Markov Random Field Structures for Facial Action Unit Intensity Estimation

Georgia Sandbach* Stefanos Zafeiriou* Maja Pantic*†
*Department of Computing
Imperial College London
180 Queen's Gate
London, UK

†EEMCS
University of Twente
7522 NB Enschede
Netherlands

{gls09,s.zafeiriou,m.pantic}@imperial.ac.uk

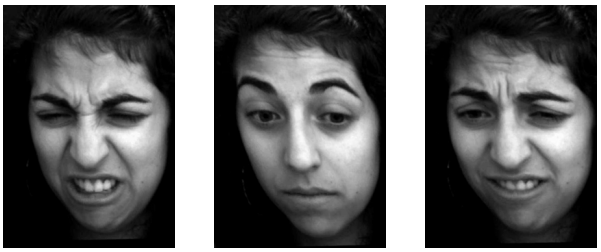
Abstract

We present a novel Markov Random Field (MRF) structure-based approach to the problem of facial action unit (AU) intensity estimation. AUs generally appear in common combinations, and exhibit strong relationships between the intensities of a number of AUs. The aim of this work is to harness these links in order to improve the estimation of the intensity values over that possible from regression of individual AUs. Our method exploits Support Vector Regression outputs to model appearance likelihoods of each individual AU, and integrates these with intensity combination priors in MRF structures to improve the overall intensity estimates. We demonstrate the benefits of our approach on the upper face AUs annotated in the DISFA database, with significant improvements seen in both correlation and error rates for the majority of AUs, and on average.

1. Introduction

Recognition of facial expressions is a challenging problem as the face is capable of complex motions, and the range of possible expressions is extremely wide. For this reason, recognition of facial action units (AUs) from the Facial Action Coding System (FACS) [5] has become a widely studied area of research. AUs are the building blocks of expressions, and are finite in number, thus allowing a comprehensive recognition system to be produced. While detection of AUs alone is an important source of information about the full expression, and thus emotional state, of a subject, knowing the full intensity of the AUs in an image or video greatly increases the richness of the information, allowing more complex emotional states to be determined. For example, in the application of pain detection, the intensity level of a subset of AUs has been shown to be important in determining the level of pain [14].

An AU rarely appears alone, and is often displayed within a combination of facial actions. The appearance of a particular action can be greatly affected by the other AUs that are active in the same region of the face, particularly those with high intensity. For example, the appearance of AU 4 (Brow Lowerer) will be hugely altered by the presence of AU 1 (Inner Brow Raiser), as can be seen in the examples shown in Fig. 1. Detecting the presence of the individual AUs may still be possible, but estimating the particular intensity of AUs in the presence of others becomes a much harder task. However, because of the nature of spontaneous expressions there are links between the presence, and intensity, of different AUs, particularly within a specific region in the face. It is knowledge about these interactions that we aim to harness in this work, through the use of Markov Random Field (MRF) structures of AUs.



(a) AU4 (+ AU9) (b) AU1 (+ AU2) (c) AU1 + AU4

Figure 1. Example that demonstrates how the combination of two AUs can be visually different from that of either one alone.

Many works have examined the problem of AU detection (e.g. [15]), in both posed and spontaneous data. Although the majority of work has looked at detection of individual AUs alone, a small amount of research has looked at modelling the semantic and dynamic relationships between AUs for the purpose of detection in posed videos [17, 18]. In these works, dynamic Bayesian networks were used to model the relationships between AUs within a time-stamp,

and also the links between different time-stamps. However, they used directed networks, to model only the co-occurrence of AUs, for the purpose of better detection.

Little research has been conducted into intensity estimation of facial action units. The main reason for this is that available data suitable for conducting research of this kind has been limited, due to the difficulties of collecting spontaneous data, which contains natural AU correlations, and fully FACS coding a database with intensities values. However, the recently acquired Denver Intensity of Spontaneous Facial Actions (DISFA) database [12] combines naturalistic data with intensity codings, and thus allows us to experiment with using intensity relationships to improve detection of intensity values. The first work to look at exploiting the interactions between AU intensity values was [10]. Here a dynamic Bayesian network was trained on the output of a multiclass SVM, to learn spatial and temporal links between AUs in order to improve over the classifier accuracy.

Other work that has looked at intensity estimation has generally only performed estimation of the individual AUs, without aiming to take into account the relationship between the intensities of different AUs. They have generally employed either posed examples, or have exploited databases that are not publicly available. The techniques previously employed include using the confidence values of Support Vector Machine (SVM) [1] or AdaBoost classifiers [6] as direct indication of intensity, employing multiple binary SVM classifiers to form a multiclass classifier which is trained on each intensity is a separate class [11], or using regression based techniques such as Relevance Vector Machines (RVMs) [9] and Support Vector Regressors (SVRs) [8, 16, 7].

In this work we propose a novel parts-based method for full estimation of AU intensities, which employs tree-based MRF structures. We adopt such a structure type because we are able to perform exact inference on the intensity values on random fields of this kind [2, 19, 20]. We exploit SVRs trained on selected Local Binary Pattern features, and combine this input with an AU combination prior to improve the estimation result. We show that this approach can significantly improve the estimation from regressors alone, and so demonstrate that harnessing the relationships in AU intensities is important for better expression recognition systems.

In summary, the contributions of this paper are:

- We propose the first AU intensity MRF structure-based approach for recognition of facial action units.
- We propose a method for building a number of tree-based models that take as root the maximum intensity AU in the expression.
- We demonstrate the effectiveness of our approach by showing significant correlation and error improvements in the intensity estimation over regression alone, when tested on the DISFA database.

2. An Action Unit Intensity Markov Random Field Structure

Here we propose a novel Markov Random Field (MRF) structure-based approach for modelling combinations of AU intensities within a particular face region in a set of images. Let I be the relevant image region, which contains N AU parts with intensities $\Lambda = \{\lambda_1, ..., \lambda_N\}$. We build a set of part-based models, $M = \{T_1, ..., T_N\}$, each of which takes the form of a tree, $T_i = (V, E)$. In these graphs, the vertices $V = \{v_1, ..., v_N\}$, are the AU parts that could be present within the image region with an intensity ranging from A-E. This set of parts is the same for all region trees. There are also a number of edges $(v_i, v_j) \in E$, which connect pairs of these parts. These can be thought of as springs which are stretched by varying degrees depending on the difference in intensities between the two AU parts, v_i and v_j.

2.1. The Model

The probabilistic model lets us assume as random variables the number of intensities of a number of AUs in a certain facial region, Λ. We formulate the posterior probability of this combination, given a feature descriptor, $\phi(I)$, of the image region, and model parameters Θ, can be written as:

$$p(\Lambda|\phi(I), \Theta) \propto p(\phi(I)|\Lambda, \Theta)p(\Lambda|\Theta) \tag{1}$$

where $p(\phi(I)|\Lambda, \Theta)$ is the likelihood of the feature descriptor given the configuration and set of model parameters, and $p(\Lambda|\Theta)$ is the intensity combination joint prior distribution over AU intensities.

We can define our likelihood probability for $\phi(I)$ in terms of the individual likelihoods given the intensity of each individual part and corresponding parameters:

$$p(\phi(I)|\Lambda, \Theta) = \prod_{i=1}^{N} p(\phi(I)|\lambda_i, \theta_i) \tag{2}$$

where $p(\phi(I)|\lambda_i, \theta_i)$ is the likelihood of the feature descriptor given that AU part v_i has an intensity λ_i. The particular choice of likelihood functions is described in the next section.

The prior probability, $p(\Lambda|\Theta)$, models the relationships between AU intensities, and can also be simply split, as in the general form for a MRF:

$$p(\Lambda|\Theta) = \frac{1}{Z} \prod_{(v_i, v_j) \in E} p(\lambda_i, \lambda_j|\theta_{i,j}) \tag{3}$$

where $p(\lambda_i, \lambda_j|\theta_{i,j})$ is the prior probability of the intensity combination λ_i and λ_j for parts v_i and v_j respectively, and Z is the partition function which is equal to

$\sum_{\Lambda} \prod_{(v_i,v_j)\in E} p(\lambda_i, \lambda_j|\theta_{i,j})$. This acts as a normalisation parameter, and in our case can simply be set to 1. Due to the fact that we adopt a tree-based MRF method in this work, exact inference can be computed. Hence our prior is simply:

$$p(\Lambda|\Theta) = \prod_{(v_i,v_j)\in E} p(\lambda_i, \lambda_j|\theta_{i,j}) \tag{4}$$

Thus the posterior distribution becomes:

$$p(\Lambda|\phi(I), \Theta) \propto \prod_{i=1}^{N} p(\phi(I)|\lambda_i, \theta_i) \prod_{(v_i,v_j)\in E} p(\lambda_i, \lambda_j|\theta_{i,j}) \tag{5}$$

We can use this to define an energy minimisation function for each model that must be minimised in order to match the model to an image region. If we take the negative logarithm of both sides, and set $a_i(\lambda_i) = -\log(p(\phi(I)|\Lambda, \Theta))$ and $c_{i,j}(\lambda_i, \lambda_j) = -\log(p(\Lambda|\Theta))$, we can write equation 5 as:

$$f(\Lambda) = \sum_{i=1}^{N} a_i(\lambda_i) + \sum_{(v_i,v_j)\in E} c_{i,j}(\lambda_i, \lambda_j) \tag{6}$$

This equation gives the energy function to be minimised in order to identify the optimal intensity combination. It consists of two components: $a_i(\lambda_i)$ is defined as an appearance function that measures mismatch for an image region when part v_i is given an intensity of λ_i, and $c_{i,j}(\lambda_i, \lambda_j)$ as a combination function that measures the deformation in the model when part v_i has an intensity of λ_i, v_j an intensity of λ_j, and v_i and v_j are a pair of connect parts.

Now we need only find the appearance likelihood functions for each part, and to calculate a suitable set of prior intensity combinations, and we can combine these to allow us to estimate the optimal intensity values. The benefit of taking only combination priors is that information about the relative intensities is encoded, rather than the absolute intensities, for which is it difficult to define a useful prior.

2.2. Tree Structures

In order to create meaningful structures that are able to model the interdependancies between AUs, without adding loops, we use a set of tree graphs, $\{T_1, ..., T_N\}$, each of which has the equivalent AU part, $\{v_1, ..., v_N\}$, as the root node, as shown in Fig. 2. The aim of this is to allow each tree to best model the cases where the root AU has the highest intensity, as this is when we would expect that it will most impact on the appearance of the other AUs in the region. For this reason, we take only training examples where this is the case to construct each tree. There are many ways to learn the tree structures, we compare two alternatives: the use of ad-hoc built star-trees, where all parts have the root as parent node, and the application of an adapted version of the Chow-Liu algorithm [3] for building tree structures.

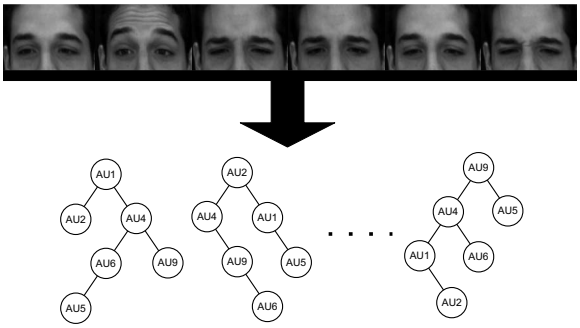


Figure 2. We build a set of Markov Random Field trees to represent the possible AU intensity combinations in a particular face region.

First the training labels are employed to calculate the mutual entropy between all pairs of AUs. We take two sets of parts: V_T are the parts in the tree, and V_L are the parts still to add, where $V_T \cup V_L = V$. Starting from the root AU, which is v_i for T_i, we add this to V_T, and let V_L consist of all other parts. In the first step, we take the two parts in V_L which have the two highest mutual entropy scores with the root. They are added as child nodes of v_i, and are moved from V_L to V_T. This step ensures that the root has at least two children, which improves the impact of the other parts on the root node. Then the highest mutual entropy score, for any pair of nodes $v_p \in V_T$ and $v_c \in V_L$, is identified. v_c is then added as a child of v_p in the tree, and moved to V_L. The algorithm then repeats this step until all parts have been added into the tree, and V_L is empty.

2.3. Model Parameters

Given a tree structure, we compute distribution parameters for each likelihood function in Equation 5. Here we describe the general approach, regardless of input. In Section 3 we will detail the particular form of the input we used in this paper. This method assumes discrete intensity labels. However, it could be extended to deal with continuous intensity labels by taking ranges of values around each integer label.

Appearance Function We need to define $a_i(\lambda_i) = -\log(p(\phi(I)|\lambda_i, \theta_i))$, for each part v_i, given the feature descriptor $\phi(I)$ extracted from our image region. Taking the subset of examples in the training set for which the intensity label is λ_i, we apply a part-specific function g_i. We then model the likelihood as a Gaussian distribution, with mean μ_{i,λ_i} and standard deviation σ_{i,λ_i}, over the outputs of this function:

$$a_i(\lambda_i) = -\log\left(\frac{1}{\sqrt{2\pi\sigma_{i,\lambda_i}^2}} exp\left(\frac{-(g_i(\phi(I)) - \mu_{i,\lambda_i})^2}{2\sigma_{i,\lambda_i}^2}\right)\right) \tag{7}$$

Hence, we calculate the parameters of this distribution for each possible intensity value, $\lambda_i = \{0, ..., 5\}$: $\theta_i = \{\mu_{i,0}, \sigma_{i,0}, ..., \mu_{i,5}, \sigma_{i,5}\}$, to allow calculation of the appearance likelihoods.

Combination Function We need to define $c_{i,j}(\lambda_i, \lambda_j) = -\log(p(\lambda_i, \lambda_j|\theta_{i,j}))$ for each connected pair of AU parts. We want to assign a prior probability for each possible intensity combination for these parts, given a set of training labels, L. For an edge $(v_i, v_j) \in E$, we can take the joint prior as the distribution of the training labels across the possible intensity combinations of the parts v_i and v_j:

$$c_{i,j}(\lambda_i, \lambda_j) = -\log\left(\frac{|L_{v_j=\lambda_j} \cap L_{v_i=\lambda_i}|}{|L|}\right) \quad (8)$$

where $L_{v_i=\lambda_i}$ is the set of training examples for which part v_i has intensity λ_i.

2.4. Inference

In order to minimise Equation 5, we employ a method based on the well known Viterbi algorithm, and exploited in [4] for efficient inference on MRF structures.

Starting from a leaf node in the tree (i.e. a part with no children), v_j, we can compute the best intensity value, λ_j, given each possible intensity value of its parent part, λ_i, by minimising simply the appearance and combination mismatches at each parent intensity:

$$f_j(\lambda_i) = \min_{\lambda_j}(a_j(\lambda_j) + c_{i,j}(\lambda_i, \lambda_j)) \quad (9)$$

and storing the intensity value λ_j at each parent intensity.

This minimum energy can then be taken as message from the child node, which summed together across all children contribute to the parent energy function. If v_j is now a non-root part with children, C_j, then the function to minimise becomes:

$$f_j(\lambda_i) = \min_{\lambda_j}\left(a_j(\lambda_j) + c_{i,j}(\lambda_i, \lambda_j) + \sum_{v_c \in C_j} f_c(\lambda_j)\right) \quad (10)$$

and the intensity values λ_j can again be stored for each parent intensity.

Finally the optimal root intensity can be computed by summing all messages from child parts with the appearance score and minimising across possible intensity values. The optimal intensities can then be found for all parts in the tree by backtracking back down the tree using the parent intensity to identify the best intensity value of each part.

3. Methodology

In the previous section we described the model framework we employ for AU intensity estimation. Here we describe the full methodology for training and testing this model. This consists of a number of steps: feature extraction and selection, SVR regression parameter optimisation and training, and finally calculating the tree parameters and testing. An overview of our system can be seen in Fig. 3.

3.1. Feature Descriptor and Selection

The first stage in our system is to use the given facial landmarks to perform alignment of the images. Exploiting the calculated positions of the eyes and nose, we transform the images into a pre-defined frame to ensure alignment of facial points suitable for feature extraction.

We employ Local Binary Pattern features [13] as the feature type. This technique provides a simple but useful way of encoding the texture shape. It works by defining a circular neighbourhood around each pixel in the image, and assigning zeros and ones to each point in this neighbourhood according to whether the intensity at these points is higher or lower than that of the central pixel. When these digits are taken together this forms a binary number which then encodes the shape around the pixel. Histograms can then be used to form feature descriptors from regions in the image, and concatenation of these gives a full feature descriptor for the image.

Feature selection is then performed in order to extract the most discriminative subset of features. We use GentleBoost, a more stable version of the AdaBoost algorithm, for this purpose. This algorithm uses weak classifiers at each iteration to choose the most discriminative feature, and then weights (boosts) the examples which are misclassified in order to focus the classifiers in the next iteration on these. To avoid overfitting, our strategy is to run the selection algorithm repeatedly, removing the previously chosen features at each stage, until the number of features selected exceeds the chosen threshold, set to 200 in this case.

The input to the feature selection algorithm is taken as the AU classification labels, i.e. presence or absence of the AU, rather than intensity values. This was shown to give a better set of discriminative features for use in regression.

3.2. Model Parameter Input

In order to calculate the appearance distributions for each part in the trees, we employ Support Vector Regressors (SVRs). This regression technique aims to fit a function to the data points that both ensures minimum error, whilst also aiming to produce as smooth an output as possible. We employ the SVRs with a histogram intersection kernel, $k(\mathbf{h}_i, \mathbf{h}_j) = \sum_{n=1}^{N} min(x_{in}, x_{jn})$. We train one SVR, r_i, for each part, v_i. They are first parameter optimised using three-fold cross-validation on a portion of a validation set, which is chosen to be AU specific. Then they are trained on the remaining portion of the validation set intensity values for v_i, including labels of zero to represent absence of the

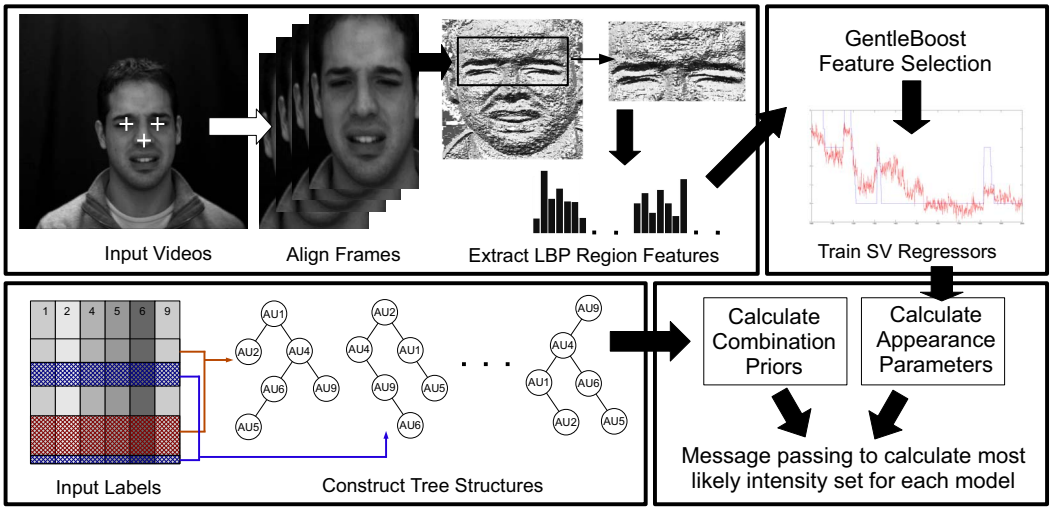


Figure 3. An overview of our full system.

AU.

The regressor is then tested against a training set, to produce output values which can be used to calculate the parameters, θ_i. First, the overall mean value of the output is subtracted from all labels. This step removes the subject specific bias introduced during the regression testing. This results in our function g_i taking the following form:

$$g_i(\phi(I)) = r_i(\phi(I)) - \hat{r_i} \tag{11}$$

where $\hat{r_i}$ is the mean regressor output.

The mean and standard deviation of the resulting output is calculated for the set of frames attached to each intensity value, λ_i. These can then be used to calculate the log likelihood of a particular normalised regressor output occurring at this intensity value. We also calculate the prior distribution of intensity combinations for each connected pair in each of the trees we have constructed, as described in Section 2.2

3.3. Testing

We then perform inference, as outlined in Section 2.4. In this case, the appearance likelihoods are calculated with the equivalent g_i function, except here the mean is calculated from the output of the regressor over the test set. In practise, the likelihoods are normalised by mapping the data values onto a standard normal distribution, to give useful likelihood probabilities.

Once inference has been performed, the output is a $N \times N$ matrix of intensity predictions, where each row corresponds to the output from one tree, and each column gives the predictions from the trees for each AU, along with a likelihood score for each tree. We examine the results from two estimation methods given this output:

1. Choosing the root value from each tree as the estimation of the intensity of the AU - as the root is the most highly influenced part in the tree, this estimation is be expected to be superior to those of the other parts in the tree.

2. Choosing the most likely tree to give estimations of the full set of AUs - allowing the model to predict which AU is root of the tree, and hence picking the most likely intensity combination.

We present the results of both methods in the next Section.

4. Experiments

We conducted experiments on the Denver Intensity of Spontaneous Facial Actions (DISFA) database [12], one of only two naturalistic databases that have been FACS coded with AU intensity values. The intensities are recorded as values between 0-5, where 0 denotes the absence of the AU, and 1-5 represent A-E intensities. This database consists of 27 subjects, each recorded whilst watching a 4-minute (242 seconds) video clip by two cameras, left and right. The FACS coding included consists of 12 AUs: 1, 2, 4, 5, 6, 9, 12, 15, 17, 20, 25, 26. 66 facial landmarks are provided for each frame in each video. In this work we utilise only the left camera view, and exploit the available landmarks in order to align all frames by transforming images in order to match the eye and nose locations to an ideal position. There are a number of frames in each video for which the automatic landmarking method was known to have failed, thus resulting in the landmarks provided for these frames being inadequate for alignment. In our experiments these frames were simply removed from all data sets. We conducted preliminary tests on only the upper region of the face in order to

	Pearson Correlation Coefficient						Root Mean Squared Error					
AU	SVR	App	Star 1	Star 2	Built 1	Built 2	SVR	App	Star 1	Star 2	Built 1	Built 2
1	0.232	0.373	0.553	0.536	0.563	**0.568**	1.553	1.341	**0.621**	0.627	**0.621**	0.624
2	0.336	0.514	0.501	0.532	0.541	**0.552**	1.231	0.663	0.590	**0.577**	0.592	0.584
4	0.395	**0.485**	0.408	0.409	0.438	0.437	1.490	1.170	1.105	1.104	1.099	**1.094**
5	0.075	0.149	0.214	0.215	**0.226**	0.172	0.981	1.230	0.304	0.302	0.338	**0.277**
6	0.045	-0.042	0.094	**0.142**	0.119	0.141	1.797	2.619	0.822	**0.770**	0.897	0.836
9	0.093	0.131	0.038	0.004	**0.168**	0.014	1.375	1.672	0.586	**0.582**	0.612	**0.582**
Ave	0.196	0.268	0.301	0.306	**0.342**	0.314	1.404	1.449	0.671	**0.660**	0.693	0.666

Table 1. Full Upper Face AU Results. We compare PCC and RMSE scores for six cases: (1) Raw regressor output (2) Most likely appearance outputs (3) Star trees root intensity values (4) Most likely star trees (5) Built trees root intensity values (6) Most likely built trees.

establish the benefits of our method. The region used shows the eyes, forehead and top of the nose. The annotated AUs active in this region are 1, 2, 4, 5, 6, and 9. Note that though 9 is a lower face AU, it still impacts on the appearance of this region, and so is included in the set.

We use the leave-one-out protocol for testing our method. This means that the video for one subject forms our testing set, with the remaining subjects available for validation and training. We form AU specific validation sets from one third of these subjects. They are used for feature selection, parameter optimisation and training of the SVRs. In these sets we aim to overcome the problem that most AUs only appear occasionally in the videos, and so the majority of frames are labelled as 0. To create a more balanced dataset for feature selection and regression training we take all of the frames for which this is not the case (i.e. the labels are 1 or higher), and then take five times this many 0 frames. We exclude subjects that do not demonstrate the AU at all. We also remove all misaligned frames of either type. This still means there will a much larger number of neutral frames than any of the other labels in the set, but results in a mostly balanced set for feature selection (where the labels are 1 or -1 as it uses just presence/absence) and also means that the regressors will be well trained for neutral frames which is desirable as they dominate the training and testing sets. For feature selection we employ the full validation set, but we then use one quarter, divided again by subject, to parameter optimise the regressors, and employ the remaining three quarters to train them.

The training set is formed from the remaining two thirds of the subjects (excluding the subject reserved for testing) and takes all frames in these videos, minus any problem frames. Hence there is a single training set for each fold, rather than being AU specific. This data set is tested against each regressor in order to create a training set of outputs for calculating the appearance parameters of each part. The labels for this dataset are also used to calculate the intensity combination priors.

4.1. Overall Performance

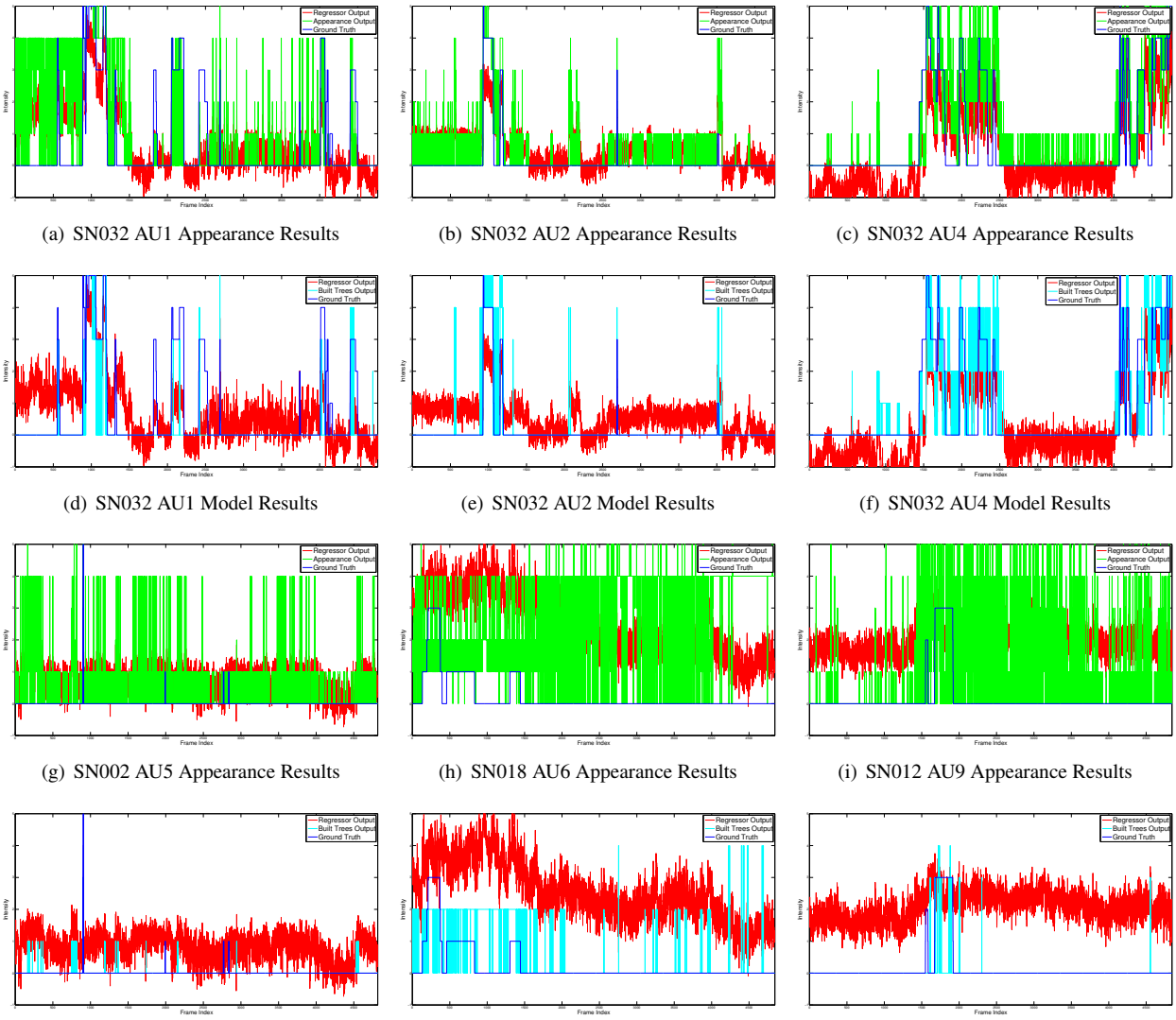
The full results are shown in Table 1. Here we show two performance measures: The Pearson Correlation Coefficient (PCC), and the Root Mean Squared Error (RMSE). In order to establish the benefits of employing our MRF-based method, we compare the results to those obtained from the appearance information alone. In the first column of each half we show the raw SVR performance, and in the second column we display the results achieved if we take only the most likely intensity based on the SVR output (i.e. taking only the first term of Equation 5 into account). We then show four sets of results for experiments, exploring both possible estimation techniques described in Section 3.3, with two possible tree structures:

1. Star-shaped trees where all child nodes are connected to the root.

2. Our automatically generated tree structures as described in Section 2.2.

As can be seen from these results, our method demonstrates a significant improvement, in terms of both correlation and error. The highest average correlation result is given by built trees where the root intensity values are taken. This method achieves a score of 0.342, far higher than those achieved with the regressors and appearance outputs alone, 0.196 and 0.268 respectively. The average correlation is also improved with all other MRF structure methods, though to a lesser extent. However, the built trees outperform star trees in both cases. The RMSE mean scores also display an improvement with our method over regression and appearance outputs, however in this case the star trees achieve a lower error than the built trees, with the maximum likelihood trees in both cases giving the lowest error.

The MRF methods achieve better results in correlation and error scores for the majority of AUs, with the correlations of AU1, AU5, and AU6 all more than doubling when built trees are employed, both when the root intensities are taken, and in the most likely trees, and improving to a lesser degree with star trees. Improvements are also seen



(a) SN032 AU1 Appearance Results (b) SN032 AU2 Appearance Results (c) SN032 AU4 Appearance Results

(d) SN032 AU1 Model Results (e) SN032 AU2 Model Results (f) SN032 AU4 Model Results

(g) SN002 AU5 Appearance Results (h) SN018 AU6 Appearance Results (i) SN012 AU9 Appearance Results

(j) SN002 AU5 Model Results (k) SN018 AU6 Model Results (l) SN012 AU9 Model Results

Figure 4. Subject examples of estimation from appearance alone versus our model. (a)-(c) and (g)-(i) show the results achieved with the regressors and appearance outputs, against the ground truth. (d)-(f) and (j)-(l) show the results from the built tree structure-based MRF model, against the regressor and ground truth.

for AU2 for built trees in both cases, and the most likely star trees, and for AU9 when the root intensities are taken from the built trees. The RMSE results show improvements for all AUs, for all tree structure and estimation combinations, which shows the benefit of the model approach for all cases. A notable exception to the general trend is the correlation scores for AU4, where the appearance output alone achieves the highest result, outperforming the built trees slightly. This suggests that the combination prior information detracts in this case. However, the RMSE results still show a small improvement. These conflicting outcomes suggest that the appearance and model perform comparably well on average, but there are large subject differences.

The correlation and error results here show that both the star-shaped trees and the automatically built structures have benefits over the regression and appearance alone, but do not give a clear indication of which method is superior. The structure of the trees does appear to impact the performance, but is not consistent across AUs. This suggests that more extensive exploration of these methods is required, with a wide range of further testing, in order to establish which tree structure, and estimation method, is superior. Though the resulting correlations are still low, particularly for AU5, AU6 and AU9, these results demonstrate that exploiting the relationships between different AUs allows a better signal of the intensity values to be extracted, and thus with better regressor inputs an accurate estimate may be possible

through methods of this kind.

4.2. Individual Subject Performance

In Fig. 4 we show an example for each AU of the improvement in intensity estimation shown by our models. Figs. 4(a)-4(c) and 4(g)-4(i) show the resulting intensities predicted by the regressors and appearance likelihoods, in red and green respectively, as compared to the ground truth shown in blue. This is compared to Figs. 4(d)-4(f) and 4(j)-4(l), where the model output, with automatically built trees, is displayed in light blue. Though there still are a large number of errors in these examples, when employing our MRF method, they show how the impact of the combination prior knowledge, and tree structure, can be dramatic, greatly reducing errors shown in the appearance output alone.

5. Conclusions and Future Work

In this work we have shown how MRF structures can be successfully applied to the problem of facial action unit intensity estimation. We have presented a method that trains SVRs on LBP features, and uses the output of these to estimate appearance likelihoods of each AU. These are then combined with AU combination priors, in the MRF structures, in order to estimate the intensity of AUs present in a region of the upper face. We have demonstrated that this approach achieves promising results when applied to a subset of AUs in the DISFA database, greatly improving correlation and error scores over the regression and appearance outputs alone. In our future work, we hope to extend this approach, using more complex structures to provide better ways of modelling the full relationships between AU intensities. This will include the move to discriminative methods which, it is hoped, will have the ability to capture a wider range of AU interactions.

References

[1] M. S. Bartlett, G. C. Littlewort, M. G. Frank, C. Lainscsek, I. R. Fasel, and J. R. Movellan. Automatic recognition of facial actions in spontaneous expressions. *Journal of Multimedia*, 1(6):22–35, 2006.

[2] C. M. Bishop et al. *Pattern recognition and machine learning*, volume 1. springer New York, 2006.

[3] C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *Information Theory, IEEE Transactions on*, 14(3):462–467, 1968.

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

[5] J. Hager, P. Ekman, and W. Friesen. Facial action coding system. *Salt Lake City, UT: A Human Face*, 2002.

[6] J. Hamm, C. G. Kohler, R. C. Gur, and R. Verma. Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders. *Journal of neuroscience methods*, 200(2):237–256, 2011.

[7] L. A. Jeni, J. M. Girard, J. F. Cohn, and F. De La Torre. Continuous au intensity estimation using localized, sparse facial feature space. In *2nd International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE)*, volume 6, 2013.

[8] L. A. Jeni, A. Lőrincz, T. Nagy, Z. Palotai, J. Sebők, Z. Szabó, and D. Takács. 3d shape estimation in video sequences provides high precision evaluation of facial expressions. *Image and Vision Computing*, 30(10):785–795, 2012.

[9] S. Kaltwang, O. Rudovic, and M. Pantic. Continuous pain intensity estimation from facial expressions. In *Advances in Visual Computing*, pages 368–377. Springer, 2012.

[10] Y. Li, S. M. Mavadati, M. H. Mahoor, and Q. Ji. A unified probabilistic framework for measuring the intensity of spontaneous facial action units. In *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition (FG'13)*, May 2013.

[11] M. H. Mahoor, S. Cadavid, D. S. Messinger, and J. F. Cohn. A framework for automated measurement of the intensity of non-posed facial action units. In *Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on*, pages 74–80. IEEE, 2009.

[12] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, page 1, 2013.

[13] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, pages 971–987, 2002.

[14] K. M. Prkachin and P. E. Solomon. The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain. *Pain*, 139(2):267–274, 2008.

[15] G. Sandbach, S. Zafeiriou, and M. Pantic. Binary Pattern Analysis for 3D Facial Action Unit Detection. In *Proceedings of the British Machine Vision Conference (BMVC 2012)*, Guildford, UK, September 2012. BMVA Press.

[16] A. Savran, B. Sankur, and M. Taha Bilge. Regression-based intensity estimation of facial action units. *Image and Vision Computing*, 30(10):774–784, 2012.

[17] Y. Tong, J. Chen, and Q. Ji. A unified probabilistic framework for spontaneous facial action modeling and understanding. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(2):258–273, 2010.

[18] Y. Tong, W. Liao, and Q. Ji. Facial action unit recognition by exploiting their dynamic and semantic relationships. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1683–1699, 2007.

[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.

[20] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2879–2886. IEEE, 2012.