Pattern Recognition Letters 112 (2018) 91–97

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Pose transforming network: Learning to disentangle human posture in variational auto-encoded latent space

Jongin Lim[a], Youngjoon Yoo[b], Byeongho Heo[a], Jin Young Choi[a,*]

[a] *Department of Electrical and Computer Engineering, ASRI, Seoul National University, 133-413 599 Gwanak-ro Gwanak-gu, Seoul 151-742, Republic of Korea*
[b] *Graduate School of Convergence Science and Technology, Seoul National University, C206-4 145 Kwangkyo-ro Youngtong-gu Suwon-si, Kyungki-do 16229, Republic of Korea*

ARTICLE INFO

Article history:
Received 4 August 2017
Available online 18 June 2018

Keywords:
Human pose transform
Disentangle hidden factors in latent space
Generative model
Variational auto-encoder

ABSTRACT

This paper proposes a novel deep conditional generative model for human pose transforms. To generate the desired pose-transformed images from a single image, a variational inference model is formulated to disentangle human posture semantics from image identity (human personality, background etc.) in variational auto-encoded latent space. A deep learning architecture is then proposed to realize the formulated variational inference model. In addition, a new loss function for the proposed training method is designed to enable pose information and identity information to be separated completely in the latent space. The proposed model is validated experimentally by demonstrating its pose-transforming capability, outperforming the existing conditional generative model.

1. Introduction

Generating photo-realistic images that transform a human subjects posture is an interesting area of research. Given an image, as shown in Fig. 1, we can transform the posture within the same scene or generate a different person with the same posture. This can be used as to predict human motion [2,5], for re-identification [22], and for computer-aided photo editing in graphic design. However, it remains challenging to manipulate a person's posture in an image, as the controllable factors that govern the desired transformation must be separated from other entangled factors. Photographic image data are high-dimensional, and posture conveys high-level semantics in a high-dimensional space, making it difficult to clearly separate posture from a person's identity or from the background.

Image generation has been a long-standing problem for computer vision research. Many recent studies of deep generative models [4,6,7,9,13,17] have exploited the strengths of deep neural networks. In particular, Variational Auto-Encoders (VAE) [9,17] and Generative Adversarial Networks (GAN) [6] have made remarkable progress in generating natural images. Building on these generative models to manipulate the generative process, several studies [3,10,12] have attempted to disentangle meaningful hidden factors. Kulkarni et al. [10] proposed a training method enabling VAE to disentangle scene structure (rotation, elevation, and light) and to generate faces and chairs. Chen et al. [3] introduced an additional objective term that encourages GAN to learn disentangled representation in an unsupervised manner. Mathieu et al. [12] extended VAE with adversarial training to disentangle hidden factors from observations. While these approaches represent meaningful progress in that there is no need for strong supervision, the amount of variation they can handle is confined to one-dimensional control, making it difficult to apply existing methods to the extraction of human posture, which is represented by a high-dimensional vector.

More plausibly, recent work on conditional generative models [14–16,19,20] has shown promising results, adding attributes on latent space and generating an image corresponding to those attributes. Introducing conditional VAE, Yan et al. [19] learned to generate a desired image from the visual attributes; specifically, they succeeded in generating plausible images of faces from facial attributes. In studies using conditional GAN [15,16,20], text-to-image synthesis [15,20] has shown promise in generating images from text descriptions. Reed et al. [16] generated images by controlling object locations with key point and bounding box annotations. By adopting a conditional approach, conditional generative models demonstrate the possibility of transforming the image by changing only target attributes and fixing the others. However, even the most advanced methods take no account of interference between conditional attributes and the remaining variations. For that reason, as our experiments will illustrate, they fail to preserve person identity when changing posture.

[*] Corresponding author.
E-mail address: jychoi@snu.ac.kr (J.Y. Choi).

https://doi.org/10.1016/j.patrec.2018.06.020
0167-8655/© 2018 Elsevier B.V. All rights reserved.

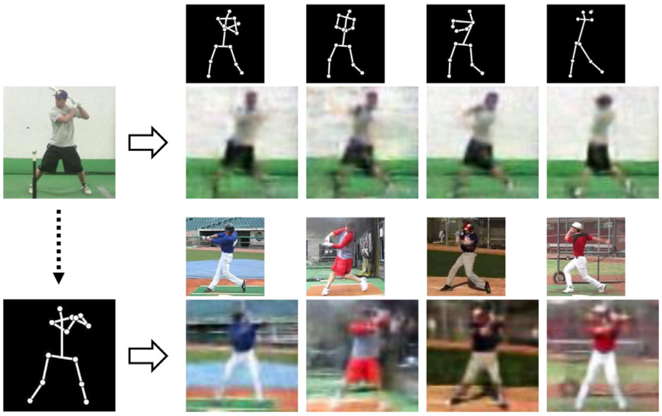

Fig. 1. Example results of the proposed Pose Transforming Network.

In this paper, we propose a novel deep conditional generative model for human pose transforms, namely *Pose Transforming Network*. We assume that an image is generated from two factors of variation: posture, governing posture semantics in an image, and identity, which captures identity factors such as human personality and background appearance. In our work, the posture factor is given as conditional information to represent the desired posture and should therefore capture all the variations that give rise to change of posture. As compared to existing conditional approaches, the key difference is that our model can generate arbitrarily pose-transformed images while preserving the identity property of the input image.

Our model consists of two encoders and one decoder, which means that the posture and identity factors can be learned separately. The first encoder handles pose embedding. In many cases, the given conditional pose description (joint etc.) may not represent similarity or distinctiveness due to large variations in body shapes. To mitigate this problem, we introduced pose embedding technique to produce more reliable conditioning factors for the generative model. The other encoder, for image encoding, learns to extract only identity information that is not related to the human posture in the image. A decoder network is then designed to generate corresponding images from the posture and identity factors.

To achieve this, a novel loss function and training method are proposed. The goal of training is not only to find the best parameters for generating a plausible target image but also to disentangle the images identity and pose factors. To this end, three loss terms are proposed: reconstruction loss, pose transform loss, and identity switch loss. By means of the new loss function, the network can be trained to prevent mutual interference between the two factors (posture and identity). The whole network is trained in end-to-end stochastic gradient descent [8] by use of the reparameterization trick [9]. The validity of the proposed approach is verified by investigating the models pose-transforming capability and comparing it with the recent conditional VAE approach (C-VAE) [19].

2. Pose transforming network

2.1. Overall scheme

Fig. 2 depicts the overall scheme of the proposed method. Let $x \in \mathcal{X}$ be an image of certain person and φ_t be a description of the target posture. For the pose description φ_t, it is reasonable to represent the posture by positions of body parts. In this work, joint positions [21] are used. Then, our goal is to generate an image x_t in which the posture of the person in x is transformed according to the target pose description φ_t.

In this work, a latent variable z is divided into two parts: a pose variable c that controls the pose variation in the image, and an identity variable u that characterizes the remaining variations such as personality and background. That is, $z = [u, c]$. By varying u and c, we can manipulate the human posture while the model generates images corresponding to z. Thus, the goal is to clearly disentangle two changes in the image which is governed by each part of the latent variable z. To do this, we present two encoders that make u and c represent only the corresponding variations from the observed data x and φ_t, respectively.

The first encoder $E_c(\varphi_t;\ W_C)$ is defined for pose embedding, yielding a pose variable c from the observed data φ_t. Sometimes, the pose description φ_t is inaccurate in case that the image is not aligned well. Moreover, there is no one who has the exactly the same posture. Even in the same posture, the vector value is different because each person has a different height and shape. With limited amount of data, it usually causes over fitting on pose description φ_t, which is not desirable for learning the generative model. To mitigate the problem, we define a pose variable c as a random variable, instead of using the value φ_t directly. We randomly sample the pose variable c from the distribution given as

$$c \sim N(m_c(\varphi_t), diag(\sigma_c^2(\varphi_t))), \tag{1}$$

where a deterministic encoder $E_c(.)$ can be defined ,which maps pose description φ_t to their mean function $m_c(\varphi_t)$ and standard deviation function $\sigma_c(\varphi_t)$. That is,

$$[m_c(\varphi_t), \sigma_c(\varphi_t)] = E_c(\varphi_t; W_C). \tag{2}$$

The proposed formulation encourages robustness to small perturbations along similar postures, and thus can avoid the over fitting on the pose description φ_t.

The second encoder $E_u(x;\ W_U)$ is for image encoding which compresses image x to identity variable u. The identity variable u characterizes the whole remaining variations which encrypts the identification and the background of the image x. For the observed image x, the encoder $E_u(.)$ produces the Gaussian mean function $m_u(x)$ and the standard deviation function $\sigma_u(x)$ for the identity variable u,

$$[m_u(x), \sigma_u(x)] = E_u(x; W_U). \tag{3}$$

During training (see Section 3), this network forces identity variables obtained from a same person to be close to each other regardless of his posture. If so, only available source for transforming a posture in the image comes from the variation of pose variable c. While the identity semantics in x could be maintained in x_t by transferring the corresponding information to the decoder via u.

Given u and c, the latent variable z is described by the Gaussian distribution,

$$z \sim N(m_z(x, \varphi_t), diag({\sigma_z}^2(x, \varphi_t))), \tag{4}$$

where the mean function and the standard deviation are given by $m_z(x, \varphi_t) = [m_u(x), m_c(\varphi_t)]$, and $\sigma_z(x, \varphi_t) = [\sigma_u(x), \sigma_c(\varphi_t)]$. Using z sampled from the Gaussian distribution in Eq. (4), the proposed decoder $D(z,\ W_D)$ generates the proper output response $\hat{x}_t$ relates to z,

$$\hat{x}_t = D(z; W_D). \tag{5}$$

The whole system can be derived by VAE formulation (see Section 2.2). Thus, by maximizing the variational lower bound of likelihood $p_\theta(x_t|z)$, the model can generate images conditioned on z. However, the purpose of learning is not only to generate plausible image corresponds to the latent variable, but also to disentangle the information which the pose variable and the identity variable convey. To achieve the purpose, we also propose novel training method and loss function in Section 3.