Supplementary Material for
Learning to Annotate Facial Action Units for Online Images

Anonymous ICCV submission

Paper ID 2547

Abstract

This material includes more results that cannot be fitted into the main paper due to space limitation. In specific, we provides more qualitative analysis of the proposed weakly supervised clustering method.

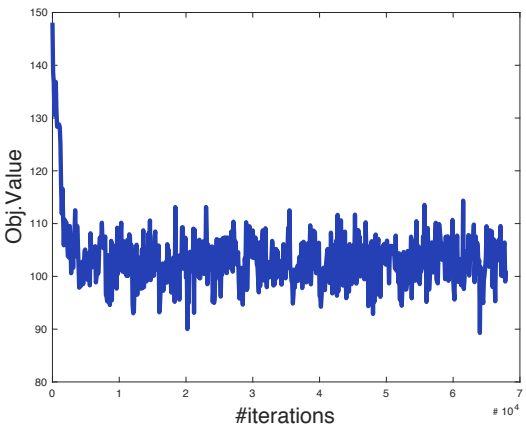
1. Qualitative analysis on weakly supervised clustering

This section shows more qualitative results of the proposed scalable weakly supervised clustering (WSC), t-SNE visualization of WSC embeddings and the corrections of annotations. Recall that we modeled WSC as a variant of spectral clustering with scatterness constarints, which learns an optimal embedding space where images with similar appearance and weak annotations are more likely to be in the same neighborhood. Details can be found in Section 3.1 of the main paper. Weak annotations can be query string or meta data *etc*. As mentioned in Section 4.1 of the main paper, the weak annotations used in this paper are the AU predictions of the pre-trained classifiers on BP4D.

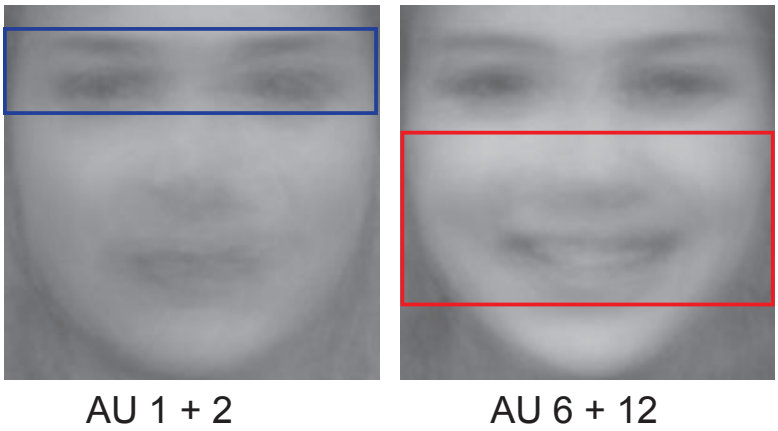
Scalable weakly supervised clustering: Fig. 1(a) illustrates the objective value of scalable weakly spectral clustering (sWSC) optimized by the accelerated stochastic gradient method, as discussed in Section 3.2 in the main paper. The figure was obtained by running sWSC on clustering ∼200,000 images of EmotioNet into two clusters using weak annotations of AU12. Although sWSC is stochastic-based, we can observe that convergence happened at around #5000 iterations. Fig. 1(b) shows the average images of two different clusters. As can be seen, each of these two clusters tends to group images from different modes of facial actions. On images of the cluster denoted as "AU 1+2", we observed more frequent AUs 1 and 2. The other cluster, denoted as "AU 6+12", exhibits strong lips corner puller and cheek raiser that directly corresponds to AU 6 and 12. This observation justifies that the proposed WSC can preserve both visually similar images while at the same time meaningful AU combinations, and meanwhile is scalable to large amount of images.

t-SNE visualization of WSC embeddings: To further understand the learned WSC embedding (denoted as $\mathbf{W}$ in the main paper), we examined the closeness of the images using t-SNE embedding. Fig. 2 shows the embedding and zoomed-in neighborhood to facilitate examination. As can be seen, samples with cheek raiser (AU6) and lips corner puller (AU12) bounded with red boxes are grouped in the lower area of the figure. Regardless of lighting conditions and head poses, most images within this neighborhood contain either subtle or obvious appearance of AU6 or AU12. On the other hand, Samples with eyebrow puller bounded with blue boxes are shown in the upper areas of the embedding space. This shows WSC embedding obtains meaningful representations for AU annotation.

Corrections of annotations: EmotioNet is collected from the online images with query strings. Using the weak annotations (*e.g*. queries or AUs obtained by pre-trained model) directly as the supervision easily leads to get the sub-optimal generalization by the inaccurate classifier. This scenario even happens for manually annotations. The proposed WSC can correct the possible inaccurate annotations for weak annotations and even for humans annotations. We show more qualitative results of using WSC to correct manually annotations here. Partial results were shown in Section 4.1 of the main paper. The left most column of Fig. 3 shows the canonical faces of the 6 AUs used in the paper. On the right of each row, we show images that WSC was able to correct their annotations. In other words, these images were originally annotated as inactive with respective to each AU. WSC discovered these patterns and demonstrated good consistency with FACS-coding, showing that noisy annotations can be further cleaned using the proposed WSC.



(a) Objective value



(b) Averaged image of different clusters

(c) Images in cluster (AU 1 + 2)

(d) Images in cluster (AU 6 + 12)

Figure 1. Visualization of averaged images of different clusters upon convergence. (a) shows the objective value v.s. the number of iterations of scalable weakly spectral clustering (sWSC). The experiments are conducted on ∼200,000 images of EmotioNet with weak annotations of AU12. (b) shows the averaged images of two different clusters. We denoted different averaged images as "AU1+2" and "AU6+12" to indicate WSC's power in grouping visually and semantically similar images. (c) and (d) are the images in these two clusters.

2. Derivation of accelerated stochastic optimization method

This section provides more details in deriving the accelerated stochastic algorithm as discussed in Section 3.2 of the main paper. Specifically, during the group-wise optimization for each $\mathbf{W}_g$, we write:

$$\min_{\mathbf{W}_g} \frac{1}{2}\|\mathbf{W}_g - \mathbf{V}_g\|_F^2 + \frac{\tilde{\lambda}}{n_g}\mathrm{Tr}(\mathbf{W}_g\mathbf{C}_g\mathbf{W}_g^\top) \tag{1}$$

Taking derivative of equation (1) as:

$$\mathbf{W}_g - \mathbf{V}_g + \frac{2\tilde{\lambda}}{n_g}\mathbf{C}_g\mathbf{W}_g = 0$$

$$(\mathbf{I}_{n_g} + \frac{2\tilde{\lambda}}{n_g}\mathbf{C}_g)\mathbf{W}_g = \mathbf{V}_g \tag{2}$$

For $\mathbf{C}_g$ is a centering matrix, which has the eigenvalue 1 of multiplicity $n-1$ and eigenvalue 0 of multiplicity 1. In addition, $\tilde{\lambda} > 0$. Thus, $\det(\mathbf{I}_{n_g} + \frac{2\tilde{\lambda}}{n_g}\mathbf{C}_g) \neq 0$. The optimal solution for (1) can be obtained as $\mathbf{W}_g^\star = (\mathbf{I}_{n_g} + \frac{2\tilde{\lambda}}{n_g}\mathbf{C}_g)^{-1}\mathbf{V}_g$.

Figure 2. t-SNE visualization of the learned WSC embedding on ~200,000 images of EmotioNet.

Figure 3. An illustration of samples that the proposed WSC is able to correct for human annotations in the EmotioNet dataset. Each row shows the corrected samples of different AUs. The left most columns are definitions of AUs and the samples in correspondence to FACS. The samples corrected in EmotioNet are shown in the right column.