AU Classification using AAMs and CRFs

Laurens van der Maaten [a] Emile Hendriks [a]

[a] *Delft University of Technology, Mekelweg 4, 2628 CD Delft*

Automatic facial expression recognition is an important problem in social signal processing that has applications ranging from treatment of autistic children to monitoring of conflict situations [6]. In psychology, facial expressions are generally described using the Facial Action Coding System (FACS; [2]), in which each facial muscle is referred to as an action unit (AU) that is present (i.e. muscle contracted) or not present (i.e. muscle relaxed). We developed a system that automatically classifies AUs based on (variations in) facial texture and shape features. The feature extraction is performed with the help of an active appearance model [1]. Detection of AU presence is performed by training a newly developed structured prediction algorithm [5] on the features thus obtained. A complete description of our system was published in [4].

Facial feature point detection. Active appearance models (AAMs) are deformable template models that consist of two main parts: (1) a facial shape model and (2) a facial texture model. The shape model is obtained by performing PCA on a collection of manual feature-point annotations that are Procrustes-aligned, i.e. by fitting a linear-Gaussian latent variable model to normalized feature points. The resulting model can be used to generate likely configurations of facial feature points. The texture model is obtained by performing PCA on a collection of shape-normalized texture images, i.e. by using the feature point annotations to warp the annotated face images to a single coordinate frame and fitting a linear-Gaussian latent variable model to the resulting normalized face images. The texture model can be used to generate likely facial appearances that are normalized for shape variations in the face. AAMs combine the shape and texture model by warping the generated facial texture onto the generated facial shape. Identification of facial feature points using AAMs is performed by maximizing the likelihood of the observed face image under the AAM with respect to its latent variables, i.e. by minimizing the squared error between the observed face image and the facial appearance generated by the AAM.

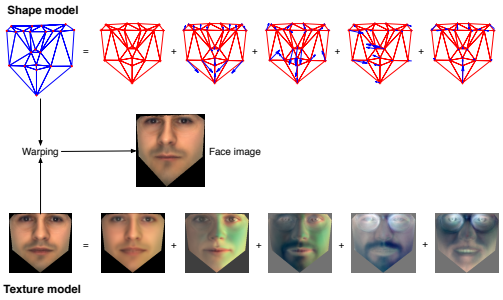

Figure 1: Active appearance modeling: (1) the face shape is made by adding a linear combination of shape components to the base shape, (2) the facial texture is made by adding a linear combination of the texture components to the mean texture, and (3) the final face image is made by warping the texture onto the shape.

Feature extraction. Using the identified facial feature points, we extract two main types of features: (1) normalized shape variations and (2) shape-normalized texture variations. Normalized shape variations (NSV) measure the difference between feature point locations in the current frame and feature point locations in the first frame of each movie. NSV features are computed by Procrustes-aligning the identified face shape to the base shape to remove rigid transformations and, subsequently, subtracting the resulting coordinates from the coordinates in the first frame. Whilst shape variations are important in AU detection predicting the presence of, e.g., action unit 24 (lip pressor) requires textural information on wrinkles. Shape-normalized texture variations (SNTVs) capture such texture information by warping each face image onto the base shape (see Figure 3), and subtracting the resulting image from the the shape-normalized image in the first frame.

Action unit detection. The detection of AU presence based on the extracted facial shape and texture variation features is performed via an extension of the conditional random field (CRF) model called hidden-unit CRF [5]. Hidden-unit CRFs model latent structure in the data that is relevant for classification in

a collection of binary stochastic variables that are conditionally independent given the data and the label sequence (see Figure 2). This conditional independence property facilitates efficient inference and learning. The resulting model can represent much more complex decision boundaries than standard CRFs.

Experiments. We performed experiments on the Cohn-Kanade data set, which comprises 593 short movies of 123 subjects performing a single posed expression. The movies are labeled for the AUs present in the movie. We measure generalization errors of classifiers trained to predict the presence of single AU in terms of the area under the ROC curve (AUC) via leave-one-subject-out cross-validation (see Table 4) on both feature sets. The results show that both types of features provide a lot of information on the presence of AUs: with these performances, our system will likely pass official FACS-certification exams. For AUs that cause large movements of feature points (like AU1, AU2, and AU25), shape variations (NSV) are the most informative features, whereas texture variations (SNTV) are most informative in recognizing the other AUs.

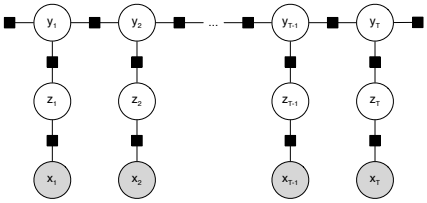

Figure 2: Factor graph of hidden-unit CRF: $\mathbf{x}$ denotes data units, $\mathbf{z}$ denotes binary stochastic hidden units, and $\mathbf{y}$ denotes 1-of-K label units.

(a) Example face 1. (b) Example face 2.

Figure 3: Shape-normalized face textures: feature points are in the same location.

AU	Name	NSV	SNTV
1	*Inner Brow Raiser*	**0.8947**	0.8834
2	*Outer Brow Raiser*	**0.9278**	0.9270
4	*Brow Lowerer*	0.8277	**0.8667**
5	*Upper Lip Raiser*	0.8857	**0.9070**
6	*Cheek Raiser*	**0.8740**	0.8691
7	*Lip Tightener*	0.8484	**0.8633**
9	*Nose Wrinkler*	**0.9415**	0.9401
11	*Nasolabial Deep.*	0.8818	**0.9270**
12	*Lip Corner Puller*	0.9171	**0.9222**
15	*Lip Corner Depr.*	0.9178	**0.9239**
17	*Lower Lip Depr.*	0.9017	**0.9125**
20	*Lip Stretcher*	0.8713	**0.8918**
23	*Lip Tightener*	0.9399	**0.9412**
24	*Lip Pressor*	0.9275	**0.9408**
25	*Lips Part*	**0.9075**	0.8961
26	*Jaw Drop*	0.8847	**0.8876**
27	*Mouth Stretch*	0.9455	**0.9459**
ALL	*Averaged*	0.8997	**0.9086**

Figure 4: Average AUC of hidden-unit CRFs on both features. Best performance is boldfaced.

Future work. An important issue of the current system is that its performance is not robust under out-of-plane rotations or partial occlusions of the face. We did not test our system in such situations, because to the best of our knowledge, there are no publicly available databases that contain FACS-coded videos with out-of-plane rotations and/or occlusions. In future work, we aim to extend our CRF models to exploit correlations between action units [3]: for instance, if we detect the presence of action unit AU12, the probability that AU13 or AU14 are also present increases; our models should exploit this information. We also plan to use our system to detect basic emotions and higher-level social signals, such as agreement and disagreement, by learning mappings from action unit labels to these emotions or signals.

References

[1] T.F. Cootes, G. Edwards, and C.J. Taylor. Active appearance models. In *Proceedings of the European Conference on Computer Vision*, volume 2, pages 484–498, 1998.

[2] P. Ekman and E. Rosenberg. *What the face reveals (2nd edition)*. Oxford, New York, NY, 2005.

[3] C. Sutton, A. McCallum, and K. Rohanimanesh. Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data. *Journal of Machine Learning Research*, 8(Mar):693–723, 2007.

[4] L.J.P. van der Maaten and E.A. Hendriks. Action unit classification using active appearance models and conditional random fields. *Cognitive Processing*, 2012.

[5] L.J.P. van der Maaten, M. Welling, and L.K. Saul. Hidden-unit conditional random fields. *JMLR W&CP*, 15:479–488, 2011.

[6] A. Vinciarelli, M. Pantic, and H. Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27:1743–1759, 2009.