Disentangled Variational Representation for Heterogeneous Face Recognition

Xiang Wu, Huaibo Huang, Vishal M. Patel, Ran He, Zhenan Sun
[1] Center for Research on Intelligent Perception and Computing, CASIA
[2] National Laboratory of Pattern Recognition, CASIA
[3]School of Artificial Intelligence, University of Chinese Academy of Sciences
[4] Johns Hopkins University, 3400 N. Charles St, Baltimore, MD 21218, USA

Application

- Facial authentication on mobile devices
- Video surveillance

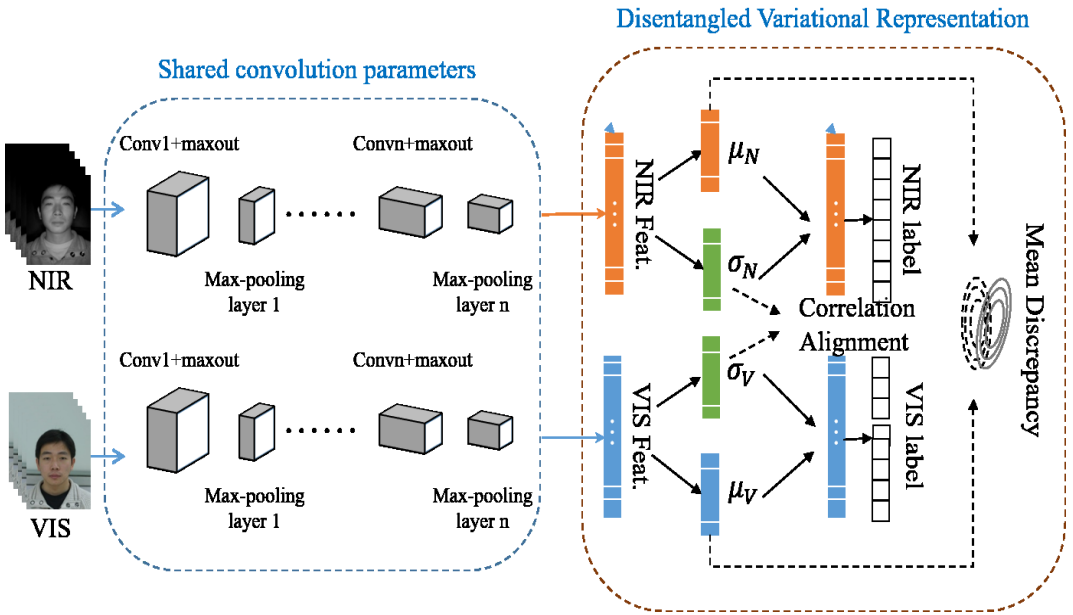
Challenges

- modality discrepancy.
- lack of training samples

Contributions

- An end-to-end Disentangled Variational Representation (DVR) is developed, aiming at disentangling the heterogeneous face representations.
- The identity information minimization and the relaxed correlation alignment constraint facilitate modeling the compact and discriminative disentangled latent variable spaces for heterogeneous modalities.
- An alternative optimization is proposed to provide mutual promotion between HFR network and disentangled variational representation part.
- DVR significantly improves the performance of HFR on three datasets.

Framework



An overview of the proposed DVR approach. The NIR and VIS representations x_N and x_V are disentangled into (μ_N, σ_N) and (μ_V, σ_V), respectively. We assume that there is a linear relationship, P, between lighting variations, i.e. $\sigma_V = P\sigma_N$. The mean discrepancy is used to measure the difference between NIR and VIS distributions in the latent space. The reconstructions $\hat{x}_N$ and $\hat{x}_V$ are obtained from the likelihood $p(x_N|z_N)$ and $p(x_V|z_V)$, respectively and are constrained by the cross-entropy loss.

Formulation

Disentangled Variational Representation

$$\mathcal{J}_{\text{DVR}} = \underbrace{-\frac{1}{2}\sum_j \left(1 + \log\sigma_{Nj}^{2(i)} - \mu_{Nj}^{2(i)} - \sigma_{Nj}^{2(i)}\right)}_{\text{NIR approximate posterior estimator}} \underbrace{-\frac{1}{2}\sum_j \left(1 + \log\sigma_{Vj}^{2(i)} - \mu_{Vj}^{2(i)} - \sigma_{Vj}^{2(i)}\right)}_{\text{VIS approximate posterior estimator}} \underbrace{+\mathbb{E}\left[\log p(x_N^{(i)}|z_N)\right] + \mathbb{E}\left[\log p(x_V^{(i)}|z_V)\right]}_{\text{reconstruction parts}}$$

→ *Variational part*

$$\underbrace{+\lambda_1\|\mu_N^{(i)} - \mu_V^{(i)}\|_2^2}_{\text{mean discrepancy part}} \underbrace{+\lambda_2\|\sigma_V - P\sigma_N\|_2^2 + \lambda_3\|P^\top P - I\|_F^2,}_{\text{correlation alignment constraint}}$$

→ *Constraints*

Heterogeneous Recognition Network

$$\mathcal{J}_{\text{cls}} = \underbrace{\text{softmax}(x_i, y; W, \Theta)}_{\text{Original features from true images}} + \underbrace{\text{softmax}(\hat{x}_i, y; W, \Theta)}_{\text{Synthesis features sampling from DVR}}, i \in \{N, V\}.$$

Performance