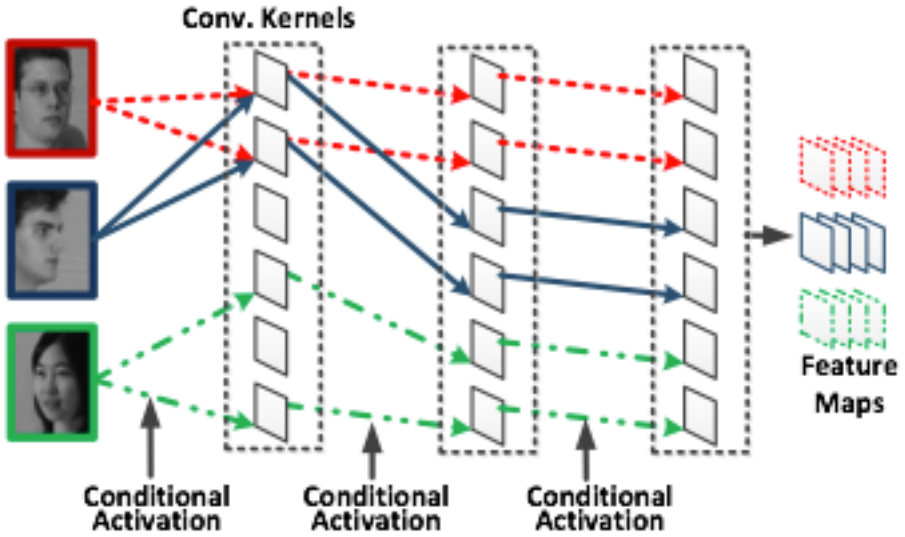
MULTIMODAL FACE RECOGNITION WITH CONDITIONAL CNN



By Erol Kazancli
based on a paper by Chao Xiong, Xiaowei Zhao, Danhang Tang, Karlekar Jayashree, Shuicheng Yan, and Tae-Kyun Kim, 2015

MULTIMODALITY
IN FACE RECOGNITION

- Pose differences
- Occlusion
- Illumination change
- Expression

DIFFICULTIES IN MODALITY

- Need for a shared feature space to compare different modalities
- Manually designed features insensitive to modality changes
- Unsuitability of the generic features like SIFT, HOG and LBP
- Information lost in the extraction stage
- Difficulty to find a linearly separable feature space

SOLUTION:
CONVENTIONAL CNN?

- With conventional CNN modality information is needed during training, which is difficult to achieve.
- Even if we have the ground truth, modalities are vague and difficult to define.

PROPOSED SOLUTION: CONDITIONAL CNN

- No prior knowledge on modality
- Automatic learning of the inherent modality distribution
- Conditional computation of different routes for different modalities
- Activation of certain kernels depending on the modality

CONDITIONAL CNN

- Deciding the route through activation and disactivation of kernels with MODALITY AWARE PROJECTION TREE
- Calculating the activation levels along the route with CONVOLUTIONAL NEURAL NETWORK

TRAINING
MODALITY AWARE PROJECTION TREE

Splitting of segments is learned in an unsupervised manner with regard to modality:

- Split function:

$$\boldsymbol{x} = \begin{cases} \mathcal{S}^{\mathcal{L}}, & \varphi(\boldsymbol{x}) \geq 0 \\ \mathcal{S}^{\mathcal{R}}, & \varphi(\boldsymbol{x}) < 0 \end{cases}$$

$$\varphi(\boldsymbol{x}) = \boldsymbol{x}^T \cdot \boldsymbol{P}^{(i,j)} + \tau^{(i,j)}.$$

- Loss function:

$$\mathcal{L} = \frac{\frac{1}{N} \sum_{\boldsymbol{x} \in \mathcal{S}} \varphi(\boldsymbol{x})^2}{(\frac{1}{N_{\mathcal{L}}} \sum_{\boldsymbol{x} \in \mathcal{S}^{\mathcal{L}}} \varphi(\boldsymbol{x}) - \frac{1}{N_{\mathcal{R}}} \sum_{\boldsymbol{x} \in \mathcal{S}^{\mathcal{R}}} \varphi(\boldsymbol{x}))^2},$$

TRAINING
CONVOLUTIONAL NEURAL NETWORK

Internal representation of CNN is learned:

- Forward function:

$$\widetilde{\boldsymbol{X}_n^{(i,j)}} = \sigma(\boldsymbol{W}^{(i,j)} * \boldsymbol{X}_n^{(i,j)} + b^{(i,j)}),$$

- Conditional forward function:

$$\begin{cases} \boldsymbol{X}_n^{(i+1,2j)} & = \mathbb{1}(\varphi(\widetilde{\boldsymbol{X}}_n^{(i,j)}) \geq 0) \cdot \widetilde{\boldsymbol{X}}_n^{(i,j)} \\ \boldsymbol{X}_n^{(i+1,2j+1)} & = \mathbb{1}(\varphi(\widetilde{\boldsymbol{X}}_n^{(i,j)}) < 0) \cdot \widetilde{\boldsymbol{X}}_n^{(i,j)} \end{cases},$$

TRAINING
JOINT LEARNING

Joint learning is achieved with a unified loss function and back propagation with SGD:

- Loss function:

$$\mathcal{L} = \sum_{n} \mathcal{J}(\boldsymbol{x}_n, y_n) + \beta \sum_{i} \sum_{j} \mathcal{L}^{(i,j)},$$

where the first term is softmax loss and the second term is node-wise loss.

TESTING

The activation of the kernels is jointly determined by the input representation in that layer and the activated route till that layer:

$$\boldsymbol{X}_{n,k}^{(i+1)} = g_{n,k}^{(i)} \cdot \sigma(\widetilde{\boldsymbol{W}}_k^{(i)} * \boldsymbol{X}_n^{(i)} + b^{(i)})$$

where x is the intermediate representation of the input and g is the activation indicator which follows a Bernouilli distribution with

$$p_{n,k}^{(i)} = Pr(\theta_{n,k}^{(i)} | \boldsymbol{X}_n^{(i)}, \boldsymbol{\theta}_n^{(i-1)}, ..., \boldsymbol{\theta}_n^{(0)})$$

RESULTS
WITH DIFFERENT POSES

	Avg.	±90°	±75°	±60°	±45°	±30°	±15°	pose
Fisher Vector [24]	66.60	24.53	45.51	68.71	80.33	87.21	93.30	×
FIP_20 [31]	67.87	34.13	47.32	61.64	78.89	89.23	95.88	√
FIP_40 [31]	70.90	31.37	49.10	69.75	85.54	92.98	96.30	√
CNN_40	70.81	32.08	47.79	69.48	85.99	93.04	96.60	×
Cluster_CNN	69.87	36.80	47.36	68.20	82.43	90.67	93.75	×
Tree_CNN	71.16	39.90	50.29	67.21	83.63	91.31	94.66	×
c-CNN	73.54	41.71	55.64	70.49	85.09	92.66	95.64	×
c-CNN Forest	76.89	47.26	60.66	74.38	89.02	94.05	96.97	×

RESULTS
WITH DIFFERENT OCCLUSIONS

	1	2	3	4	5	6	7	8	9	10	Avg.
HDLBP [2]	69.77	68.79	66.39	69.09	67.45	66.89	67.70	67.26	66.71	69.85	67.99
Fisher Vector [24]	70.83	72.90	73.21	72.83	71.80	73.44	73.33	72.29	72.96	73.29	72.68
PEM [16]	62.87	65.08	65.44	63.17	62.70	65.50	63.08	61.58	64.46	63.81	63.76
CNN_20	74.40	73.12	71.69	72.94	71.38	74.65	72.63	74.63	71.27	72.40	72.91
CNN_40	75.40	73.83	74.12	73.30	72.74	76.20	72.36	76.20	71.43	73.50	73.90
c-CNN	77.63	75.09	75.00	75.03	73.69	76.55	76.16	76.85	74.80	74.43	75.52
c-CNN Forest	77.65	75.16	75.00	76.17	73.71	77.67	77.27	77.81	76.10	75.83	76.24

Table 2. Comparisons of precision (%) with some prior methods on occluded LFW for ten folds.

ADVANTAGES

- No need for modality information during training
- Modality variation handled efficiently with automatic learning of modality and activation-disactivation of kernels
- No need for specific modality invariant features
- Joint training of the conditional and the convolutional part with a unified loss function

DISADVANTAGES

- Training time may be slightly longer
- Training data needs to be representative enough of the modality variation
- More parameters to define

FUTURE WORK

- Flexible assignments of convolution kernels in each layer
- Hand-crafted features instead of raw-pixel inputs