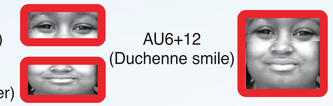
Deep Region and Multi-label Learning for Facial Action Unit Detection

Kaili Zhao[1] Wen-Sheng Chu[2] Honggang Zhang[1]

[1]Beijing University of Posts and Telecommunications [2]Carnegie Mellon University

Problem

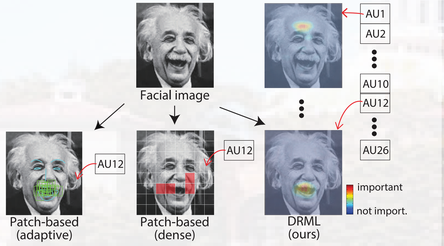
▲ Facial Action Unit (AU) detection



▲ Multi-label learning

- AUs have strong correlations: Likely or unlikely co-occur.
- Different from multi-class classification in holistic expression recognition, AU detection is a multi-label classification problem.

▲ Region learning v.s. patch learning



▲ Contributions

- A unified framework: Address region learning and multi-label learning using one convolutional network.
- A new region layer: An alternative layer between a standard convolutional layer [1] and a locally connected layer [2].

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep CNNs. In NIPS, 2012.

[2] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In CVPR, 2014.

DRML

▲ Input & output

- Multi-label loss (cross-entropy):

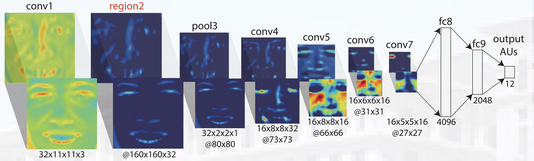
$$L(\mathbf{Y},\widehat{\mathbf{Y}}) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C}\{[\mathbf{Y}_{nc}>0]\log\widehat{\mathbf{Y}}_{nc} + [\mathbf{Y}_{nc}<0]\log(1-\widehat{\mathbf{Y}}_{nc})\}$$

- Input:

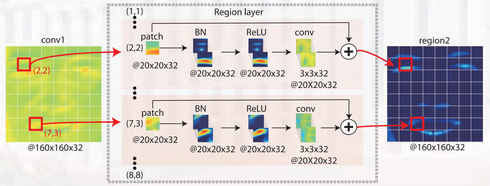
$\mathbf{Y}_{nc} \in \{-1, 0, 1\}$

@170x170x3

▲ Network architecture



▲ Region layer



- Patch clipping: Slice a 160x160 response map into an 8x8 uniform grid.
- Local convolution: Force the learned weights to be *updated independently* in each patch, to capture local appearance changes.
- Identity addition: Avoid vanishing gradients by adding *skip connections*.

▲ Comparisons to the state-of-the-art

Methods	AU relation	Feature representation	PL & ML optimization
JPML [3]	Statistically derived	Manually-crafted	Alternative
DRML (ours)	Learned	Learned	Joint

[3] K. Zhao et al. Joint patch and multi-label learning for facial action unit detection. In CVPR, 2015.

Alternative Methods

▲ Alternative CNNs

- AlexNet [1]: An AlexNet trained with multi-label loss.
- Locally Connected Network (LCN) [2]: ConvNet architecture with conv5-conv7 replaced with locally connected layers.
- Regular ConvNet: DRML without the region layer.
- DRML: The proposed architecture.

▲ Comparisons

Methods	End-to-end trainable	Multi-label learning	Learning represent.	Non-linearity	Online update
APL [4]	✗	✗	✗	✗	✗
AUDN [5]	✗	✗	✓	✓	✗
BDBN [6]	✓	✗	✓	✓	✓
JPML [3]	✗	✓	✗	✗	✗
DRML	✓	✓	✓	✓	✓

▲ Practical performance

- ConvNet/AlexNet shares kernels across an entire image.
- LCN confines kernels to individual pixel locations.
- DRML serves as an alternative between the ConvNet/AlexNet and LCN.
- Number of parameters (M: million)

	ConvNet	AlexNet	DRML	LCN
#Params	≤56M	≤60M	≤56M	≤120M

- Running time (ms)

	ConvNet	AlexNet	DRML	LCN
Train	9.7 ± 0.003	5.8 ± 0.005	31.2± 0.001	74.7± 0.006
Test	3.3 ± 0.002	2.2 ± 0.002	12.1± 0.001	34.7± 0.008

* All experiments were performed on one NVIDIA Tesla K40c GPU and Caffe toolbox [7] with modifications to support the region layer and multi-label loss.

[4] L. Zhong et al. Learning multiscale active facial patches for expression analysis. IEEE Transactions on Cybernetics, (99), 2014.

[5] M. Liu, S. Li, S. Shan, and X. Chen. AU-aware deep networks for facial expression recognition. In AFGR, 2013.

[6] P. Liu, S. Han, Z. Meng, and Y. Tong. Facial expression recognition via a boosted DBN. In CVPR, 2014.

[7] Y. Jia et al. Caffe: Convolutional architecture for fast feature embedding. In ACMMM, 2014.

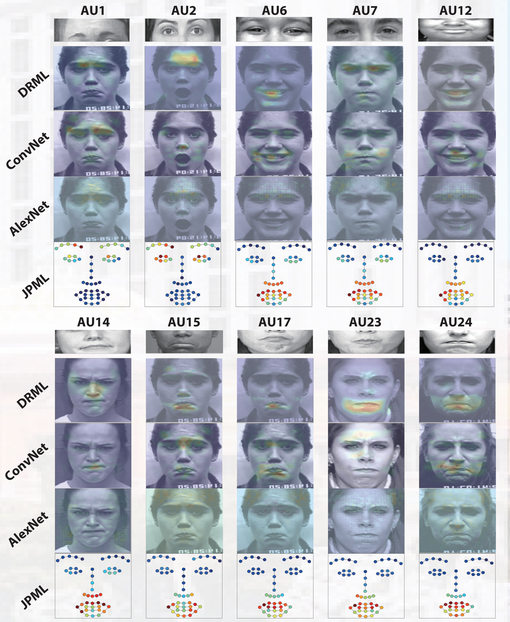
Regions v.s. Patches

▲ Importance between regions and patches

- Active region: Magnitude of per-pixel gradients with respect to a particular AU, ie, saliency map [8].
- Active patch: L2 norm for per-AU classifier, ie, JPML [3] and APL [4].

▲ Learned regions and patches

- The face images were selected manually from the CK+ dataset [9].



[8] K. Simonyan et al. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv:1312.6034, 2013.

[9] Lucey et al. The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In CVPRW, 2010.

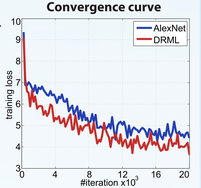
Experiments

▲ Settings

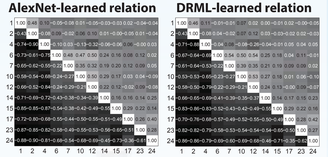
- Prediction: fc9 feature + Linear SVM.
- Metrics: F1 score and AUC.

▲ Scenarios

- Within-dataset: 3-fold partition for train/val/test sets.
- Cross-dataset: BP4D --> DISFA.



▲ Convergence of DRML



▲ BP4D [10]: 328 videos from 41 participants

AU	F1-frame LSVM	JPML	AlexNet	ConvNet	LCN	DRML	AUC LSVM	JPML	AlexNet	ConvNet	LCN	DRML
1	23.2	32.6	27.0	**40.4**	[45.0]	36.4	20.7	40.7	34.9	49.4	**51.9**	[55.7]
2	22.8	25.6	25.5	[46.1]	41.2	**41.8**	17.7	42.1	25.8	[51.3]	50.9	[54.5]
4	23.1	37.4	31.9	**42.8**	42.3	[43.0]	22.9	46.2	36.1	47.4	**53.6**	[58.8]
6	27.2	42.3	51.4	51.8	[58.6]	**55.0**	20.3	40.0	48.3	52.2	**53.2**	[56.6]
7	47.1	50.5	**55.4**	54.3	52.8	[67.0]	44.8	50.0	54.3	[64.8]	**63.7**	61.0
10	[77.2]	**72.2**	52.8	54.0	54.0	66.3	**73.4**	[75.2]	54.3	61.4	62.4	53.6
12	63.7	[74.1]	49.0	61.0	54.7	**65.8**	55.3	**60.5**	50.0	60.2	[61.6]	60.8
14	**64.3**	[65.7]	51.7	56.7	59.9	54.1	46.8	**53.6**	47.7	29.8	[58.5]	57.0
15	18.4	38.1	25.5	[44.1]	**36.1**	33.2	18.3	50.1	34.9	**50.6**	49.9	[56.2]
17	33.0	40.0	41.4	38.3	**46.6**	[48.0]	36.4	42.5	48.5	[53.5]	48.4	**50.0**
23	19.4	30.4	26.1	[41.8]	**33.2**	31.7	19.2	**51.9**	40.5	49.5	50.3	[53.9]
24	20.7	[42.3]	23.5	32.8	**35.3**	30.0	11.7	**53.2**	31.7	52.5	47.7	[53.9]
Avg.	35.3	45.9	38.4	**47.0**	46.6	[48.3]	32.2	50.5	42.2	51.8	**54.4**	[56.0]

▲ DISFA [11]: 27 videos from 27 participants

AU	F1-frame LSVM	APL	AlexNet	ConvNet	LCN	DRML	AUC LSVM	APL	AlexNet	ConvNet	LCN	DRML
1	10.8	11.4	12.0	11.7	**12.8**	[17.3]	21.6	32.7	**47.8**	44.2	44.1	[53.3]
2	10.0	12.0	11.6	12.0	**12.0**	[17.7]	15.8	27.8	52.1	37.3	**52.4**	[53.2]
4	21.8	**30.1**	27.6	28.9	29.7	[37.4]	17.2	37.9	44.0	**47.9**	47.7	[60.0]
6	15.7	12.4	22.6	21.4	**23.1**	[29.0]	8.7	13.6	**44.3**	38.5	39.7	[54.9]
9	11.5	10.1	11.5	11.5	[12.4]	10.7	15.0	[64.4]	48.7	49.5	40.2	**51.5**
12	[70.4]	**65.9**	31.1	31.0	26.4	37.7	**93.8**	[94.2]	55.3	54.8	54.7	54.6
25	12.0	21.4	**44.4**	40.7	[46.2]	38.5	3.4	[50.4]	**50.2**	48.4	48.6	45.6
26	22.1	26.9	**28.2**	27.7	[30.0]	20.1	20.1	[47.1]	45.8	45.8	**47.0**	45.3
Avg.	21.8	23.8	23.6	23.1	**24.0**	[26.7]	27.5	46.0	**49.1**	45.8	46.8	[52.3]

[10] Zhang et al. A high-resolution spontaneous 3d dynamic facial expression database. In AFGR, 2013.

[11] S. M. Mavadati et al. Disfa: A spontaneous facial action intensity database TAFFC, 4:151–160, 2013.