Latent Trees for Estimating Intensity of Facial Action Units

Sebastian Kaltwang[1], Sinisa Todorovic[2], Maja Pantic[1]
[1]Imperial College London. [2]Oregon State University.

This paper formulates a new generative graphical model, called Latent Tree (LT), for estimating intensity levels of Facial Action Units (FAUs) in videos. FAU intensity estimation is an important step toward interpreting facial expressions. As input features, we use locations of facial landmark points. To address uncertainty of input, we formulate the LT model, its inference, and novel algorithms for efficient learning of both LT parameters and structure. Our structure learning iteratively builds LT by adding either a new edge or a new hidden node to LT, starting from initially independent nodes of observable features. A graph-edit operation that increases maximally the likelihood and minimally the model complexity is selected as optimal in each iteration. For FAU intensity estimation, we derive closed-form expressions of posterior marginals of all variables in LT, and specify an efficient bottom-up/top-down inference. Our evaluation on the benchmark DISFA [4] and ShoulderPain [3] datasets, in subject-independent setting, demonstrate that we outperform the state of the art, even under significant noise in facial landmarks. Effectiveness of our structure learning is demonstrated by probabilistically sampling meaningful facial expressions from the LT.

To implement the aforementioned LT model, we consider a Bayesian generative framework. We formalize our problem as that of jointly predicting multiple FAU targets, $\mathbf{T} = \{\mathbf{x}_1, ..., \mathbf{x}_T\}$, given a set of image features, $\mathbf{F} = \{\mathbf{x}_{T+1}, ..., \mathbf{x}_{T+F}\}$. Every target $\mathbf{x}_m \in \mathbf{T}$ can be defined as a vector of various attributes associated with mth FAU, and in a special case for our problem as FAU intensities. Image features $\mathbf{x}_m \in \mathbf{F}$ are defined as local descriptors of the face, which can be appearance based (e.g., patches) or locations of facial landmarks detected in a video frame.

We specify a graphical model for representing the joint distribution of targets and features, $p(\mathbf{T},\mathbf{F})$, and use the Bayes' rule to derive an elegant solution to FAU intensity estimation as

$$\hat{\mathbf{T}} = \max_{\mathbf{T}} \frac{p(\mathbf{T},\mathbf{F})}{\sum_{\mathbf{T}'} p(\mathbf{T}',\mathbf{F})}. \quad (1)$$

Our formulation has a number of advantages over existing approaches . They typically adopt the discriminative framework for directly predicting FAU intensities given the features, e.g., using Support Vector Classification (SVC), Relevance Vector Machine, AdaBoost, or ordinal Conditional Random Fields. While discriminative approaches are generally robust, we experimentally demonstrate in this paper that they underperform under the aforementioned challenges. In particular, due to frequent partial occlusions of the face or large out-of-plane head movements in non-staged video, some input features might be missing or very unreliable. Our results show that our model can robustly handle missing input features by marginalizing them out, unlike the competing discriminative approaches. Also, our model is less likely to overfit to training human subjects, due to the joint modeling of all FAUs $\mathbf{T}$ and features $\mathbf{F}$.

For effectively capturing statistical dependencies among $\mathbf{T}$ and $\mathbf{F}$, our model has hidden (latent) random variables. Also, for ensuring modelling efficiency (e.g., few model parameters) and efficient inference of $\hat{\mathbf{T}}$, we organize the hidden variables in a tree structure, and hence call our model Latent Tree (LT). In LT, leaf nodes represent $\mathbf{T}$ and $\mathbf{F}$, and all other nodes correspond to the hidden variables (also called hidden nodes). Importantly, no other restrictions are placed on the model structure beyond the tree structure, defined by the total number of hidden nodes and edges.

LT structure is unknown a priori. We specify a new algorithm for efficient learning of both model parameters and model structure on training data. Our structure learning iteratively builds LT by introducing either new parent nodes or new connections between existing hidden nodes, depending on the resulting increase in the joint likelihood $p(\mathbf{T},\mathbf{F})$. Our key contribution here is a heuristic algorithm for efficiently computing the maximum likelihood increase.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.

For FAU intensity estimation, we derive closed-form expressions of posterior marginals of all variables in LT, and specify an efficient inference of $\hat{\mathbf{T}}$ given $\mathbf{F}$ in two passes – bottom-up and top-down.

We evaluate LT on the benchmark DISFA [4] and ShoulderPain [3] datasets, which provide per-frame FAU intensity labels for spontaneous facial expressions. We compare our LT model for predicting all FAU targets simultaneously (LT-all) to Support Vector Classification (SVC), Support Vector Regression (SVR) (both using LIBSVM [1]) and Binary Latent Trees (BLT) [2]. Additionally to the evaluation on clean data, we create random noise to corrupt the test features. The noise is created with different severity levels: 50% noise features means that for every testing instance, we randomly select 50% of the feature dimensions and replace them with a randomly sampled value from a Gaussian distribution that has the same mean and variance as the overall training dataset. The noise is only influencing the test data, i.e. the models are trained on clean data.

Fig. 1 shows the results for FAU17 and the average of all FAUs (avg) as the noise level varies. The highest result is at 0% noise and the performance deteriorates as the noise increases. The performance drop of our LT model is slower than the other models, and it is even possible to beat other models as the noise increases, see FAU17: at about 50% noise, out performance is better than SVR, although SVR has the better performance on clean data. This is due to our generative model, which is able to ignore noisy data that is not consistent with the clean dimensions.

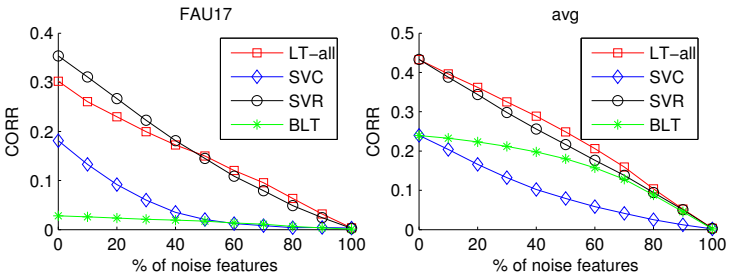

Figure 1: Results on the DISFA data for FAU17 and the average of all FAUs (avg). The LT-all model is compared to SVC, SVR and BLT. Each graph shows the correlation (CORR) as the percentage of noise feature varies.

In conclusion, our novel latent tree (LT) model shows superior performance, even under significant noise introduced to facial landmark points. We also demonstrate effectiveness of our structure learning by probabilistically sampling locations of facial landmark points, conditioned on a given FAU intensity. Our generative sampling produces plausible facial expressions.

[1] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27, 2011.

[2] S Harmeling and C K I Williams. Greedy Learning of Binary Latent Trees. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(6):1087–1097, 2011.

[3] Patrick Lucey, J.F. Cohn, K.M. Prkachin, P.E. Solomon, and Iain Matthews. Painful data: The UNBC-McMaster shoulder pain expression archive database. In *Int. Conf. Autom. Face Gesture Recognit.*, pages 57–64. IEEE, 2011.

[4] S Mavadati, M Mahoor, K Bartlett, P Trinh, and J F Cohn. DISFA: A Spontaneous Facial Action Intensity Database. *IEEE Trans. Affect. Comput.*, 4(2):151–160, 2013.