. This pre-computation of the inverse is not possible in our case. Furthermore, it is necessary to perform a vector comparison

Fig. 9. Examples of fitting outputs. Errors measured in IOD-normalised distance. Rows (top to bottom): GP + ConsEx on LFPW, Proposed + ConsEx on LFPW, GP + ConsEx on Helen, and Proposed + ConsEx on Helen. Columns (left to right): best fit, first tertile, second tertile and worst fit (the first tertile is the image with an error better than 2/3 of the images and worse than the remaining 1/3, conversely for the second tertile). Error (top to bottom, left to right): 0.031, 0.031, 0.039, 0.028, 0.062, 0.056, 0.060, 0.056, 0.085, 0.075, 0.100, 0.084, 0.475, 0.295, 0.344, and 0.209.

per example in the training set. This is however a relatively low number (approx. 700 in LFPW and 1100 in Helen). This is done once for all the points and, furthermore, it is possible to use algorithms for efficient nearest neighbours. In contrast, once the examples used for inference are selected, our method needs less vector comparisons. Specifically, the covariance between two examples is $\mathcal{O}(d)$, where d is the feature dimensionality. Therefore, inference has (approximately) complexity $\mathcal{O}(dn + n^2)$. The bottleneck (the Cholesky decomposition) depends on n. When applying the proposed inference method, n is much lower (as low as 300) than when performing inference with standard GP, compensating to a large degree for the need of matrix inversion.

6. Conclusions

We have proposed new method for facial landmarking that is particularly suited for dealing with the highly varying nature of in-the-wild images. In particular, our work lies within the part-based models, and focuses on improving the quality of the response maps obtained from local appearance. In the first place, we have corroborated that regressors are suitable for constructing precise response maps and, at the same time, we have identified one problem downgrading performance for such approaches. Our solution combines a local regression framework with a measure that captures global image similarity, and has the ability to perform inference with models specific to the test sample at hand. Several experiments confirm the improvement attained by such approach when combined with two state of the art shape alignment strategies. Furthermore, we provide extensive comparisons with the state-of-the-art methods. Finally, we also provide cross-dataset experiments that show that our results generalise well to unseen data.

Acknowledgements

This work has been funded in part by the Engineering and Physical Sciences Research Council (EPSRC) project EP/J017787/1 Analysis of Facial Behaviour for Security in 4D (4D-FAB). The work of Brais Martinez is also funded in part by the EPSRC grant EP/H016988/1: Pain rehabilitation: E/Motion-based automated coaching.

References

[1] I. Matthews, S. Baker, Active appearance models revisited, Int. J. Comput. Vis. 60 (2) (2004) 135–164.
[2] G. Tzimiropoulos, J. Alabort, S. Zafeiriou, M. Pantic, Generic active appearance models revisited, Asian Conference on Computer Vision, 2012, pp. 650–663.
[3] G. Tzimiropoulos, M. Pantic, Optimization problems for fast AAM fitting in-the-wild, IEEE International Conference on Computer Vision, 2013.
[4] T.F. Cootes, C.J. Taylor, D.H. Cooper, J. Graham, Active shape models—their training and application, Comput. Vis. Image Underst. 61 (1) (1995) 38–59.
[5] J.M. Saragih, S. Lucey, J.F. Cohn, Deformable model fitting by regularized landmark mean-shift, Int. J. Comput. Vis. 91 (2) (2011) 200–215.
[6] P.N. Belhumeur, D.W. Jacobs, D.J. Kriegman, N. Kumar, Localizing parts of faces using a consensus of exemplars, IEEE Conference on Computer Vision and, Pattern Recognition, 2011, pp. 545–552.
[7] X. Zhu, D. Ramanan, Face detection, pose estimation, and landmark localization in the wild, IEEE Conference on Computer Vision and, Pattern Recognition, 2012, pp. 2879–2886.
[8] X. Cao, Y. Wei, F. Wen, J. Sun, Face alignment by explicit shape regression, IEEE Conference on Computer Vision and, Pattern Recognition, 2012, pp. 2887–2894.
[9] X. Xiong, F.D. la Torre, Supervised descent method and its applications to face alignment, IEEE Conference on Computer Vision and, Pattern Recognition, 2013, pp. 532–539.
[10] A. Asthana, S. Cheng, S. Zafeiriou, M. Pantic, Robust discriminative response map fitting with constrained local models, IEEE Conference on Computer Vision and, Pattern Recognition, 2013, pp. 3444–3451.
[11] M.F. Valstar, B. Martinez, X. Binefa, M. Pantic, Facial point detection using boosted regression and graph models, IEEE Conference on Computer Vision and, Pattern Recognition, 2010, pp. 2729–2736.
[12] T.F. Cootes, M.C. Ionita, C. Lindner, P. Sauer, Robust and accurate shape model fitting using random forest regression voting, European Conference on Computer Vision, 2012, pp. 278–291.
[13] B. Martinez, M.F. Valstar, X. Binefa, M. Pantic, Local evidence aggregation for regression based facial point detection, IEEE Trans. Pattern Anal. Mach. Intell. 35 (5) (2013) 1149–1163.
[14] R. Gross, I. Matthews, J.F. Cohn, T. Kanade, S. Baker, Multi-pie, Image Vis. Comput. 28 (5) (2010) 807–813.
[15] H. Zhang, A. Berg, M. Maire, J. Malik, Svm-knn: discriminative nearest neighbor classification for visual category recognition, IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, 2006, pp. 2126–2136.
[16] L. Ladicky, P.H.S. Torr, Locally linear support vector machines, International Conference on, Machine Learning, 2011, pp. 985–992.
[17] D. Nguyen-Tuong, M. Seeger, J. Peters, Local Gaussian process regression for real time online model learning, Neural Information Processing Systems, 2008, pp. 1193–1200.
[18] R. Urtasun, T. Darrell, Sparse probabilistic regression for activity-independent human pose inference, IEEE Conference on Computer Vision and Pattern Recognition, 2008.
[19] P. Dollár, P. Welinder, P. Perona, Cascaded pose regression, IEEE Conference on Computer Vision and, Pattern Recognition, 2010, pp. 1078–1085.
[20] Y. Cheng, Mean shift, mode seeking, and clustering, IEEE Trans. Pattern Anal. Mach. Intell. 17 (8) (1995) 790–799.
[21] M.A. Carreira-Perpinan, Gaussian mean-shift is an EM algorithm, IEEE Trans. Pattern Anal. Mach. Intell. 29 (5) (2007) 767–776.
[22] C.E. Rasmussen, C. Williams, Gaussian Processes for Machine Learning, The MIT Press, 2006.
[23] E. Snelson, Z. Ghahramani, Sparse Gaussian processes using pseudo-inputs, Neural Information Processing Systems, 2005, pp. 1257–1264.
[24] C. Shan, S. Gong, P.W. McOwan, Facial expression recognition based on local binary patterns: a comprehensive study, Image Vis. Comput. 27 (6) (2009) 803–816.
[25] T. Ahonen, A. Hadid, M. Pietikainen, Face description with local binary patterns: application to face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 28 (12) (2006) 2037–2041.
[26] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, IEEE Conference on Computer Vision and, Pattern Recognition, 2005, pp. 886–893.
[27] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution grey-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 971–987.
[28] B. Jiang, M. Valstar, B. Martinez, M. Pantic, A dynamic appearance descriptor approach to facial actions temporal modeling, IEEE Trans. Cybern. 44 (2) (2014) 161–174.
[29] V. Le, J. Brandt, Z. Lin, L.D. Bourdev, T.S. Huang, Interactive facial feature localization, European Conference on Computer Vision, 2012, pp. 679–692.
[30] X. Zhu, C. Vondrick, D. Ramanan, C.C. Fowlkes, Do we need more training data or better models for object detection? British Machine Vision Conference, 2012.