Learning Facial Action Units from Web Images with Scalable Weakly-supervised Spectral Clustering

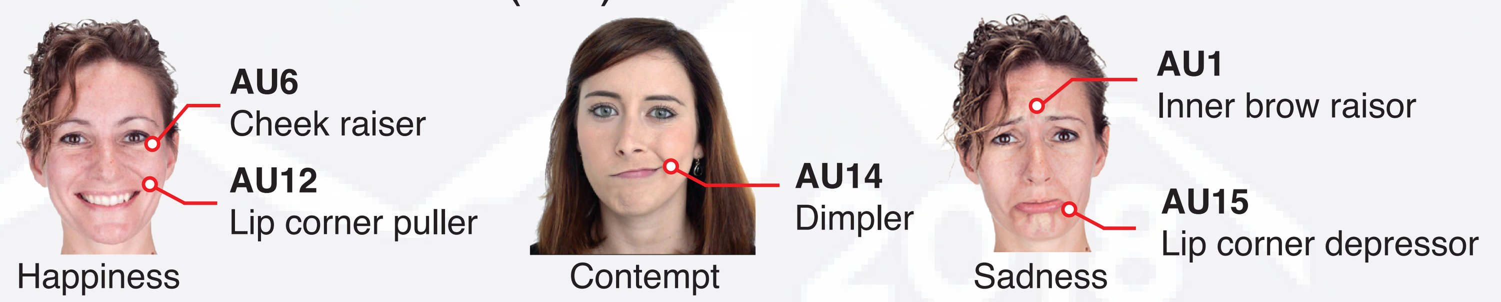
[1]Beijing University of Posts and Telecomm. [2]Carnegie Mellon University [3]The Ohio State University

[1]Kaili Zhao, [2]Wen-Sheng Chu, [3]Aleix M. Martinez

CVPR 2018 SALT LAKE CITY • JUNE 18-22

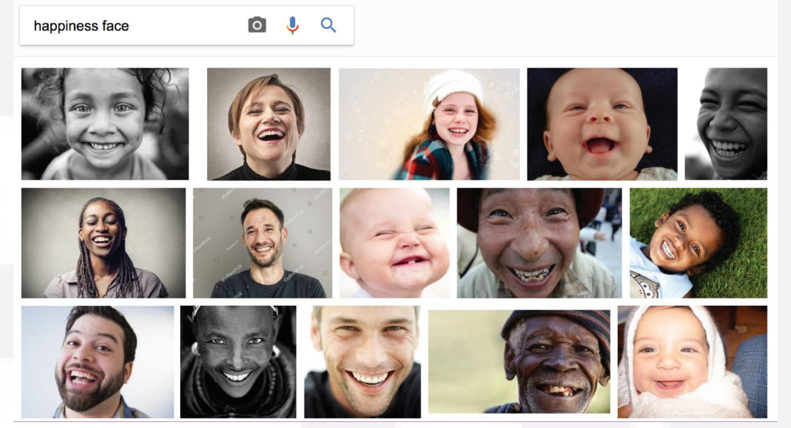
Problem

▲ Facial Action Unit (AU) detection



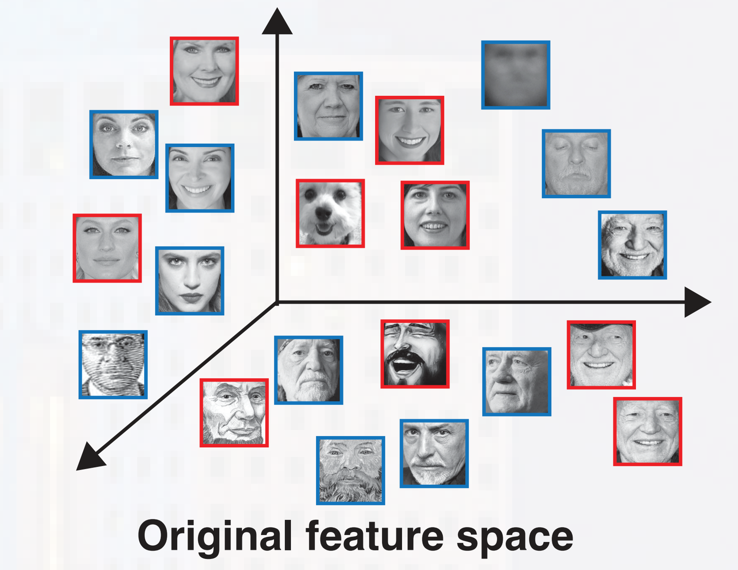
▲ Motivation

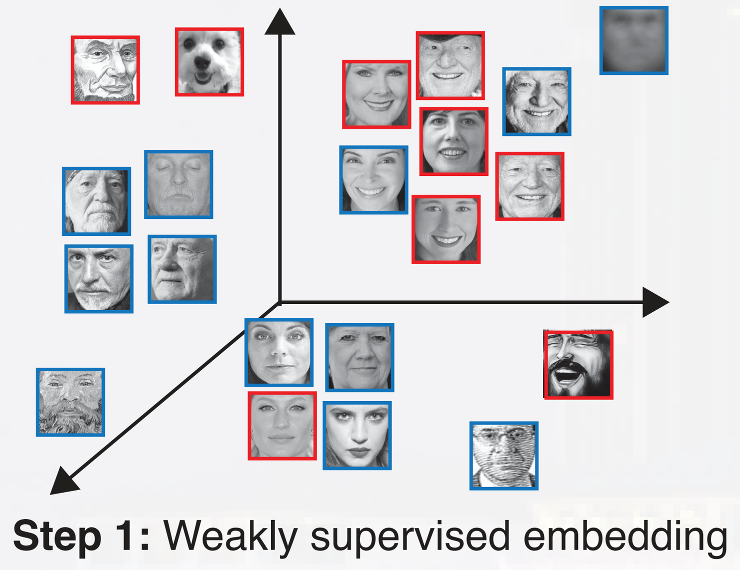
1. Utilize large and freely available web images
2. Avoid manual annotation laborious and error-prone
3. Improve model performance with free unannotated data

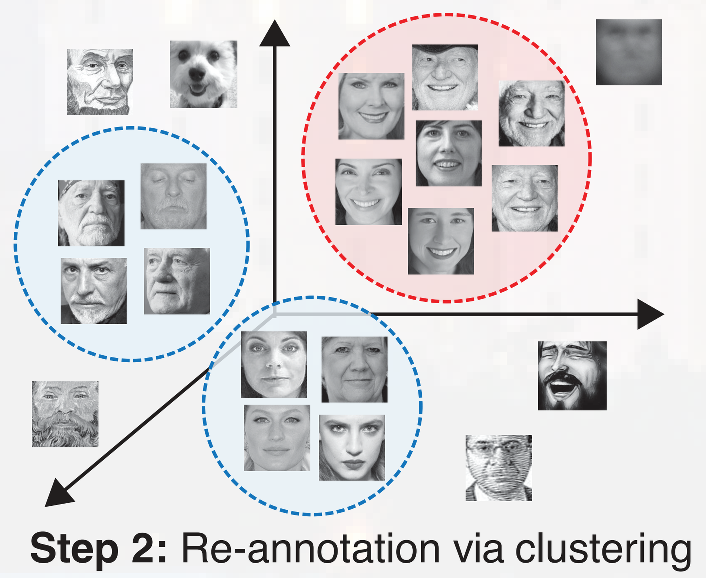


▲ Weakly-supervised Spectral Clustering (WSC)

Step 1. Weakly supervised embedding (WSE)
Step 2. Re-annotation via rank-order clustering







▲ Alternative methods

Methods	UD	PN	SL	IE
STM, CPM [1]	✓	✗	✗	✗
GFK, LapSVM [2]	✓	✗	✗	✓
Spectral/K-means clu.	✓	✗	✗	✓
WSC	✓	✓	✓	✓

UD: Unannotated data, **PN**: Pruning noises, **SL**: Scalability, **IE**: Identity exemption

[1] "Selective transfer machine for personalized facial action unit detection," in CVPR, 2013.
[2] "Geodesic flow kernel for unsupervised domain adaptation," in CVPR, 2012.

Scalable weakly-supervised spectral embedding

▲ Objective: Learn an embedding space with coherence among visual similarity and weak annotation in 1 million images

- **Formulation:** $$\min_{\mathbf{W}\in\mathbb{R}^{N\times K}} \underbrace{f(\mathbf{W},\mathbf{L})} + \frac{\lambda}{|\mathcal{G}|}\underbrace{\psi(\mathbf{W},\mathcal{G})} \quad \text{s.t.} \quad \mathbf{W}^\top\mathbf{W}=\mathbf{I}_K$$
- **Visual similarity:** (Spectral clustering) $$f(\mathbf{W},\mathbf{L})=\mathrm{Tr}(\mathbf{W}^\top\mathbf{L}\mathbf{W}),\ \mathbf{L}=\mathbf{D}-\mathbf{A},\ A_{ij}=\begin{cases}\exp(-\gamma d(\mathbf{x}_i,\mathbf{x}_j)), & \text{if } \mathbf{x}_i\in\mathcal{N}_k(\mathbf{x}_j)\\ 0 & \text{otherwise}\end{cases}$$
- **Weak annotation:** (Agreement) $$\psi(\mathbf{W},\mathcal{G})=\frac{1}{n_g}\sum_{\mathbf{w}_i\in\mathcal{G}_g}(\mathbf{w}_i-\overline{\mathbf{w}}_g)^\top(\mathbf{w}_i-\overline{\mathbf{w}}_g)=\frac{1}{n_g}\sum_{\mathbf{w}_i\in\mathcal{G}_g}\mathrm{Tr}(\mathbf{W}^\top\mathbf{C}_g\mathbf{W})$$

▲ Solution: Address the nonsmooth nature of $\psi(\mathbf{W},\mathcal{G})$ using first-order Taylor expansion and group decomposition

- **Analytical solution:** $\mathbf{W}_g^\star=(\mathbf{I}_{n_g}+\frac{2\tilde{\lambda}}{n_g}\mathbf{C}_g)^{-1}\mathbf{V}_g$ ⟵ Inverse is slow and numerically unstable!
- **10x-Faster solution:** $$\mathbf{W}_i=\frac{1}{a}\mathbf{V}_i-\frac{b}{a(a+bn_g)}\sum_j\mathbf{V}_j,\quad a=1+\frac{2\tilde{\lambda}}{n_g},\ b=\frac{2\tilde{\lambda}}{-n_g^2}$$

▲ Optimization: Accelerated gradient descent + stochastic extension

Algorithm 1 Weakly Supervised Spectral Embedding

Input: Laplacian matrix $\mathbf{L}\in\mathbb{R}^{N\times N}$, orthonormal matrix $\mathbf{W}_0\in\mathbb{R}^{N\times K}$, stepsize η, update ratio γ, and tuning parameter λ
Output: An orthonormal matrix $\mathbf{W}\in\mathbb{R}^{N\times K}$

```
a_0 = 1, t = 0
while not converge do
    if f(W_t) + λψ(W_t, G) ≥ Q_L(W_t, V) then
        η = γη
    end if
    V = W_t − η(2LW_t)
    for G_g ∈ G do
        W_g = (I_{n_g} + (2λ/n_g) C_g)^{-1} V_g // Update each group of W
    end for
    a_t = 2/(t+3)
    W_t = W_t + ((1−a_{t−1})/a_{t−1}) · a_t(W_t − W_{t−1})
    W_t = orth(W_t) // Enforce W_t to be orthonormal
end while
W = W_t
```

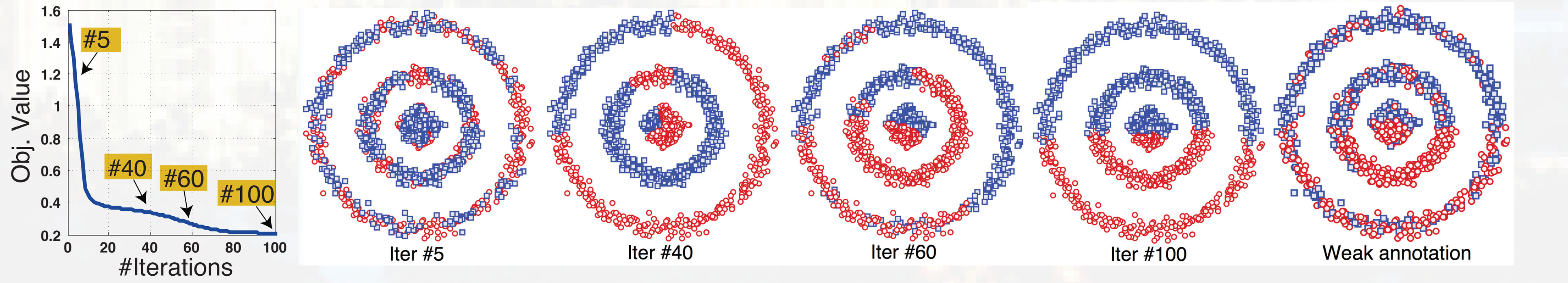
Algorithm 2 Stochastic Spectral Embedding

Input: Laplacian matrix $\mathbf{L}\in\mathbb{R}^{N\times N}$, orthonormal matrix $\mathbf{W}_0\in\mathbb{R}^{N\times K}$, number of batches B, number of iterations T, stepsize η, update ratio γ, and tuning parameter λ
Output: An orthonormal matrix $\mathbf{W}\in\mathbb{R}^{N\times K}$

```
while t ≤ T do
    for b = 1, ..., B do
        L̃_t = sampling(L) // Perform edge sampling
        Solve W_t using Algorithm 1 with (L̃_t, W_{t−1}, η, γ, λ)
        W_t = orth(W_t) // Enforce W_t to be orthonormal
    end for
end while
W = W_t
```


[3] "Accelerated gradient method for multi-task sparse learning problem," in ICDM, 2009.
[4] "Spectral clustering with a convex regularizer on millions of images," in ECCV, 2014.

▲ Toy EX: Group neighboring samples with high weak annotation agreement



Re-annotation via clustering

▲ Re-annotation pipeline

Build a graph on WSE → Hierarchical clustering via BFS → Quality index uMQI → Majority voting

▲ Build a graph based on WSE embedding

- **Rank-order distance** [5]: Same-class samples have similar neighbors
- Standard L1/L2 distance is sensitive to biased distribution of AU data

▲ Hierarchical clustering

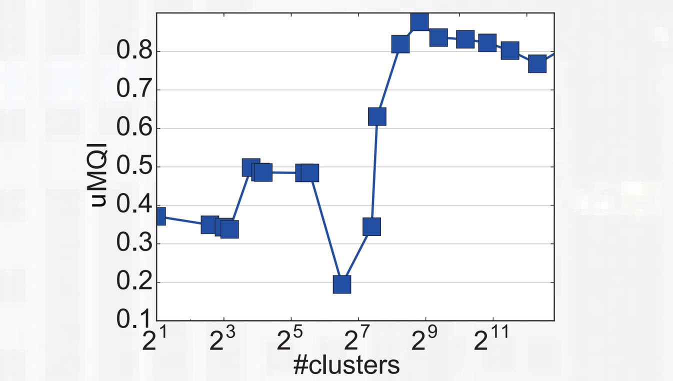
- **Scalable to 1M images**: $\mathcal{O}(Nk)+\mathcal{O}(N)$ vs k-means $\mathcal{O}(tKNd)$
- **Noise/outlier pruning** by identifying clusters with rare samples
- Avoid the non-trivial **#cluster** as input

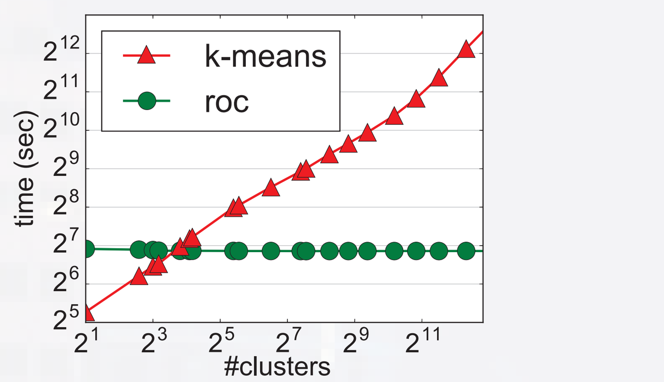
▲ Undirected modularization quality index (uMQI)

- **Intra-cluster connectivity** vs **inter-cluster isolation**

$$\mathrm{uMQI}(\mathcal{C})=\underbrace{\mathbb{E}[\mathrm{intra}(\mathcal{C})]}-\underbrace{\mathbb{E}[\mathrm{inter}(\mathcal{C})]},\ \mathcal{C}=\{\mathcal{C}_i,\forall|\mathcal{C}_i|>\delta\}_{i=1}^K$$

$$\mathrm{intra}(\mathcal{C})=\frac{1}{K}\sum_i\frac{m_i}{n_i^2-n_i}\qquad \mathrm{inter}(\mathcal{C})=\frac{1}{K(K-1)/2}\sum_{ij}\frac{m_{ij}}{2n_in_j}$$





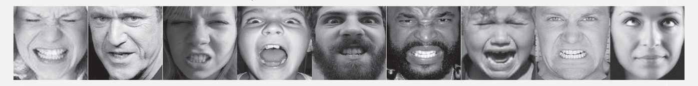
[5] "Clustering millions of faces by identity," TPAMI, 2017.

- **Experiment** ⑤: WSC by design can naturally prune nosie/outlier samples.

Findings:
- It took 2 min on a Intel i7-CPU machine to get WSC for 200K images
- WSC is able to rectify incorrent weak annotations by pre-trained classifiers.

AU12: wlb=−1, wsc=+1

AU12: wlb=+1, wsc=−1



Experiments

▲ EmotioNet dataset [6]

- 1M web images: 50K images (5%) were manually labeled by experts.
- 7 AUs with base rate > 5% were chosen for experiments.

▲ Settings

- 25K/25K partition of labeled images for training/test (following [6])
- Weak annotations were obtained by an AlexNet pre-trained on BP4D.
- The remaining 950K demonstrates the use of unlabeled images.

▲ Comparisons

- **Annotations:**
  wlb: weak annotation
  wsc: our annotation
  gt: human annotation
- **Experiment** ①: {gt25k} + {[**wlb** | **wsc** | **gt**]10k}
  Findings:
  - wsc >> wlb (> 20%)
  - wsc ~= gt (~4%)
- **Experiment** ②: AlexNet vs DRML [7]
  Findings:
  - DRML >~ AlexNet
  - More unlabeled data plus WSC helps both
- **Experiment** ③: SSL methods
  Findings:
  - SSL suffers from noisy data due to smoothness
  - LapSVM is slow and fails to scale up
  - WSC is efficient and best performer
- **Experiment** ④: Large-scale evaluation
  Findings:
  - WSC scales to 1M!
  - More improvement with more WSC-annotated images

AU	① AlexNet {gt15k, wlb10k}	① AlexNet {gt15k, wsc10k}	① AlexNet {gt25k}	② DRML {gt25k}	② AlexNet {gt25k, wsc25k}	② DRML {gt25k, wsc25k}
1	11.8	19.8	24.2	**25.3**	**25.3**	**[26.3]**
4	23.9	32.5	34.7	**[35.7]**	34.5	**35.5**
5	26.6	37.6	39.5	**40.0**	39.3	**[40.3]**
6	58.8	73.5	73.1	75.3	**75.6**	**[78.7]**
12	82.1	87.1	86.8	86.6	**87.4**	**[88.1]**
25	82.1	84.3	88.5	**[88.9]**	**88.8**	**[88.9]**
26	24.3	40.2	45.6	46.2	**47.7**	**[49.1]**
Avg.	44.2	53.6	56.1	56.9	**57.0**	**[58.1]**

③ AU	F1 GFK {gt25k}	F1 LapSVM {gt25k, ulb10k}	F1 TSVM {gt25k, ulb25k}	S score GFK {gt25k}	S score LapSVM {gt25k, ulb10k}	S score TSVM {gt25k, ulb25k}
1	19.3	1.2	24.1	66.1	82.3	70.2
4	31.0	25.7	32.3	61.1	85.3	62.5
5	31.8	23.1	40.3	61.1	60.7	80.6
6	73.8	58.3	75.7	71.7	70.0	79.1
12	85.1	57.7	87.4	75.5	50.9	80.2
25	85.8	88.9	88.2	72.4	79.4	78.5
26	39.0	5.0	47.0	69.5	83.2	78.4
Avg.	52.2	37.0	56.4	68.2	73.1	75.6

④ AU	20k wlb	20k wsc	200k wlb	200k wsc	400k wlb	400k wsc	1M wlb	1M wsc
1	17.6	18.3	17.8	19.3	16.9	**[21.3]**	17.6	**21.2**
4	20.3	20.5	19.0	20.4	18.9	**21.3**	18.4	**[22.1]**
5	28.5	28.9	30.1	30.8	31.5	**33.4**	30.8	**[41.6]**
6	72.4	74.1	75.9	76.9	76.3	**78.6**	77.4	**[79.3]**
12	76.7	85.8	79.1	86.4	79.3	**87.8**	81.4	**[88.2]**
25	84.7	85.7	85.4	85.9	79.4	**86.1**	86.1	**[89.1]**
26	32.4	34.9	32.7	36.0	33.3	**36.1**	33.3	**[47.2]**
Avg.	47.5	49.7	48.5	50.8	47.9	**52.1**	49.3	**[55.5]**

[6] "EmotioNet challenge: Recognition of facial expressions of emotion in the wild," in CVPRW, 2017.
[7] "Deep region and multi-label learning for facial action unit detection," in CVPR, 2016.