Unsupervised Disentangling of Viewpoint and Residues Variations by Substituting Representations for Robust Face Recognition





Minsu Kim Joanna Hong Junho Kim Hong Joo Lee Yong Man Ro*

Image and Video Systems Lab, School of Electrical Engineering, KAIST, South Korea

*Corresponding author

Introduction

- Face is generally captured along with diverse factors of variations such as identity, viewpoint, and illumination. These variations pose challenges in face recognition methods in having robust performance in a wild environment.

- In order to handle this challenge, several works [1, 2] have proposed disentangling methods that achieve robust performance in a wild environment by disentangling identity and non-identity variations (i.e., viewpoint and illumination)

- However, they need annotations of non-identity variations such as viewpoint and illumination. It is not easy to collect such pose or illumination information for all subjects in facial databases

[1] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning gan for pose-invariant face recognition," CVPR, 2017, pp. 1415–1424.
[2] X. Peng, X. Yu, K. Sohn, D. N. Metaxas, and M. Chandraker,"Reconstruction-based disentanglement for pose-invariant face recognition," ICCV, 2017, pp. 1623–1632.

Introduction

- In this paper, we propose a learning method of disentangling identity and viewpoint representations **without any auxiliary supervision of the variations**.
- Furthermore, we disentangle not only the identity and viewpoint but also residues (e.g., illumination and color variations) that inevitably exists in a face.
- By disentangling the non-identity variations from a face, we set a new state-of-the-art face recognition mehod on CFP and Multi-PIE datasets that have large pose variations.

Visual comparison of identity, viewpoint, and residues representations.
The three rows of images are synthesized by interpolating each representationfrom source to target image.
From top to bottom, identity, viewpoint, and residues are represented.

Proposed method

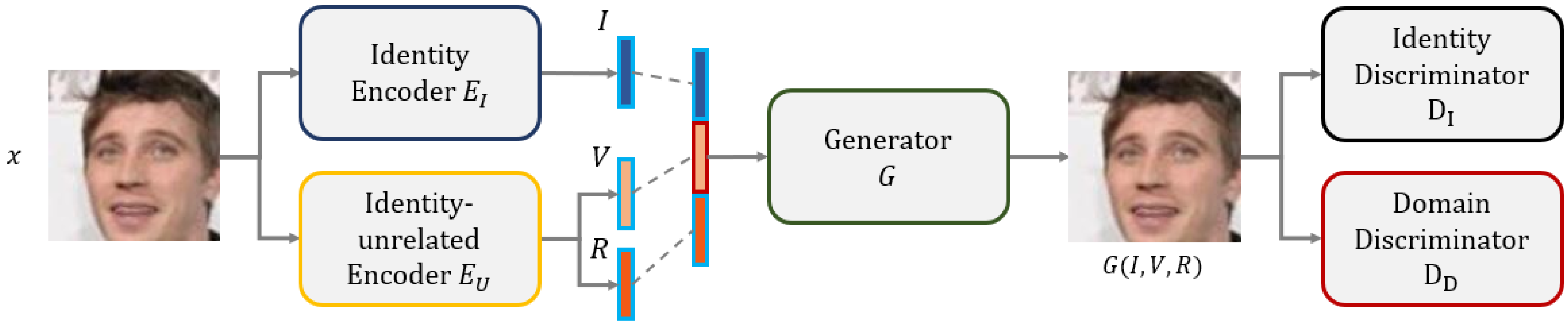
- We design the learning problem as **finding a generative function** which is conditioned on the three representations (i.e., identity, viewpoint, and residues) that **wield independent effects on the output**.

- To this end, we propose two learning schemes.
  - **Viewpoint substitution**
  - **Identity substitution**

- Also we suggest **a disentangling loss function using distance covariance**.

Proposed method

- **Overall framework**



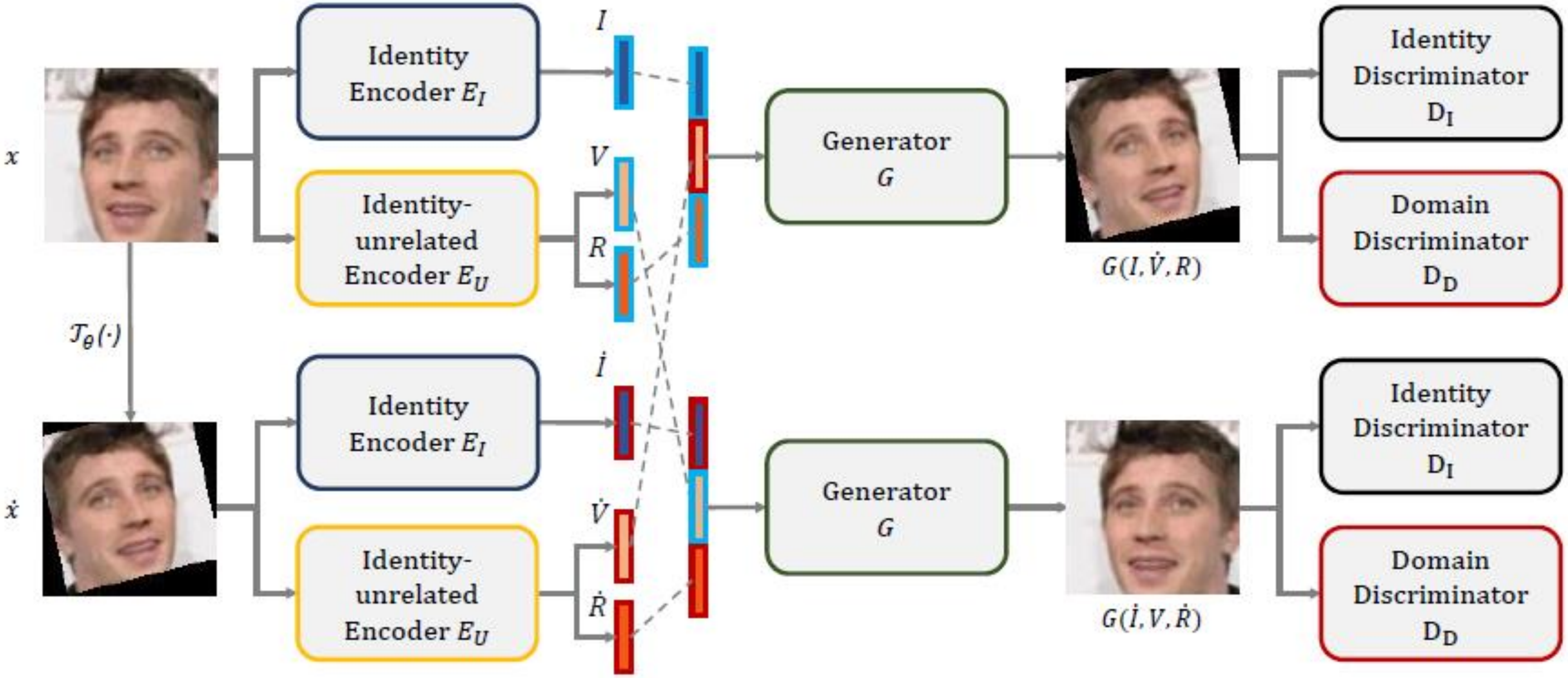
- Consists of 5 modules
  - ID encoder E_I
  - ID-unrelated encoder E_U
  - Generator G
  - ID-discriminator D_I
  - Domain-discriminator D_D

Proposed method

1. **Learning the viewpoint representation**
   - If we have the labels of viewpoint or the pairs of images that have the same identity but different viewpoint, learning the viewpoint representation of a face becomes easy.
   - However, it is very challenging when such information is absent.
   - To alleviate this challenge, we propose to use **a simple transformation** that changes the viewpoint while maintaining the identity of a face image.
   - The transformation can be affine, perspective, or thin plate spline transformation.
   - In this paper, we use affine transformation which is simple but effective.
   - By using the transformation, we can access the pair of images that contain different viewpoints.

Proposed method

1. Learning the viewpoint representation

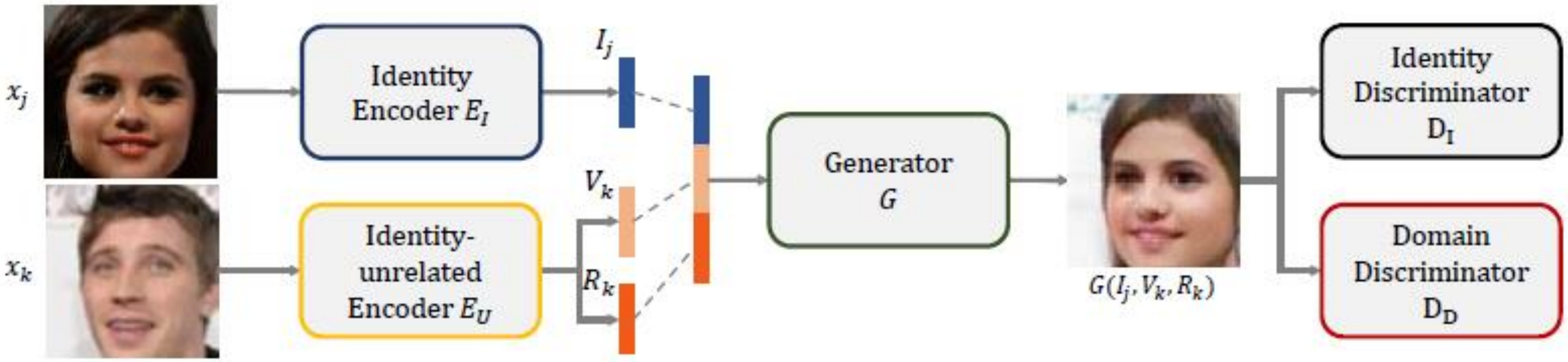


- The transformed image differs only in the viewpoint aspect from the input image.
- Viewpoint substitution loss : $L_v = \frac{1}{2}(\|x - G(\dot{I}, V, \dot{R})\|_2 + \|\dot{x} - G(I, \dot{V}, R)\|_2).$

Proposed method

2. Learning the Identity representation

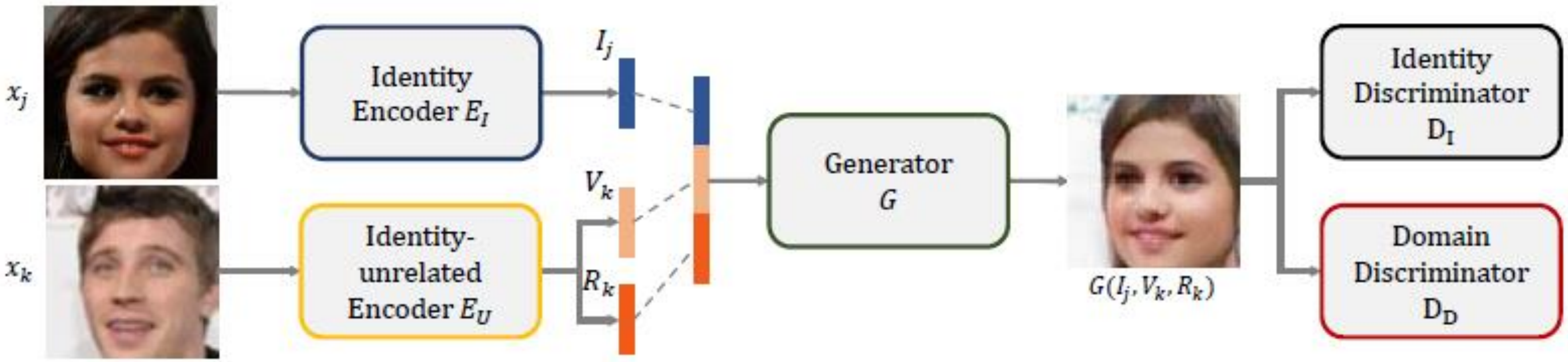
- To guide the identity encoder to work with the identity information, we bring the advantage of recent advances of face recognition. That is, ID-encoder E_I is trained with identification loss. Note that we don't train the E_I with the transformed images but the original images.
- To achieve the identity disentangled representation, we substitute identity representation from another.



- The generated image should contain the identity of the source identity image while other representations keep remaining.

Proposed method

2. Learning the Identity representation



- Identity substitution loss:

$$L_d = -\log D_I^{y_k}(x_k) - \log D_D(x_k) - \log(1 - D_D(G(I_j, U_k)))$$

$$L_g = -\log D_I^{y_j}(G(I_j, U_k)) - \log D_D(G(I_j, U_k))$$

$$L_r = \begin{cases} \|x_j - G(I_j, U_k)\|_2 , & \text{if } j = k \\ \alpha \|x_k - G(I_j, U_k)\|_2 , & \text{if } j \neq k \end{cases}$$

Proposed method

3. Disentangling loss

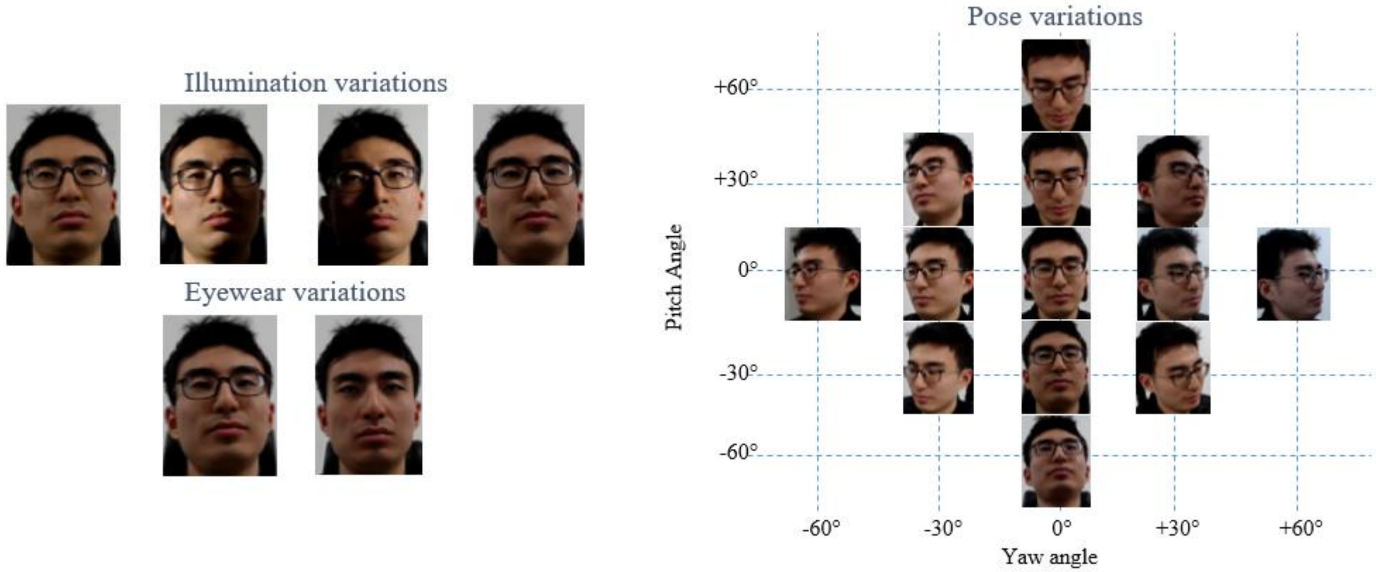
- The disentangled representations should contain different information from one another.

  To guarantee the independency between learned representations, we use distance covariance as a disentangling loss function.
- Distance covariance, $dCov_n^2(X, Y)$, is a metric that measures dependency between random vectors and becomes zero when the two random vectors are independent from each other.

- Disentangling loss: $L_c = \frac{1}{3}(dCov_n^2(I, V) + dCov_n^2(V, R) + dCov_n^2(R, I))$

Experiments

- We used ResNet-14 (7 blocks) for the identity encoder.
- For the implementation details, please refer to our paper.
- Datasets
  - Casia-WebFace, Multi-Pie for training
  - CFP, LFW, IJB-A, YTF, KAIST-MPMI for evaluation



Variations in KAIST-MPMI Data

Experiments

- Ablation Study

Identification loss	Method	LFW	CFP-FP	YTF
Softmax	*baseline*	98.02	91.60	91.58
	no cov/pose	**98.08**	91.53	92.16
	no cov	97.87	91.84	92.92
	full	97.95	**92.04**	**93.22**
CosFace [5]	*baseline*	98.28	91.86	92.14
	no cov/pose	98.83	92.83	93.06
	no cov	98.78	93.28	93.52
	full	**98.98**	**93.59**	**93.98**
ArcFace [4]	*baseline*	98.33	92.66	92.22
	no cov/pose	98.83	93.10	93.18
	no cov	98.90	93.59	93.36
	full	**99.03**	**94.03**	**94.02**

  - Verifying the effectiveness of each component of the proposed methods

Experiments

- Comparison on benchmark databases

VERIFICATION ACCURACY COMPARISON ON CFP.

Method	Frontal-Frontal	Frontal-Profile
Sengupta et al. [35]	96.40	84.91
Sankarana et al. [36]	96.93	89.17
Chen et al. [34]	**98.67**	91.97
DR-GAN [8]	97.84	93.41
Peng et al. [2]	**98.67**	93.76
Human	96.24	94.57
Ours	98.66	**94.03**

PERFORMANCE COMPARISON ON IJB-A.

Method	Verification Accuracy	
	@0.01 FPR	@0.001 FPR
Wang et al. [37]	72.9	51.0
PAM [38]	73.3	55.2
DCNN [39]	78.7	-
DR-GAN [8]	77.4	53.9
Ours	**81.0**	**64.4**

- Compared to the supervised disentangling methods [2,8], our proposed method shows better results especially in large pose environment.

Experiments

- Comparison on benchmark databases

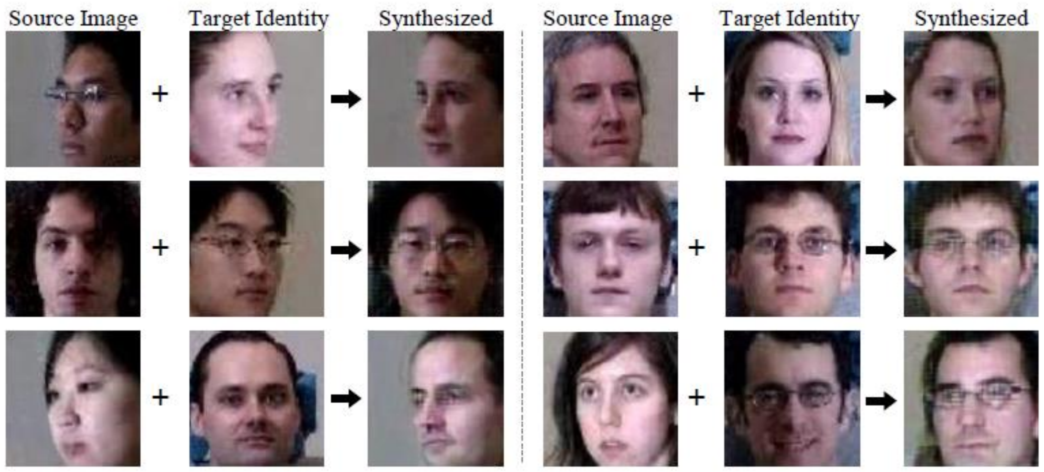
IDENTIFICATION ACCURACY COMPARISON ON MULTI-PIE.

Method	0°	15°	30°	45°	60°	Average
Zhu et al. [40]	94.3	90.7	80.7	64.1	45.9	72.9
Zhu et al. [13]	95.7	92.8	83.7	72.9	60.1	79.3
Yim et al. [41]	**99.5**	**95.0**	88.5	79.9	61.9	83.3
DR-GAN [8]	97.0	94.0	90.1	86.2	83.2	89.2
Ours	94.9	92.9	**92.4**	**88.8**	**85.5**	**90.5**

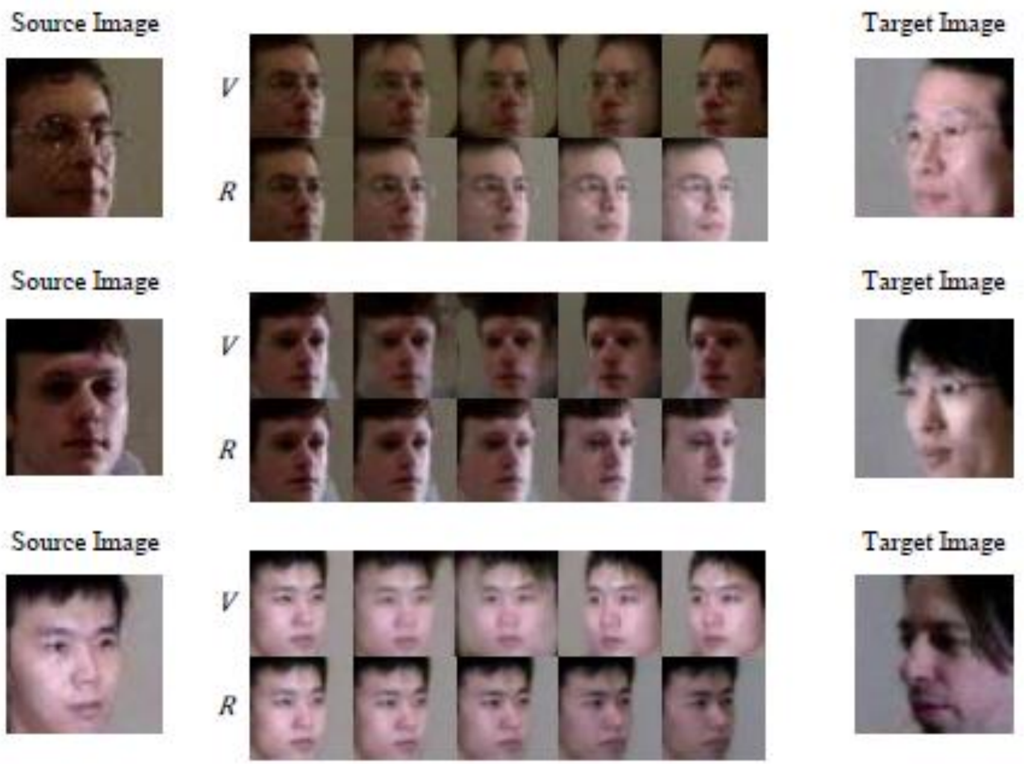
- Our proposed method shows better results especially in large pose environment.

Experiments

- Visualizing the disentangled representations



Synthesized images by substituting the identity representation of target identity image



Synthesized images by interpolating viewpoint and residues representation from source to target image.

Conclusion

- In this paper, we introduced a novel framework to learn disentangled representations for robust face recognition
- In particular, we propose two learning schemes, veiwpoint substitution and identity substitution.
- Also we show that using distance covariance as a disentangling loss enforces the disentanglement.

- By disentangling identity, viewpoint, and residues representation, we set a new state-of-the-art on benchmark databases in terms of face recognition performance.