Multi-View Constrained Local Models for Large Head Angle Facial Tracking

Georgia Rajamanoharan Timothy F. Cootes
Centre for Imaging Sciences
The University of Manchester, UK
{georgia.rajamanoharan, timothy.f.cootes}@manchester.ac.uk

Abstract

We propose Multi-View Constrained Local Models - a simple but effective technique for improving facial point detection under large head angles, such as in a car driving setting. Our approach combines a global shape model with separate sets of response maps targeted at different head angles, indexed on the shape model parameters. We explore shape-space division strategies and show that, as well as outperforming the traditional method, our approach also provides a marked speed-up which demonstrates the suitability of this technique for real-time face tracking.

1. Introduction

Accurate facial point detection is a crucial part of any facial analysis system. It allows for consistent feature extraction either through facial alignment, or targeting of areas around the feature points identified. Many methods have been developed for this task, but in recent years the majority of approaches have been based on one of three main methods: Active Appearance Models (AAMs) [8], pictorial structures [12], Constrained Local Models (CLMs) [11]. Though recently Supervised Descent Method (SDM) has been proposed [18], which employs a sequence of linear regressors to predict shape representation. Advances on these basic techniques include Bayesian AAMs [1], Histogram of Oriented Gradient (HOG) AAMs [2], the use of Structured Support Vector Machines (SSVMs) [19, 20], and parallel cascades of linear regressors [4] which allow updating of the sequence. Recent extensions of the basic CLM idea include the use of more discriminative response maps [3], Local Neural Fields (LNFs) [5] and Random Forest Regression-Voting (RFRV) techniques [9]. These approaches have demonstrated that CLMs are highly accurate and robust, and able to deal with unseen identities as well as a large number of expressions and lighting conditions.

In a number of applications, such as the tracking of driver faces in cars, the facial point detection must be accurate over a wide range of head angles, due to the regular turning of the head that occurs. However, traditional methods often fail when there is large yaw (horizontal turning) or pitch (vertical nodding) of the head. In this paper we propose a novel method to tackle the problem of facial feature point detection and tracking in large head angles: Multi-View Constrained Local Models (MV-CLMs). Our approach is simple, but gives improved performance on large head angle face images, whilst offering a significant speed-up over the traditional CLM techniques. We employ shape-space division in order to select pose-specific training sets that allow a number of models to be built, each targeted at a narrower head angle range. We then propose an efficient algorithm for switching between these models that allows for improved accuracy without the need to test all models on every image.

In summary, the contributions of the paper are as follows: we propose a novel Multi-View CLM approach that combines one shape model with response maps that are specifically targeted at different head orientation examples, and adapt the CLM search algorithm to allow for switching to the appropriate response maps. We demonstrate how this approach improves performance on large head angle test images, as well as giving a significant speed-up over the traditional approach. Finally, we investigate a number of strategies for division of the shape-space for training of our response maps and shape models, and test all strategies on appropriate datasets containing large head angles.

2. Random Forest Regression Voting CLMs

CLMs are a widely employed approach for detection of feature points, particularly in faces [11]. They exploit a combination of a statistical shape model with point-specific local response maps, in order to match points against an image. The shape model represents the position of the full set of points, $\mathbf{X}$, through a linear combination of a set of modes of variation:

$$\mathbf{X} = T(\bar{\mathbf{X}} + \mathbf{P}\mathbf{a}) \tag{1}$$

where $\bar{\mathbf{X}}$ is the mean positions of the set of points in a suitable reference frame, $\mathbf{P}$ the set of orthogonal modes of vari-

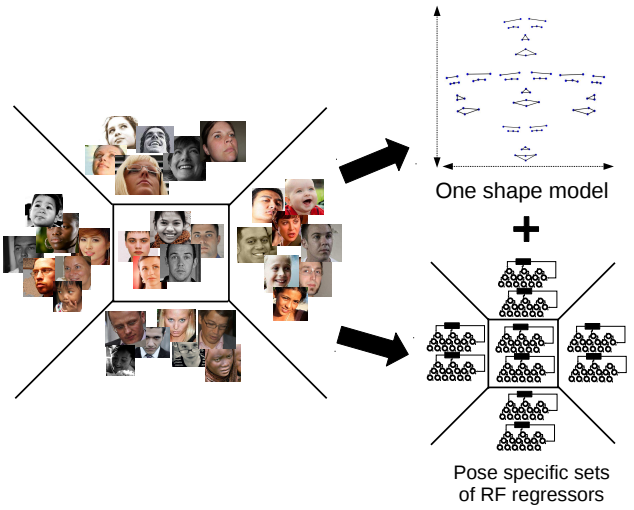

Figure 1: Overview of the Multi-View CLM approach.

ation, $\mathbf{a}$ a set of shape parameters, and T a global transformation (typically similarity).

Point response maps are generated from the approximated positions of the points which are used to provide a quality of fit, $C_j(\mathbf{x}_j)$, of each point. The full objective function for optimisation of the shape parameters and transformation parameters, $\mathbf{t}$, given image I is then:

$$f(\mathbf{a}, \mathbf{t}) = -\log p(\mathbf{a}, \mathbf{t}) + \alpha \sum_{j=1}^{N} C_j(\mathbf{x}_j) \qquad (2)$$

where α is a normalising constant, and $p(\mathbf{a}, \mathbf{t}) = p(\mathbf{a})$ as all poses are equally likely. Minimisation of this cost function will therefore lead to the best fit of all facial points. A number of functions are possible for C, such as normalised correlation, but here we are interested in the use of RFRV methods, which have been shown to lead to fast, robust and accurate performance in the application of facial point tracking [9].

Regression-Voting is a technique for building response maps via exploitation of a regressor which has been trained to predict the position of a point relative to each coordinate in a regular grid, around which features are sampled. The confidence of predictions for all possible positions are aggregated over all of the regressors in an accumulator array. The resulting map is then smoothed to allow for uncertainty in the predictions. The Random Forest Regression-Voting Constrained Local Model (RFRV-CLM) technique exploits Hough Forests which provide multiple regressors employed per point response map, with each random forest also able to used make multiple predictions. A number of different voting procedures were explored in [9], however a single unit vote per tree was found to produce the most accurate results as well as being the most cost efficient method.

3. Multi-View Constrained Local Models

RFRV-CLMs have been shown to be very robust when applied to frontal and near-frontal face images, able to deal with a wide range of expressions [9]. However, larger head-turns can cause inaccuracy in the facial point detection. At large head angles, the appearance of the patches around each point are very different from those in frontal or near frontal images. As the Regression Forests (RFs) are built on a wide range of examples, the majority of which are frontal, this can lead to inaccuracy in their predictive performance at large head angles. The aim of this work is to overcome these limitations by building facial orientation specific regressors. By training each set of regressors on examples of a limited range of facial angles only, we can build RFs that are better targeted at point detection in particular head angle images. Hence, we propose a novel MV-CLM approach that employs one global shape model along with pose-specific response maps. Fig. 1 gives an overview of our method.

3.1. Training

The training process for the MV-CLM is similar to the training of a traditional RFRV-CLM. However, the difference here is that we divide the training set according to two shape model parameters of the global shape model - the parameters that govern head pose. We employ a set of N images, $\mathrm{L} = \{I_1, ..., I_N\}$, for which the feature points of interest, $\mathrm{X} = \{\mathbf{X}_1, ..., \mathbf{X}_N\}$, have been annotated. The first stage in training is to construct a global statistical shape model. This is trained by applying Principal Component Analysis (PCA) to the aligned shapes to produce the major modes of variation [10]. This shape model is common to all the local models, and employed to select which model should be applied as well as for constraining the feature points.

Then, given a training image, after estimation of the global pose and resampling into a reference frame, the shape model is fitted to the annotated facial points and the two relevant shape parameters taken: the first corresponding to rotation about a horizontal axis, (i.e. variation in head pitch), a^v, and the second to rotation about a vertical axis (i.e. variation in head yaw), a^h. Given these parameters we construct a model index, m_i, for each training image:

$$m_i = g(a_i^v, a_i^h) \qquad (3)$$

where g defines an indexing function. This allows the training set to be divided into model specific sets according to index.

We construct a new set of training examples by taking only the images with the appropriate model index: $\mathrm{L}_k = \{I_i | I_i \in \mathrm{L}, m_i = k\}$ with annotated points $\mathrm{X}_k = \{\mathbf{X}_i | \mathbf{X}_i \in \mathrm{X}, m_i = k\}$, where $k = \{1, ..., M\}$ and M is the number of regions in the division strategy. These data subsets can then be used to train our response maps and shape

Algorithm 1 MV-CLM search algorithm.

Require: r_{min}, r_{min}, radius reduction parameter, γ, and initial shape parameters, $\mathbf{a}$
Initialise $\mathbf{t} = \mathbb{I}$, radius $r = r_{max}$, $\mathbf{X}_i = \bar{\mathbf{X}} + \mathbf{Pa}$, $m = g(a^v, a^h)$
for each model $k \in \mathrm{M}$ **do**
  Calculate response maps C_j^k for each point
end for
for each iteration **do**
  while $r \geq r_m$ **do**
    for each point $\mathbf{x}_p$ in point set $\mathbf{X}_i$ **do**
      Search for point, $\mathbf{x}'$, within disk of radius r which gives best QoF: $\mathbf{x}_p \rightarrow \arg\min_{\mathbf{x}' : \, |\mathbf{x}' - \mathbf{x}_p| < r} C_j^m(\mathbf{x}')$
    end for
    Fit shape parameters, $\mathbf{a}$, and transformation, $\mathbf{t}_{ref}$, to points.
    if $\mathbf{a}^T \mathbf{S}_\mathbf{a}^{-1} \mathbf{a} > \lambda$ **then**
      Set $\mathbf{a}$ to nearest point such that $\mathbf{a}^T \mathbf{S}_\mathbf{a}^{-1} \mathbf{a} = \lambda$
    end if
    Update model index $m \rightarrow g(a^v, a^h)$
    Update points according to new parameters: $\mathbf{X}_i \rightarrow T(\bar{\mathbf{X}} + \mathbf{Pa})$
    Reduce radius size: $r \rightarrow \gamma r$
  end while
  Update transformation parameters: $\mathbf{t} \rightarrow \mathbf{t} \circ \mathbf{t}_{ref}$
  Map points to the image frame: $\mathbf{X}_i \rightarrow T(\mathbf{X}_i)$
  If pose changes significantly, recalculate C_j^k.
end for

models for pose-specific models. We then train a number of models, $\mathrm{M} = \{\mathcal{M}_k\}$. Given the training subset for each model, $\mathcal{M}_k$, training of the point-specific response maps, C_j^k, proceeds as in the original RFRV-CLMs method [9] which we summarise here. The response map for a single feature point is trained by generating samples around the true position of this point in the reference frame. Features, $\mathbf{f}_j$, are extracted at a set of random displacements, $\mathbf{d}_j$, drawn from a flat distribution around this position. The scale and orientation of the pose is also randomly perturbed to allow for inaccuracy of the initial estimate. Randomised decision trees are then constructed from pairs of $\{\mathbf{f}_j, \mathbf{d}_j\}$, choosing the feature to split the data at each node through minimisation of mutual entropy at each stage. Haar-like features are employed, as in the original implementation, as these are efficient to calculate from integral images.

3.2. Testing

Here we detail the testing procedure for the MV-CLM. Again, the method is similar to that of a traditional RFRV-CLM search algorithm, but here the response maps utilised in each iteration are updated based on the current shape parameters. This allows at each stage for the appropriate RF model to be selected based on the current shape parameters, before predicting the most likely position of each point based on the response maps for that model.

First, we assume that we have an initial estimate of the pose and shape parameters, either from the previous frame in a sequence (during tracking), or from initialisation. We then aim to minimise Equation 2 and thus identify the optimal shape parameters. Assuming a flat distribution for the model parameters, $\mathbf{a}$, within hyper-ellipsoidal bounds, and that all poses are equally likely, this equation becomes equivalent to:

$$f(\mathbf{a}) = \sum_{j=1}^{N} C_j^m(\mathbf{x}_j) \quad subject \quad to \quad \mathbf{a}^T \mathbf{S}_\mathbf{a}^{-1} \mathbf{a} \leq \lambda \tag{4}$$

where $\mathbf{S}_\mathbf{a}$ is the parameter covariance matrix and λ is the threshold on the Mahalabanobis distance and $m = g(\mathbf{a})$.

The MV-CLM search algorithm then proceeds by resampling using the current pose, to transform the image into the reference frame, and computing the cost images for all models, C_i^k. These are then employed to search within the disks defined in Equation 4. The shape parameters are then updated by fitting the model to these points, and the model index m recalculated. If pose changes significantly then the response maps for all models are recalculated. This algorithm is is summarised in Alg. 1. When there is no prior knowledge about the shape parameters (e.g. face detection only), we initialise the shape parameters in multiple configurations, and perform the search in parallel from these starting points, choosing the initialisation that gives the best final quality of fit. This ensures that the algorithm does not get stuck in local minima, and thus fail to switch to the correct model. This is not required in tracking as the shape parameters are initialised based on the previous frame.

An added benefit of this method is that the size of the trees produced is greatly reduced from those of the traditional RFRV-CLM, due to the specificity of the RFs produced. Added to the fact that only one set of RFs is employed for testing in each iteration, our approach thus offers a marked speed-up over the original RFRV-CLM.

3.3. Shape-Space Division Strategy

We explore a number of strategies for division of the shape-space defined by the vertical and horizontal shape parameters, a^v and a^h. Fig. 2 summarises these strategies.

One Dimensional Division Here we divide training examples according to one parameter only, as shown in Fig. 2a. Taking either the vertical parameter, a^v, or horizontal parameter, a^h, we define the division by the number of regions, n, a central parameter value, c, that defines the centre of the middle region, and the width of each shape-space region, w. Then in the case of vertical division in one dimension, we can set m according to the following function:

$$g^v(a^v) = \begin{cases} 0 & if \quad a^v < c^v - \frac{w^v(n^v-2)}{2} \\ n-1 & if \quad a^v > c^v + \frac{w^v(n^v-2)}{2} \\ x & otherwise \end{cases} \tag{5}$$

where x satisfies $a^v > c - \frac{w^v(n^v-2x)}{2}$ and $a^v < c^v - \frac{w^v(n^v-2(x+1))}{2}$. $g^h(a^h)$ is defined similarly. In this paper we set $n^v = n^h = 3$, $c^v = c^h = 0$ and $w^v = w^h = 0.1$.

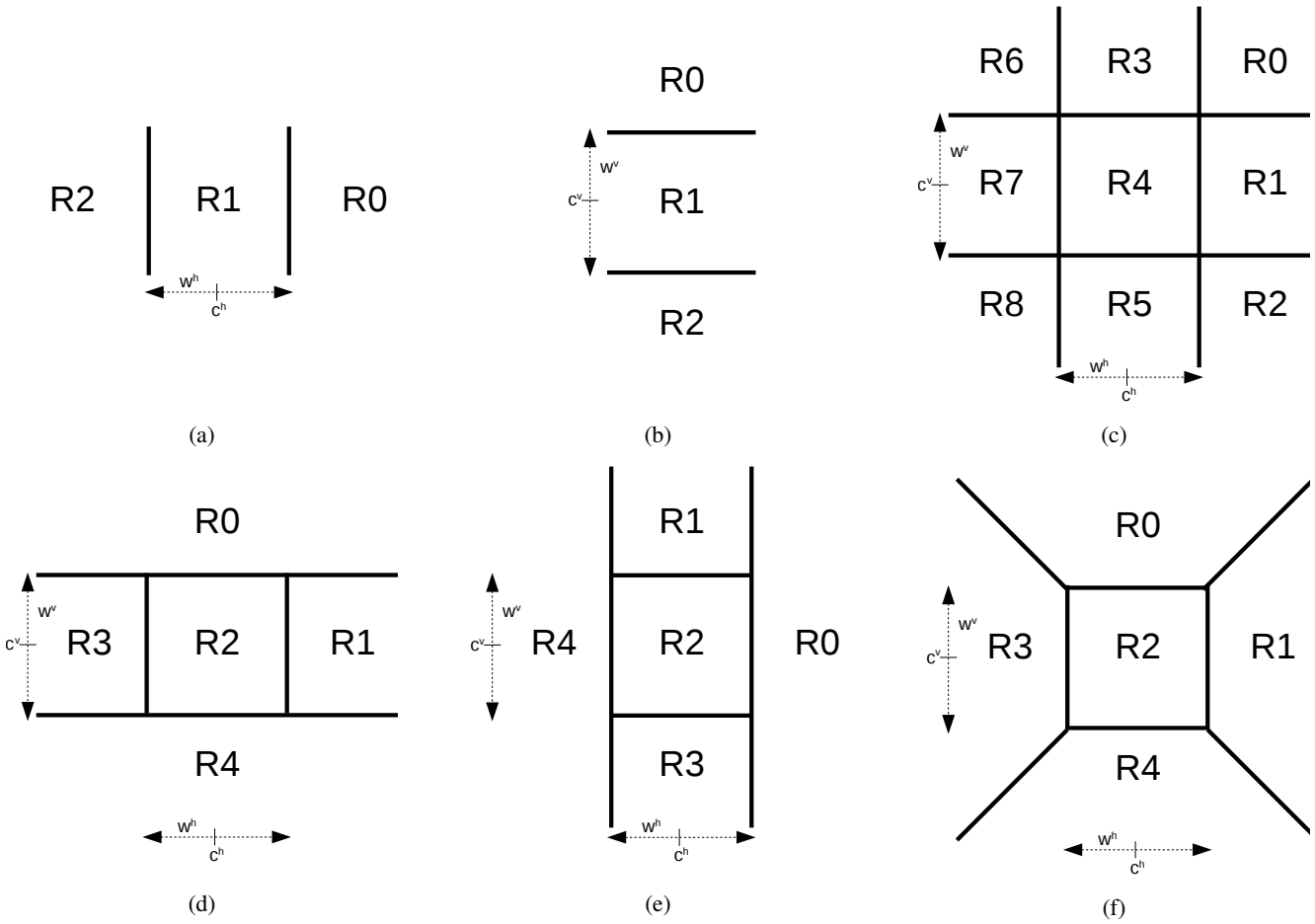

Figure 2: The different shape-space division strategies employed with $n^v = n^h = 3$. (a) 1D Horizontal (b) 1D Vertical (c) 2D Grid (d) 2D Grid-Merge Horizontal (e) 2D Grid-Merge Vertical (f) 2D Diagonal.

We now define the two dimensional strategies in terms of the one dimensional horizontal and vertical functions, $g^v(a^v)$ and $g^h(a^h)$, which we now shorten to g^v and g^h.

2D Grid Division In the two dimensional case, the most obvious strategy is to divide the space into a grid of rectangular regions, as shown in Fig. 2d. The space is split in each dimension using the 1D method, and a unique index given to each resulting region. Hence, $g(a^v, a^h)$ is defined as:

$$g(a^v, a^h) = g^v + n^v g^h \tag{6}$$

where n^v is the number of regions in the vertical dimension.

2D Grid-Merge Division However, one difficulty with the grid approach can be that there are often not enough training examples for the corner cases (e.g. positive yaw plus positive pitch) of the shape-space to allow for adequate training of RFs for these models. Hence, we want a strategy that creates bigger shape-space regions to incorporate more examples into one model, while still dividing into useful regions. One way to do this is to merge a number of the grid regions into a single region, in order to collate regions for which there are not an adequate number of examples for building accurate RFs, as shown in Fig. 2f.

We can then define $g(a^v, a^h)$ in each of these cases. In the horizontal case we take $g(a^v, a^h)$ to be defined as only dependent on the vertical index as long as it is not the central index. In this latter case it is defined as the sum of the vertical and horizontal indices:

$$g(a^v, a^h) = \begin{cases} g^v & if \quad g^v < \hat{n}^v \\ g^v + g^h & if \quad g^v = \hat{n}^v \\ g^v + n^h - 1 & otherwise \end{cases} \tag{7}$$

where $\hat{n}^v = \frac{(n^v - 1)}{2}$ is the central vertical index. The vertical case is similarly defined, except here m is dependent on the horizontal index only, apart from in the case where it is equal to the central index.

2D Diagonal Division The final strategy is to separate the frontal examples from all other head orientations. This region is defined as the same as in all other cases - a rectangular region defined by (c^v, c^h), w^v and w^h. However, the rest of the space is divided by diagonal lines to create trapezium-shaped regions that capture the central examples in all four directions (e.g. no yaw but positive pitch) but also a portion of the corner cases (e.g. positive pitch plus positive yaw), as shown in Fig. 2e. This method creates a more even division strategy which we would expect would give improved performance. Hence, we define $g(a^v, a^h)$ as:

$$g(a^v, a^h) = \begin{cases} \hat{n}^v + \hat{n}^h & if \quad g^v = \hat{n}^v, g^h = \hat{n}^h \\ g^v & if \quad x^v > x^h, g^v < \hat{n}^v \\ g^h + \hat{n}^v & if \quad x^h >= x^v \\ g^v + \hat{n}^h - 1 & otherwise \end{cases} \tag{8}$$

where $x^h = \frac{|a^h - c^h|}{w^h}$ and $x^v = \frac{|a^v - c^v|}{w^v}$ are the normalised distances of the horizontal and vertical parameters respectively from the central point.

4. Experiments

We assess the performance of our algorithm by conducting experiments on two different test sets, constructed from different databases. For training in both cases we employ a mixture of two datasets: the Annotated Facial Landmarks in the Wild (AFLW) database [14], and a set of driver videos that contain sequences of subjects in a car environment, performing a wide range of expressions and head movements. The training set is constructed from 970 images from the AFLW database, as well as 662 images from four driver video sequences. This allows for a wide range of subjects and expressions from the AFLW data, as well as a number of large head orientation training examples from the driver videos. We train a model on facial points, that cover the eyes, eyebrows, nose and mouth. Firstly, we construct a test set by taking the remaining 22 sequences from the driver videos, each of which contains 160-180 annotated images. No subject appears in both the training and test sets. Secondly, we participated in the 300-VW Challenge [16] to explore our best method when employed on 68 points and on a wide-range of data, and assess how it compares to a baseline method in three different categories of test data. Here, we trained on every 10^{th} frame of the training videos provided, which totalled 9330 images. The training images in this challenge were annotated in a semi-automatic fashion that employed two methods [7, 17].

We are particularly interested in how our approach performs on examples of large head angles, and traditional face detectors struggle to accurately initialise on data of this kind. Hence, we perform the driver video experiments by taking the ground truth eye points, and randomly translating these by up to 10 pixels in either dimension, scaling by up to 20%, and rotating by up to $10°$, before applying the facial point search algorithm. This is done five times for each image. Table 1 summarises the results of all of our approaches as compared to the traditional RFRV-CLM method. For the 300-VW challenge, a face detector is employed to initialise the model, thus this experiment demonstrates the performance of our algorithm in a real-world scenario. In both cases, we perform 5 search iterations using each model.

4.1. Discussion

Table 1 shows the overall performance of the different MV-CLM division strategies, when compared to the original RFRV-CLM method trained on the same data. The code for this baseline method was provided by the authors for this test. Rows 2-3 display the 1D methods, while rows 4-7 display the 2D methods. The final column in this table also shows the comparative computational cost, on average, of a single search run on an Intel Quadcore 3.1GHz machine. This table demonstrates how an overall increase in performance can be achieved when dividing the shape-space using three of the strategies: 1D Horizontal (1D-H), 2D Grid-Merge Vertical (2D-GMV), and 2D Diagonal (2D-D). At the same time a marked improvement in speed is achieved, particularly in the case of 2D Diagonal strategy in which a speed-up of more than 25% is observed.

As the majority of our test images are frontal, we can further investigate the benefits of the optimal 1D and 2D methods by looking at performance for different ranges of head angles. We divide the test set automatically, but fitting the global shape model to the ground truth points and then taking the first two shape model parameters. Images with $a^t > 0.1$ or $a^t < -0.1$, where $t = \{h, v\}$, are deemed to be non-frontal (labelled as +ve/-ve yaw/pitch). Otherwise they are deemed to be frontal (labelled as 0 yaw/pitch). The resulting cumulative distribution function (CDF) of average point errors for these different image types are shown in Fig. 3. The graphs demonstrate how the improvement is concentrated into the large positive and negative yaw examples (subject looking left and right) and the large positive pitch (subject looking down). The performance on frontal images and negative pitch (subject looking up) are comparative to the RFRV-CLM method in both cases. These results highlight the significant improvement achieved by the MV-CLM approach on large head angles.

Finally, the 300-VW results in Fig. 4 provide further evaluation of the optimal approach found in the first experiment, the 2D-D MV-CLM. It is clear from the top row, showing results for the 49 facial points, that our method is far superior to the baseline method [4]. The baseline here was trained on the LFPW [6], Helen [15] and MultiPie Databases [13], rather than the training set for the challenge, so, though it is not possible to say for certain that our method significantly outperforms this method, the

Division Type	Method	Med	Mean	Std	Maximum Error			Time (ms)
					50%	80%	90%	
No Division	RFRV-CLM [9]	5.73	7.54	6.82	5.73	9.77	12.50	195
1D Division	1D-V MV-CLM	5.85	7.88	7.34	5.85	10.13	13.29	***158***
	1D-H MV-CLM	**5.49**	**7.25**	**6.85**	**5.49**	**9.17**	***11.57***	179
2D Division	2D-G MV-CLM	5.79	7.84	7.54	5.79	10.00	12.83	172
	2D-GMV MV-CLM	5.72	7.50	7.20	5.72	9.49	11.95	161
	2D-GMH MV-CLM	5.71	7.67	7.15	5.71	9.81	12.89	169
	2D-D MV-CLM	***5.68***	***7.42***	***7.27***	***5.68***	***9.23***	**11.50**	**143**

Table 1: Driving Video Statistical Results

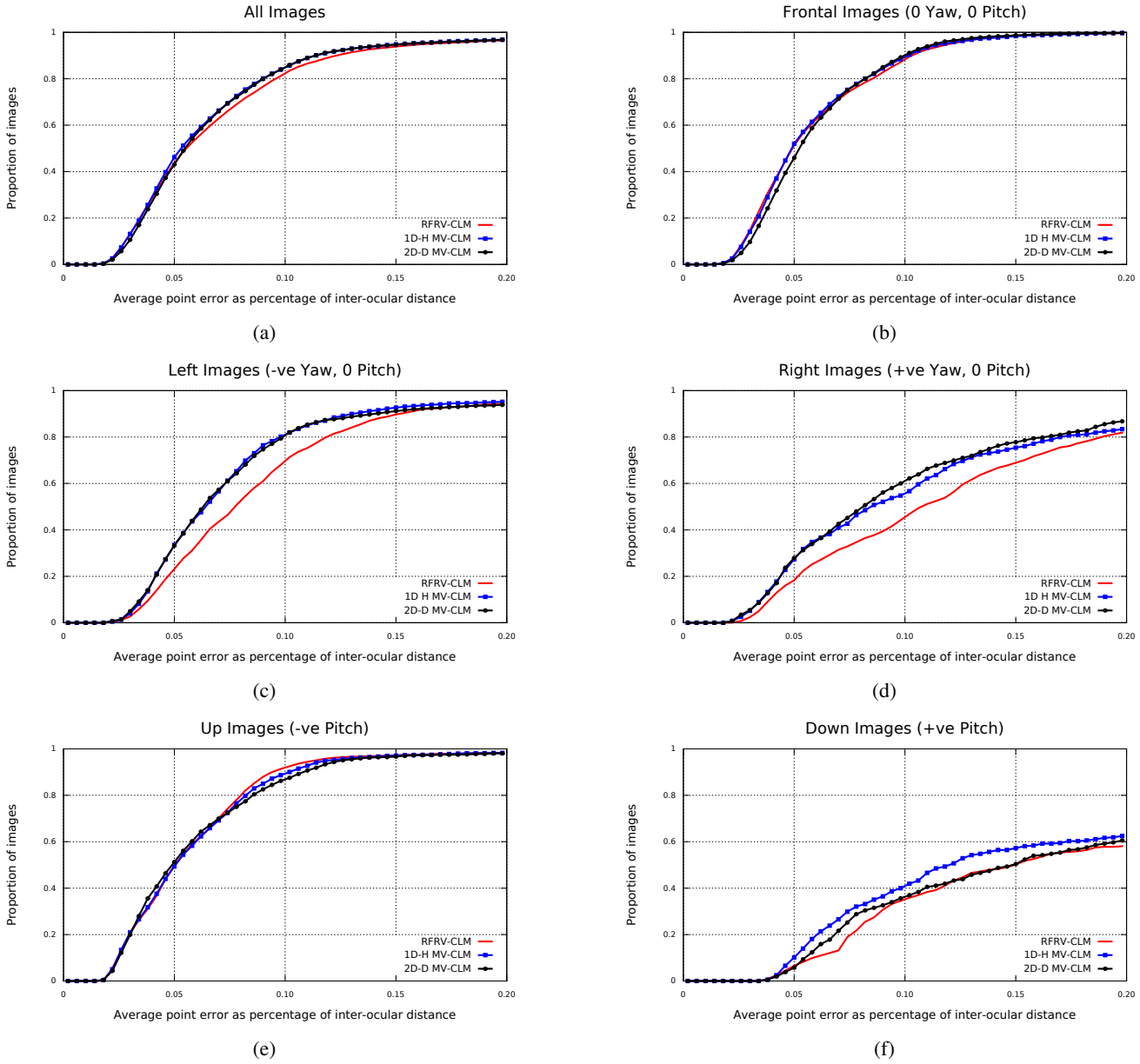

Figure 3: Driver video CDF curves for different head orientations. (a) All images. (b) Frontal images. (c) Negative yaw (left facing) images. (d) Positive yaw (right facing). (e) Negative pitch (up facing). (f) Positive pitch (down facing).

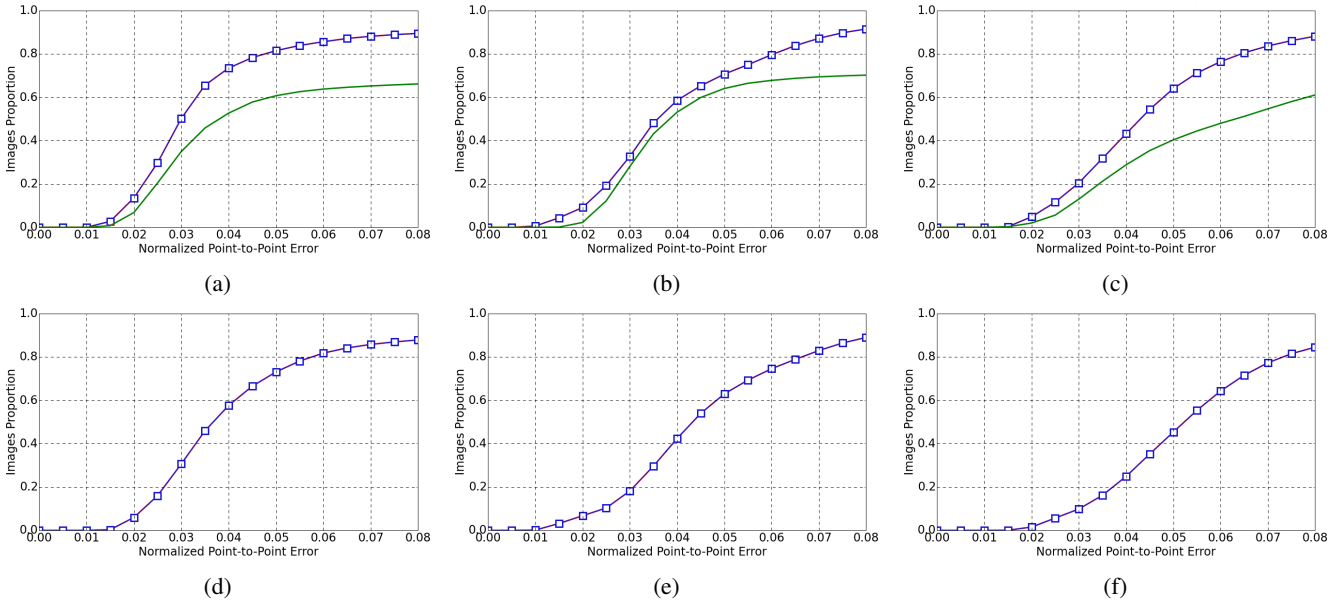

Figure 4: 300-VW Challenge CDFs of relative average point error. Green - baseline, Blue - our results. Column 1 - Category 1: Well-lit conditions with no occlusions. Column 2 - Category 2: Unconstrained conditions but without large occlusions. Column 3 - Category 3: Arbitrary conditions including occlusions, make-up and different illuminations. (a-c) Results for 49 facial points only. (d-f) Results for all 68 points included in challenge.

Figure 5: Fitting achieved by our MV-CLM as compared to the RFRV-CLM on a number of examples. Columns 1-2: Example driver video images. Columns 3-4: Example of Category 2 300-VW test video. First row: RFRV-CLM. Second row: 2D-D MV-CLM.

large amount of training data employed suggests that this is the case. In addition, is it clear that our method performs very well in all three categories, including the third category which contains examples of people recording in completely unconstrained conditions, including different illuminations, expressions and occlusions. Even in this case, our algorithm is able to achieve a success rate of more than 80% at an error of 0.07 or less for the 49 facial points, and this only increases to 0.075 in the 68 point case (including points around the face). As further illustration, we also show some examples of the improved performance of our method over the RFRV-CLM on driver images and one video from Category 2 of the 300-VW Challenge in Fig. 5.

5. Conclusions

In this paper we presented a simple but effective technique for improving the accuracy of the traditional RFRV-CLM technique on large head angles, while decreasing the computational cost. This method works by combining a global shape model with separate response maps, indexed on the shape model parameters, specifically targeted at particular head angles. We explored alternative shape-space division strategies, and identified the two optimal approaches. The performance improvements of these methods was demonstrated on two datasets, particularly on large positive and negative yaw examples. The method was also shown to give a marked speed-up over the traditional technique.

6. Acknowledgements

This research was funded by Toyota Motor Europe.

References

[1] J. Alabort-i Medina and S. Zafeiriou. Bayesian active appearance models. In *Proceedings of IEEE Intl Conf. on Computer Vision and Pattern Recognition (CVPR 2014)*, 2014. 1

[2] E. Antonakos, J. Alabort-i Medina, G. Tzimiropoulos, and S. Zafeiriou. Hog active appearance models. In *Image Processing (ICIP), 2014 IEEE International Conference on*, pages 224–228. IEEE, 2014. 1

[3] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Robust discriminative response map fitting with constrained local models. In *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2013)*, Portland, Oregon, USA, June 2013. 1

[4] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Incremental face alignment in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1859–1866. IEEE, 2014. 1, 5

[5] T. Baltrusaitis, P. Robinson, and L.-P. Morency. Constrained local neural fields for robust facial landmark detection in the wild. In *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on*, pages 354–361. IEEE, 2013. 1

[6] P. N. Belhumeur, D. W. Jacobs, D. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 545–552. IEEE, 2011. 5

[7] G. Chrysos, S. Zafeiriou, E. Antonakos, and P. Snape. Offline deformable face tracking in arbitrary videos. In *IEEE International Conference on Computer Vision Workshops (ICCVW), 2015*. IEEE, 2015. 5

[8] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6):681–685, 2001. 1

[9] T. F. Cootes, M. C. Ionita, C. Lindner, and P. Sauer. Robust and accurate shape model fitting using random forest regression voting. In *Computer Vision–ECCV 2012*, pages 278–291. Springer, 2012. 1, 2, 3, 6

[10] T. F. Cootes, C. Taylor, D. Cooper, and J. Graham. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995. 2

[11] D. Cristinacce and T. Cootes. Automatic feature localisation with constrained local models. *Pattern Recognition*, 41(10):3054–3067, 2008. 1

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

[13] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010. 5

[14] M. Kostinger, P. Wohlhart, P. M. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 2144–2151. IEEE, 2011. 5

[15] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In *Computer Vision–ECCV 2012*, pages 679–692. Springer, 2012. 5

[16] J. Shen, S. Zafeiriou, G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *IEEE International Conference on Computer Vision Workshops (ICCVW), 2015*. IEEE, 2015. 5

[17] G. Tzimiropoulos. Project-out cascaded regression with an application to face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3659–3667, 2015. 5

[18] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 532–539. IEEE, 2013. 1

[19] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1385–1392. IEEE, 2011. 1

[20] X. Yu, J. Huang, S. Zhang, W. Yan, and D. Metaxas. Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model. In *14th IEEE International Conference on Computer Vision*, 2013. 1