Facial Point Detection using Boosted Regression and Graph Models

Michel Valstar
Department of Computing
Imperial College London
michel.valstar@imperial.ac.uk

Brais Martinez
ICT Department
Universitat Pompeu Fabra
brais.martinez@upf.edu

Xavier Binefa
ICT Department
Universitat Pompeu Fabra
xavier.binefa@upf.edu

Maja Pantic
Imperial College London, Department of Computing
Twente University, EEMCS
m.pantic@imperial.ac.uk

Abstract

Finding fiducial facial points in any frame of a video showing rich naturalistic facial behaviour is an unsolved problem. Yet this is a crucial step for geometric-feature-based facial expression analysis, and methods that use appearance-based features extracted at fiducial facial point locations. In this paper we present a method based on a combination of Support Vector Regression and Markov Random Fields to drastically reduce the time needed to search for a point's location and increase the accuracy and robustness of the algorithm. Using Markov Random Fields allows us to constrain the search space by exploiting the constellations that facial points can form. The regressors on the other hand learn a mapping between the appearance of the area surrounding a point and the positions of these points, which makes detection of the points very fast and can make the algorithm robust to variations of appearance due to facial expression and moderate changes in head pose. The proposed point detection algorithm was tested on 1855 images, the results of which showed we outperform current state of the art point detectors.

1. Introduction

Facial point detection is an important step in tasks such as face recognition, gaze detection, and facial expression analysis. The performance of these tasks is usually to a large degree dependent on the accuracy of the facial point detector, yet the perfect facial point detector is yet to be developed. In this paper, we propose a novel method that brings us a step closer to this goal.

Many existing works consider the objects to detect to be entire facial features, such as an eye, the nose, or the mouth

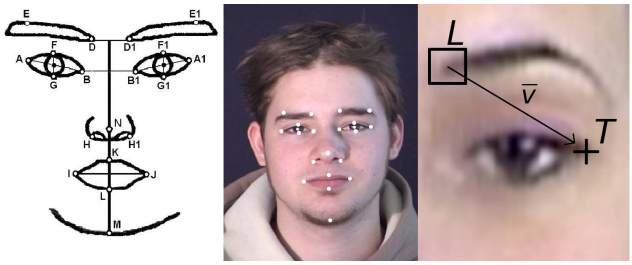

Figure 1. Point model of 22 fiducial points. The right image shows the relationship between a patch drawn at location L and the target location T.

[16]. We will denote those detectors as facial component detectors. However, the cues for tasks like facial expression recognition or gaze detection lie in the more detailed positions of points within these facial components. For example, a smile can be detected by analysing the positions of the mouth corners, not by the position of the mouth itself.

In this paper we present a novel point detector which we apply to detect 22 fiducial facial points in order to obtain an experimental performance comparison of the method. The points we aim to detect are shown in figure 1. They include 20 fiducial locations which provide useful information for automatic expression recognition, such as the upper eyelid, the eye corners, the mouth corners and the nostrils. We will denote such locations as facial points. Besides the facial points we also detect the pupils, so that in addition to facial expression analysis the gaze direction can be estimated.

Previous methods for facial feature point detection can be classified into two categories: texture-based and shape-based methods. Texture-based methods model the local texture around a given feature point, for example the pixel values in a small region around a mouth corner. Shape-based methods regard all facial feature points as a shape, which

978-1-4244-6985-7/10/$26.00 ©2010 IEEE

is learned from a set of labelled faces, and try to find the proper shape for any unknown face.

Typical shape-based methods include detectors based on active shape or active appearance models [10, 2]. These methods detect shapes of facial features instead of separate facial points. A number of approaches that combine texture and shape-based methods have been proposed as well, for example [3], which use PCA on the grey level images combined with Active Shape Models (ASM), and [14] that extends the ASM with Constrained Local Model. Chen *et al*. proposed a method that applies a boosting algorithm to determine facial feature point candidates for each pixel in an input image and then uses a shape model as a filter to select the most probable position of five feature points [1]. Of the works described above, [3, 14] have been evaluated on the same publicly available database: the BioID database [11]. This allows us to compare our work with the shape based approaches mentioned above.

Typical texture-based methods include a grey-value, eye-configuration and Artificial Neural-Network-based method that detects 8 facial points around the eyes [19], a log-Gabor filter based facial point detection [9] method to detect 7 facial points, and a two-stage method for detecting 8 facial points that uses a hierarchy of Gabor filter networks [7]. Vukadinovic and Pantic [23] presented a work that aims to detect 20 facial points. It uses Gabor filters to extract features from heuristically determined regions of interest. A GentleBoost classifier is learned on these features. During testing, a sliding window is applied to every location in this region, and the point with the highest response to the classifier is selected as the detected point. An implementation of [23] is publicly available from Dr. Pantic's website. This allows us to compare it with the method proposed in this work.

Many of the methods described above apply a sliding-windows-based search in a region of interest (ROI) of the face. A classic example of this is [23]. In this approach, a binary classifier or some other function of goodness that determines how well a location represents the target facial point is applied to every location in the ROI. However, this is a slow process, as the search time increases linearly with the search area. Depending on the type of classifier used, this approach may also lead to either multiple points classified as the target point, or to an incorrect maximum. Proposals to use gradient descent techniques to speed up this process have reportedly failed [13], as the learned functions tend to have local extremes, which can result in incorrect detections. Recently, a method was proposed to tune the classifiers in such a way that the output is a smoother function, without local extremes [15]. However, the authors reported that their method was not entirely successful in eliminating all local extremes. Another method to speed up the search was proposed by Lampert et al. [12]. In their work they proposed a branch-and-bound scheme that finds a global optimal solution over all possible sub-images.

Recently, there have been a number of approaches that use local image information and regression based techniques to locate facial points. Classifiers can only predict whether the tested location is the target location or not. Regressors on the other hand can provide much more detailed information.

By using regression we can eliminate the need for an exhaustive sliding window based search, as every patch close enough to the target point can provide an estimate of the target's location relative to that patch. Zhang et al.[24] use regression to address deformable shape segmentation. They applied an image-based regression algorithm that uses boosting methods to find a number of contours in the face. Based on these contours, they could also compute the locations of 20 facial points. Cristinacce and Cootes [4] use GentleBoost regression within the Active Shape Model (ASM) search framework to detect 20 facial points. Seise et al. [20] use the ASM framework together with a Relevance Vector Machine regressor to track the contours of lips. However, their approach was tested on only a single image sequence. Also, Relevance Vector Machines are notoriously slow and hard to train.

In summary, although some of these detectors have been reported to perform quite well when localising a small number of facial feature points such as the corners of the eyes and the mouth, there are three major issues with all existing previous work. First of all, none but [23] is able to detect all 20 facial feature points necessary for automatic expression recognition (see Fig. 1). To wit, none are able to detect the upper and lower eyelids. This is despite the fact that the upper and lower eyelid are instrumental in detecting four frequently occurring facial expressions: eye blinks, winks, widening of the eye aperture (e.g. in an expression of surprise) and narrowing of the eye aperture (e.g. in sleepy or angry expressions). Also, no previous work has reported to be able to robustly handle large occlusions such as glasses, beards, and hair that covers part of the eyebrows and eyes. Lastly, non have reported to detect facial points robustly in the presence of facial expressions. We will show that the approach proposed in this paper overcomes all three shortcomings, while retaining high accuracy and low computational complexity.

We propose a novel method based on Boosted Regression coupled with Markov Networks, which we coin BoRMaN. BoRMaN iteratively uses Support Vector Regression and local appearance based features to provide an initial prediction of 22 points, and then applies the Markov Network to ensure we sample new locations to apply the regressor to from correct point constellations. Our method thus exploits the property that objects which have a regular structural composition are made up of a combination of dis-

Figure 2. Some typical results on the FERET and BioID databases.

tinct parts whose relative positions can be described mathematically. The face, with the eyes, mouth, eyebrows etc. as parts, is a good example of this type of object.

Our approach is cast in a probabilistic framework. To determine the location of a point, we use three independent sources of information: the first is an *a priori* probability of a point's location based on the location of the detected face. Secondly we use the regression predictors, and thirdly we use Markov Random Fields (MRFs) to model the points' relative positions. Our method has lower computational complexity than existing point detectors, and is robust to facial expressions and a certain degree of head pose variations. The BoRMaN point detector will be made publicly available for download from the authors' websites.

The main contribution of the work presented here is the combination of SVRs for local search with MRFs for global shape constraints. We believe that this is a novel approach to face point localisation. In addition, to the best of our knowledge, this is the first time that feature selection by Boosting is applied to Support Vector Regression. Regarding the MRFs, we note three methodological novelties:

Firstly, a node is defined to be a spatial relation between two facial points rather than being a facial point itself. This allows a representation that is invariant to in-plane rotations, scale changes and translations (see below). It also produces a more compact set of training examples, since now only the anthropomorphic differences between subjects are encoded.

Secondly, our method proposes a novel way of defining the relations between nodes. For example, modelling the vector of two angles is difficult, since both values can be affected by in-plane rotations. By modelling the difference between two angles, and the ratio of two vector lengths, we achieve the desired invariance to in-plane rotations, isotropic scaling and translations.

Thirdly, using Gaussian Mixture Models (GMMs) to model the relations produces a bias in the final estimate towards the mean values. Yet, most of the state of the art methods use GMMs for setting spatial relations. Instead, we define a new metric which only penalises improbable configurations.

The remainder of this paper is structured as follows: In section 2 we explain the BoRMaN method we use to detect facial points. In section 3 we present an evaluation study performed on three different databases, 1500 images of frontal faces in total. Finally, in section 4 we present our closing remarks.

2. BoRMaN point detection

2.1. A priori probability

To make sure we start testing our regressors close to the target location, we need some prior information about the locations of the points. This is particularly important because we cannot test the regressor on just any image position, and still expect a reasonable result. The better the prior is, the more likely it is to obtain a good regressor estimate. In our approach we base our *a priori* probability on the bounding box returned by a face detector (the face box).

Because of its proven success, we apply a modified Viola & Jones face detection method [6] to grey-scale versions of the input images. Some postprocessing is afterwards applied to the detected face: it is enlarged by 40 % at the bottom so that every chin of our training set was included, it is resized to a 200 x 280 pixels face box, and a global illumination normalisation is applied so the worst effects of varying illumination conditions are removed. We will denote the normalised grey-scale image as F.

We divide our points into two groups: stable fiducial points and unstable fiducial points. The difference between these points is that stable points do not change their position due to facial expression or speech. In our case the set of stable points is $S_s = \{p_A, p_{A1}, p_B, p_{B1}, p_H, p_{H1}, p_N\}$ (see fig. 1). These points are detected first, as they are auxiliary for the detection of the unstable points.

After the face box has been found, we can model the prior probability of the x- and y-position of each facial point relative to the coordinate system of the detected face. Using the correct target locations T for all points in each image (obtained from manual annotation), we can map their positions to this new coordinate system based on the face box. This results in a set of points T_{fb}, for which we calculate the mean and standard deviation of their x- and

y-coordinates. We thus have a bivariate Gaussian prior probability P_i^s of the location of a facial point i, where $i \in \{p_A, p_{A1}, p_B, p_{B1}, p_H, p_{H1}, p_N\}$, relative to the coordinate system of a detected face box. This model automatically takes into account the error made by the face detector.

After detection of the stable points it is possible to use them to perform a face registration by applying a non-reflective similarity image transformation on the image F, resulting in an image that is registered to remove in-plane head rotation and, to a large effect, individual face shape differences. We denote the resulting registered face by F_r. The *a priori* probabilities of the locations of the unstable points are modelled in the same way as the stable point locations, but relative to the registered face coordinate system. We thus also have a bivariate Gaussian prior probability P_j^u of the location of each unstable facial point j, where $j \in \{p_{eyeR}, p_{eyeL}, p_D, p_{D1}, p_E, p_{E1}, p_F, p_{F1}, p_G, p_{G1}, p_I, p_J, p_K, p_L, p_M\}$.

2.2. Regression Prediction

We formulate our localisation problem as finding the vector v that relates a patch location L, selected according to some probability distribution function, to the target point T (see Fig. 1). We decompose this problem into two separate regression problems. Regressor R_α is tasked with finding the angle α of v and the regressor R_ρ is to predict the length ρ of the vector, i.e. the distance of L to T. We will denote the estimate of v provided by the regressors R_α and R_ρ by $\hat{v_L}$. This gives us the predicted target location $\hat{T} = L + \hat{v_L}$.

As regressor we have chosen Support Vector Regressors (SVRs). The reason for this is the capability of dealing with nonlinear problems, and a reportedly high generalisation capability. An early pilot study ruled out using multi ridge regression for this problem. The SVRs use a Gaussian RBF kernel. We thus need to optimise for the regression sensitivity ϵ, the kernel parameter γ and the slack variable C. Parameter optimisation is performed in a separate cross-validation loop during training, i.e. independently from the test data.

Fig. 3 shows the output of R_α and R_ρ for detection of a pupil. The regressor in this example is applied on patches located at every second pixel in every second row in an area three times the standard deviation of the prior location of the pupil. As we can see, the regressors give a good yet not a perfect indication of where the target point is. Note that although the location of the pupil is a global minimum, the predicted distance at that location is not zero.

The error of the estimates provided by the regressor can be grouped into two types. Most of the estimates contain errors that result from imprecisions in the regressor output. Such errors can be removed by using an iterative procedure, where the point is detected in several iterations. The final

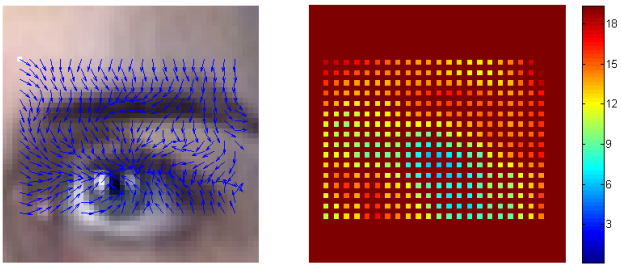

Figure 3. The output of the SVRs to detect an pupil: the estimated direction of the target (left panel) and the estimated distance to the target (right panel). The distance to the target is shown in pixels.

prediction is derived from a combination of the estimates made (see section 2.4). On the other hand, some estimates have greater errors which are not merely imprecisions. To prevent these errors from influencing the iterative process we apply spatial restrictions on the location of each facial point depending on the other facial points. This process prevents unfeasible facial point configurations. It is realised by modelling a Markov Random Field (MRF), as outlined in section 2.3. An outline of the whole algorithm is given in section 2.4.

2.3. Spatial Relations

The introduction of spatial relations between facial point positions refers to the consideration of anthropomorphical restrictions when performing facial point detection. The objective for introducing spatial restrictions is the improvement of the target position estimates by preventing unfeasible facial point combinations. The importance of such information is grounded in the richness of the problem of facial point detection: the face contains both stable and unstable fiducial points, where the latter have greatly varying positions relative to the former. Also, some points are more distinctive than others, e.g. the inference of the position of the eye corner given local image intensities is more reliable than the same task for the case of the chin position. It is therefore natural to consider the influence between facial points and derive intelligent relations, where the most reliable and stable points aid the detection of the more complicated ones.

When it comes to modelling the spatial relations, some works opt to directly model the positions of each facial point with respect to the positions of other points (e.g. [21]), using for example a coordinate system based on the head position. Instead, we propose a method were the relations between relative positions of *pairs* of points are modelled. More precisely, each relative position of a pair of points $\{i, j\}$ is a vector $r_{i,j}$ pointing from one facial point to another. The relation between two of these vectors is described by two parameters: the relation between their angles

R_α and the relation between their lengths R_ρ. Thus, if we note $r_{i,j} = (\alpha_{i,j}, \rho_{i,j})$, the objective is to model the possible relations between variables $\alpha_{i,j}$ and $\alpha_{k,l}$, and between variables $\rho_{i,j}$ and $\rho_{k,l}$. Furthermore, the obtained model should be able to deal with in-plane face rotations and imprecisions of the face detector, which affects the scale of the face box. Thus we model $R_\alpha = \alpha_{i,j} - \alpha_{k,l}$, and $R_\rho = \rho_{i,j}/\rho_{k,l}$, which obtains such an independence.

Another important difference with respect to other methods is that we model these variables with a Sigmoid function. If a variable takes its values in $[m^-, m^+]$, then $S(x) = P_{sigm}(min(v-m^-, -v+m^+))$. With this model the probability drops very fast when the value is out of the segment of possible values. Note that the value in the extremes is $S(m^-) = S(m^+) = 0.5$, which is the Sigmoid point of inflexion. An advantage of using a Sigmoid instead of a Gaussian for modelling the possible values is that a Gaussian penalises all the values but its mean, biasing the results. In contrast, modelling with a Sigmoid only penalise highly improbable constellations.

For example, in practice this model of spatial relations encodes that the line connecting points p_A and p_B is approximately orthogonal to the line connecting points p_F and p_G, or that the distance between points p_A and p_B and the distance between points p_{A1} and p_{B1} have a certain probable pre-specified length relation (See Fig 4). So although the positions of points p_F and p_G are flexible, the vector connecting them is constrained to be roughly perpendicular to the vector connecting p_A and p_B. As long as there are no out-of-plane head rotations, the lengths of vectors $p_A - p_G$ and $p_G - p_B$ are the same. We have thus obtained invariant relations from variable point positions. It is also important to note that the effectiveness and accuracy of directly modelling the point positions, P_i^s and P_i^u, depends on the accuracy of the face detector, while modelling the relative positions is independent of the face detection.

Once the pairwise relations are defined, we model the joint probability of a configuration using a Markov Random Field. In our model, the nodes correspond to each of the relative positions $r_{i,j}$ and their states are binary, coding whether the estimates are erroneous or correct. In each relation, the relative positions of points i and j, $r_{i,j} = (\alpha_{i,j}, \rho_{i,j})$, and the relative positions of points k and l $r_{k,l} = (\alpha_{k,l}, \rho_{k,l})$ are modelled as $S_{ang}(\alpha_{i,j}, \alpha_{k,l}) \cdot S_{dist}(\rho_{i,j}, \rho_{k,l})$. An example of what a node is and how the relation between two nodes is modelled is shown in fig. 4. Considering all possible relations (a fully connected net) is unfeasible for the general case due to the exponential number of relations. Some works, as [8], propose automatic ways of selecting the most informative relations and reduce the number of edges. In our case, we construct the MRF relations following a hierarchy: first the stable points are detected using a fully connected network. Afterwards, a

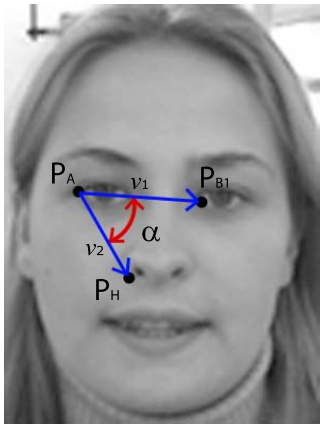

Figure 4. Vectors v_1 and v_2 are nodes $r_{p_A,p_{B1}}$ and r_{p_A,p_H}. The MRF models the relation between these two nodes: the difference between the angles of the two vectors, α, and the ratio between the lengths of the two vectors

"synthetic" facial point is created for the right eye, left eye and nose, using the mean of the stable points belonging to each of this facial component. Those points are then considered as fixed. The net generated for the left eyebrow is created using the 3 synthetic points and the two unstable points of the eyebrows. Equivalently, this process is performed to detect the unstable points for the right eyebrow, both eyes, the mouth and the chin.

Different algorithms can be used for minimising the Markov Network. We use a Belief Propagation algorithm, obtaining a probability of each point being a correct estimate.

2.4. Point detection algorithm

The BoRMaN algorithm iteratively improves its detection results. It is outlined in algorithm 2.1. The algorithm starts of with the locations of maximum prior probability as the predicted targets, as this is our best guess of the point locations, given the face detection results. We use the locations of maximum prior probability as the first locations to generate the Haar-like features from (see section 2.5), which are then used by the regressors to make the first prediction about the target locations.

We start with an empty set of predicted target locations. After each round, the predicted target locations provided by the regressors are added to a set of predictions for each point. We update the target locations as the median of this set of predictions. This updated target is then analysed by the Markov Network, which generates the patch locations to test the regressors on in the next round. To avoid repetitive results, we add a small amount of zero-mean Gaussian noise to the patch locations suggested by the Markov Nets. We repeat this for a fixed number of rounds n_r, and return

the last updated target as the final prediction of the target locations. Keeping n_r fixed allows us to guarantee a result within a fixed period of time.

Algorithm 2.1: BORMAN($priors$)

$targets \leftarrow priors$
$patches \leftarrow priors$
$predictions \leftarrow \emptyset$
for $rnd \leftarrow 1$ **to** max_rnds
do $\begin{cases} reg = regressor(patches) \\ predictions \leftarrow predictions \cup max(priors * reg) \\ targets \leftarrow median(predictions) \\ patches \leftarrow MarkovNet(targets) + N(0, \sigma) \end{cases}$

2.5. Local appearance based features and AdaBoost feature selection

For this work, we have chosen to adopt Haar-like filters as the descriptors of local appearance. The reason for this is a twofold: on the one hand, we want to show that the success of our approach is due to the idea of turning the point detection problem from a classification procedure into a regression procedure, and not due to some highly descriptive appearance feature.

On the other hand, one of our main aims of the proposed approach is to greatly improve the time required to detect all points. By computing the integral image of our input face image first, computation of each Haar-like filter is reduced to as little as four addition/subtraction operations.

The optimal patch size has empirically been determined to be 32 pixels high and wide during a pilot study. For every location in the face image F from where we want to get a prediction of the direction and distance to the target point, we compute the responses to 9 different Haar-like filters, at four different scales: the full 32 pixels, 16, 8, and 4 pixels big. All filters are square, and for the 16, 8 and 4 pixels filters, the centres of the filters were placed to overlap half of the width of the adjacent filters of the same scale. This results in 2556 dimensional feature vectors.

Although SVR regressors are able to learn a function even with very little training data, regression performance decreases when the dimensionality of the training set is too large. To be more precise, if we have a training set D with n_f features and n_s instances, then if $n_f > n_s$, it is possible to uniquely describe every example in the training set by a specific set of feature values. Our training set consists of some 400 examples (images). Considering the fact that the dimensionality of our feature set is 2556, we are indeed in danger of over-fitting to the training set. One way to overcome this problem is to reduce the number of features used to train the SVR using feature selection algorithms.

Boosting algorithms such as GentleBoost or AdaBoost are not only fast classifiers, they are also excellent feature selection techniques, as reported in [22]. As an added benefit of employing feature selection, we will have to compute fewer features at each patch location, thus speeding up our algorithm. This is in contrast with feature reduction techniques such as PCA, which are not strictly feature *selection* techniques and still require all features to be computed first.

We implemented Drucker's approach to AdaBoost regression [5], using multi-ridge regression as the weak regressors. To find the optimal number of features to select, a stop condition is usually defined based on the strong regressor output. For example, selection of features could terminate if the strong regressor output stops increasing for a predefined number of rounds. However, preliminary tests have shown that this does not produce the optimal number of selected features. Therefore, we do not use this stop criterium and instead let the AdaBoost process order all features based on their relative importance. We then optimise the number of features to use in a separate cross-validation process using SVRs.

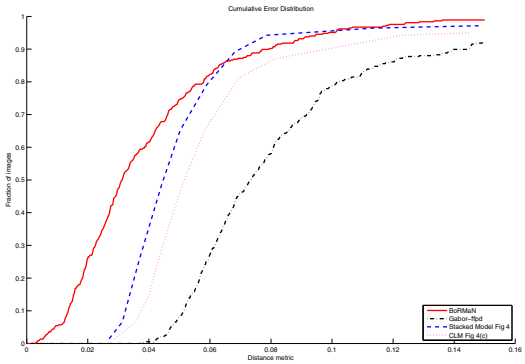

Figure 5. Comparison of the cumulative error distribution of point to point error measured on the BioID test set.

3. Experiments

We have evaluated our method in two ways: a cross-validation test on 400 images taken from the FERET and MMI-Facial Expression databases [18, 17], and a database independent test on the BioID database [11]. The first test determines how well the database copes with varying expressions and occlusions. The second test performs a benchmark comparison of our proposed method with the existing state of the art. Typical results are shown in Fig 2.

The images selected from the FERET and MMI-Facial Expression databases contain varying facial expressions, many occlusions of the mouth area by beards and moustaches, of the eyebrow area by hair, and of the eye areas by glasses. There often were significant reflections on the glasses, which made the detection of the eyes a particu-

Table 1. BoRMaN point detection results for the cross-validation test on 400 images. The classification rate C is defined as the number of times $e < 0.1$, and the mean and standard deviation of the error (e_μ ,e_σ) are measured in percentages of d_{IOD}.

Point	C	$\mathbf{e}_\mu$	$\mathbf{e}_\sigma$	Point	Cl.Rate	$\mathbf{e}_\mu$	$\mathbf{e}_\sigma$
p_A	92.25%	4.44%	4.46%	p_{G1}	96%	3.40%	4.14%
p_{A1}	90.5%	5.25%	5.86%	p_H	93.5%	3.71%	3.46%
p_B	84.5%	5.43%	5.67%	p_{H1}	93.25%	4.00%	3.48%
p_{B1}	92.25%	4.27%	4.24%	p_I	93.5%	4.40%	4.06%
p_D	90.25%	5.20%	4.73	p_J	92.5%	4.87%	5.65%
p_{D1}	91.25%	4.97%	5.02%	p_K	95%	3.94%	4.08%
p_E	89%	5.40%	4.77%	p_L	89.5%	5.23%	5.26%
p_{E1}	81%	7.10%	7.81%	p_M	19.25%	20.5%	12.0%
p_F	94.5%	3.34%	3.96%	p_N	96.25%	3.63%	3.15%
p_{F1}	94.25%	3.62%	5.15%	right pupil	94.75%	3.16%	4.06%
p_G	95%	3.41%	3.90%	left pupil	94.75%	3.21%	4.81%

larly challenging problem. On this set we applied a 10-fold cross-validation evaluation. The results of this study are shown in table 1. The table shows the mean error per point in percentages of d_{IOD} (column 2), the standard deviation of the error per point in percentages of d_{IOD} (column 3), and the classification rate per point (column 1). The detection error of a point i is defined as the Euclidian point to point distance between T_i and $\hat{T}_i$:

$$e_i = \frac{||T_i - \hat{T}_i||}{d_{IOD}} \tag{1}$$

where d_{IOD} is defined as the Inter-Ocular Distance, i.e. the distance between the pupils. The classification rate C_i is defined as:

$$C_i = \frac{\sum_{j=1}^{n} e_i^j < 0.1}{n} \tag{2}$$

where j is an image number and n the total number of images in the dataset. As we can see, all points but point p_M are detected with extremely high accuracy, even though the database includes many occlusions and expression. Point p_M has a low detection results for two reasons: Firstly, the point's appearance is not well defined. The chin is locally smooth, and we can only identify it easily if a subject has a sharp jawline. Even then, we're dependent on good lighting to make the jawline visible. Secondly, human annotators find it very difficult to consistently annotate the location of the chin. This causes a big variance in the appearance of the chin in the training data, which, in turn, causes detection of the chin to be more difficult.

The goal of our second test was to compare our facial point detector with those of others. Namely, we want to compare our point with the current state of the art: two Active Shape Model methods ([3, 14], which we denote as CLM and Stacked Model, respectively), and a Gabor-feature/GentleBoost based method that employs sliding window based search [23] (which we will call Gabor-ffpd). To make such a comparison, we are forced to use the BioID database, as neither the CLM or the Stacked Model implementations are publicly available, yet they both tested their methods on the BioID dataset. There *is* a publicly available implementation of the Gabor-ffpd. Thus, if we apply both the BoRMaN method and the Gabor-ffpd method on the BioID dataset, we can compare the performance of the various methods on a common dataset. The BoRMaN method was trained using the FERET and MMI-database training data of the first fold of the previously outlined cross-validation study.

The results of this are shown in Fig 5. The figure shows the cumulative error distribution of the m_{e17} error measure. The measure m_{e17} is defined in [3] as the mean error over all internal points, that is, all points that lie on facial features instead of the edge of the face. For our method, that would mean all points except for p_M. However, neither the CLM nor the Stacked Model approaches are able to detect the eyelids. So, to allow a fair comparison, we have excluded the points $\{p_F, p_{F1}, p_G, p_{G1}\}$ as well when calculating m_{e17}. Fig 5 shows clearly how we outperform all three other approaches. The difference between the error levels for which 50% of the images are correctly detected is twice as big when comparing BoRMaN with Stacked models than when comparing Stacked Models with CLMs. The figure also shows that a significant proportion BoRMaN predictions have an extremely low error: 26% of the images have an average point error of less than 2% of d_{IOD}, which translates to roughly 2 pixels per point. Because the BoRMaN method was trained using only images from completely different databases, we have also shown that our system generalises to unseen images from other databases.

4. Conclusions and future work

We have proposed a novel method for finding 20 fiducial points and the pupils in an input image of a frontal face, based on boosted Support Vector Regression, Markov

Random Fields and dense local appearance based features. The proposed method, which we coined BoRMaN, is robust to varying lighting conditions, facial expression, moderate variations in head pose, and occlusions of the face caused by glasses or facial hair. Our method is also more accurate than the current state of the art in facial point detection [3, 14, 23]. It is approximately twice as fast as [23].

5. Acknowledgments

This work has been funded in part by the European Communitys 7th Framework Programme [FP7/20072013] under the grant agreement no 231287 (SSPNet). The work of Michel Valstar is further funded in part by the European Community's 7th Framework Programme [FP7/2007-2013] under grant agreement no 211486 (SEMAINE). The work of Maja Pantic is also funded in part by the European Research Council under the ERC Starting Grant agreement no. ERC-2007-StG-203143 (MAHNOB). The work of Brais Martinez and Xavier Binefa was funded by the Spanish MITC under the "Avanza" Project *Ontomedia* (TSI-020501-2008-131)

References

[1] L. Chen, L. Zhang, H. Zhang, and M. Abdel-Mottaleb. 3d shape constraint for facial feature localization using probabilistic-like output. *Proc. IEEE Int'l conf. on Automatic Face and Gesture Recognition*, pages 302–307, 2004.

[2] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *IEEE European Conf. Computer Vision*, 2:484–498, Sep 1998.

[3] D. Cristinacce and T. Cootes. Feature detection and tracking with constrained local models. *Proc. British Machine Vision Conference*, pages 929–938, 2006.

[4] D. Cristinacce and T. Cootes. Boosted regression active shape models. *Proc. British Machine Vision Conference*, pages 880–889, 2007.

[5] H. Drucker. Improving regressors using boosting techniques. *Int'l workshop on Machine Learning*, pages 107–115, 1997.

[6] I. Fasel, B. Fortenberry, and J. Movellan. A generative framework for real time object detection and classification. *Computer Vision and Image Understanding*, 98(1):182–210, 2005.

[7] R. Feris, J. Gemmell, K. Toyama, and V. Kruger. Hierarchical wavelet networks for facial feature localization. *Proc. IEEE Int'l conf. on Automatic Face and Gesture Recognition*, pages 118–123, 2002.

[8] L. Gu, E. Xing, and T. Kanade. Learning gmrf structures for spatial priors. *IEEE Conf. Computer Vision and Pattern Recognition*, pages 1–6, 2007.

[9] E. Holden and R. Owens. Automatic facial point detection. *Proc. Asian Conf. Computer Vision*, 2:731–736, 2002.

[10] C. Hu, R. Feris, and M. Turk. Real-time view-based face alignment using active wavelet networks. *IEEE Int'l Workshop on Analysis and Modeling of Faces and Gestures*, pages 215–221, 2003.

[11] O. Jesorsky, K. Kirchberg, and R. Frischholz. Robust face detection using the hausdorff distance. *Lecture Notes in Computer Science*, pages 90–95, 2001.

[12] C. Lampert, M. Blaschko, and T. Hofmann. Beyond sliding windows: Object localization by efficient subwindow search. *IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[13] S. Lucey and I. Matthews. Face refinement through a gradient descent alignment approach. *Proc. of the HCSNet workshop on Use of vision in human-computer interaction*, pages 43–49, 2006.

[14] S. Milborrow and F. Nicolls. Locating facial features with an extended active shape model. *Proc. IEEE European Conference on Computer Vision*, pages 504–513, 2008.

[15] M. Nguyen and F. D. la Torre. Learning image alignment without local minima for face detection and tracking. *Proc. IEEE Int'l conf. on Automatic Face and Gesture Recognition*, pages 1–7, 2008.

[16] M. Nguyen, J. Perez, and F. D. la Torre. Facial feature detection with optimal pixel reduction svm. *Proc. IEEE Int'l conf. on Automatic Face and Gesture Recognition*, pages 1–6, 2008. Presented at FG08. CMU paper.

[17] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *IEEE International Conference on Multimedia and Expo*, pages 317–321, 2005.

[18] P. Phillips, H. Wechsler, J. Huang, and P. Rauss. The feret database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing*, 16(5):295–306, 1998.

[19] M. Reinders, R. Koch, and J. Gerbrands. Locating facial features in image sequences using neural networks. *Proc. IEEE Int'l conf. on Automatic Face and Gesture Recognition*, pages 230–235, 1996.

[20] M. Seise, S. McKenna, I. Ricketts, and C. Wigderowitz. Learning active shape models for bifurcating contours. *IEEE Trans. Medical Imaging*, 26(5):666–677, 2007.

[21] E. Sudderth, A. Ihler, and W. Freeman. Nonparametric belief propagation. *Proc. Conf. on Computer Vision and Pattern Recognition*, Jan 2003.

[22] M. Valstar and M. Pantic. Combined support vector machines and hidden markov models for modeling facial action temporal dynamics. *Lecture Notes on Computer Science*, 4796:118–127, 2007.

[23] D. Vukadinovic and M. Pantic. Fully automatic facial feature point detection using gabor feature based boosted classifiers. *IEEE Int'l Conf. Systems, Man and Cybernetics*, 2:1692–1698 Vol. 2, 2005.

[24] J. Zhang, S. Zhou, D. Comaniciu, and L. McMillan. Discriminative learning for deformable shape segmentation: A comparative study. *Proc. IEEE European Conference on Computer Vision*, pages 711–724, 2008. Based on Zhou's ICCV05 paper Is.