VAE/WGAN-BASED IMAGE REPRESENTATION LEARNING FOR POSE-PRESERVING SEAMLESS IDENTITY REPLACEMENT IN FACIAL IMAGES

Hiroti Kawai, Jiawei Chen, Janusz Konrad, and Prakash Ishwar

Department of Electrical and Computer Engineering, Boston University, Boston, MA, USA

Motivation

- Smartroom of the future: could improve energy efficiency, health outcomes and productivity by recognizing activities of occupants
- Standard approach: video cameras
- Problem: privacy concerns
- Proposed solution: seamlessly replace occupant’s appearance while preserving other useful information like expression, pose, etc.

State of the art:

PPRL-VGAN framework to preserve WGAN

Contributions:

- Inception modules
- PPRL-VGAN framework to preserve WGAN
- WGAN + modified cost function (image reconstruction cost) to improve training stability and image quality

Proposed Methodology for Head Pose Estimation

1. PPRL-VGAN for headpose estimation:

   - Inception modules: contain 3 convolutional-layer branches with different filter sizes; branch outputs are concatenated
   - Improved training method:
     - Wasserstein GAN (WGAN): leverages Earth-Mover distance (instead of Jensen-Shannon divergence) via gradient penalty in discriminator loss
     - Image reconstruction cost: compares input image and generated image to improve image quality
     - Generator Loss: encourages synthesis of realistic images with new identity and original headpose

   \[ L_G = E\left[-D_1^G(G(x, c(y)))\right] + E\left[-\log D_2^D(G(x, c(y)))\right] \]
   \[ + \frac{1}{3} \sum_{i=1}^{3} E\left[-D_1^G(D_i^{G}(G(x,c(y))))\right] + E\left[-\log D_2^{D_i}(G(x,c(y)))\right] \]
   \[ + \frac{1}{2} \sum_{i=1}^{3} |D_i^{G}(G(x,c(y))||r(x))| \]

   - Discriminator Loss: encourages accurate prediction of identity, expression and real vs synthetic detection

   \[ L_D = E\left[-D_1^D(x) + D_2^D(G(x,c(y)))\right] + E\left[-\log D_2^{D_i}(x)\right] \]
   \[ + \frac{1}{3} \sum_{i=1}^{3} E\left[-(|\nabla D_i^D(x)||_2 - 1)^2\right] \]

   \[ WGAN \text{ cost} \quad \text{Identity cost} \quad \text{Head-pose cost} \]

   \[ \text{Image Reconstruction} \quad \text{Regularization} \quad \text{Gradient penalty} \]


Generator: based on Variational Autoencoder:

- encoder converts input image into a latent vector representation
- decoder synthesizes a new realistic-looking image with specified identity from a latent vector

Discriminator: 3 prediction objectives

- \( D_1 \): Is image real or fake?
- \( D_2 \): Identity
- \( D_3 \): Facial expression

Experimental Results

Quantitative evaluation: Privacy protection evaluated by training another neural network to predict identity under 3 attack scenarios:

<table>
<thead>
<tr>
<th>Attack Scenario</th>
<th>Identification (%)</th>
<th>Headpose MAE (°)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>Ours</td>
<td>PPRL-VGAN</td>
</tr>
<tr>
<td>Privacy Unconstrained</td>
<td>99.97</td>
<td>0.69</td>
</tr>
<tr>
<td>Training: Original Dataset Test: Synthesized Images</td>
<td>10.23</td>
<td>9.92</td>
</tr>
<tr>
<td>Training: Synthesized Images Test: Synthesized Images</td>
<td>23.11</td>
<td>21.64</td>
</tr>
<tr>
<td>Training: Latent Vectors Test: Latent Vectors</td>
<td>21.33</td>
<td>23.71</td>
</tr>
</tbody>
</table>

Identity/head-pose morphing: The generative ability of our model is evaluated by identity and head-pose morphing:

Conclusions

- Our method synthesizes realistic face images with a desired identity and improved image quality compared to a state-of-the-art method.
- We achieve performance competitive with a state-of-the-art method for learning an identity-invariant image representation.
- Our model can be applied to other image tasks such as pose or face morphing.