Generated Motion Maps

Yuta Matsuzaki, Kazushige Okayasu, Akio Nakamura
Tokyo Denki University
matsuzaki.y, okayasu.k@is.dendai.ac.jp, nkmr-a@cck.dendai.ac.jp

Hirokatsu Kataoka
National Institute of Advanced Industrial Science and Technology (AIST)
hirokatsu.kataoka@aist.go.jp

Abstract

The paper presents a concept for generated motion maps to directly generate a human-specific modality such as human pose and stacked optical flow, with only one rgb-image. Although the conventional approaches have achieved a complicated estimation with a discriminative model, we find the solution with a recent generative model. The two primary contributions in this paper are as follows: (i) proposed approach directly generates a {human pose heatmap, stacked optical flow} from an rgb-image, (ii) we have collected a database which contains image pairs between RGB-channel and image modality (pose-based heatmap and stacked optical flow). The experimental results clearly show the effectiveness of our generative model, as well as its ability to generated motion maps.

1. Introduction

In the beginning of the 21st century, the vision-based algorithms have been proposed to solve several problems on human behavior analysis [6]. The human behavior analysis, e.g. face analysis and action recognition, enables to make components such as visual surveillance, traffic safety system and virtual reality. We usually utilize a discriminative model to solve the problems. However, a specific algorithm is employed in the focused problem.

On one hand, generative models are rapidly increasing by started from the GAN (generative adversarial networks) [2]. The GAN effectively trains an image generator to strengthen both discriminative and generative models. The trained generative model has a great performance on image generation. Hereafter the GAN is improved with deep convolutional networks (DCGAN; Deep Convolutional GAN [7]) and pixel-dependent mutual conversion (image-to-image or pix2pix [4]). We can mutually convert any images especially in the pix2pix, that is one of the general-purpose algorithms.

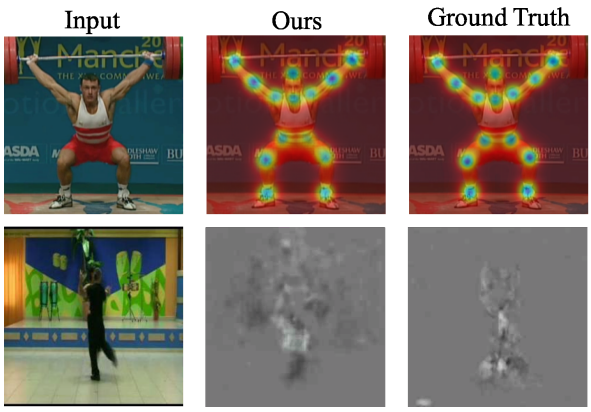


Figure 1. Our concept for generative motion maps (best view in color): The figure shows the direct transformation from rgb-image to human pose heatmap (top) and stacked optical flow (bottom). The three columns indicate (left) input RGB-image, (middle) output with our generative model and (right) ground truth with convolutional pose machines (CPM) [9] and dense optical flow [1].

Here if we can directly generate a modality (e.g. pose, flow) which is required in human behavior analysis, it is beneficial to apply an open-use application.

The paper presents a concept for generated motion maps in human behavior analysis. We develop a framework that directly generates human motion maps. In the paper we try to convert {human pose heatmap, stacked optical flow} from an RGB-image (see Figure 1).

The paper contains two main contributions:

- Our proposal directly generates a {human pose heatmap, stacked optical flow} from an RGB-image with a generative model. To achieve the challenging work, we must (i) estimate a human pose heatmap from a huge pattern-space, and (ii) get a stacked optical flow (e.g. 10-frame stacking) with only one rgb-channel.

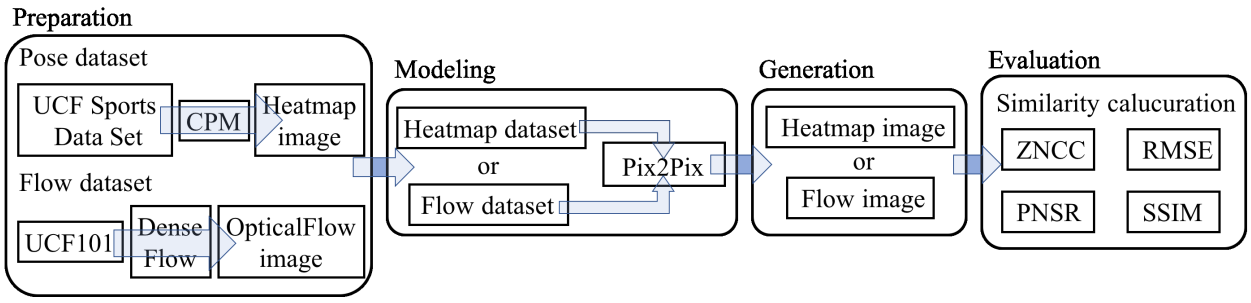


Figure 2. Process flow of our gnerative model.

- We have collected a database which contains image pairs to train pixel-level correspondences.

2. Generative model of pose heatmap and stacked optical flow

We apply the means of image generation or style transformation to pose estimation or optical flow generation by GAN without complicated models for specific processing. Regarding the pose estimation, human pose heatmap representing joint positions is generated. Regarding the optical flow generation, optical flows calculated from ten frames are accumulated into one image. It is difficult to estimate pose from enormous space without complicated models. At the same time, a generation of 10-frame stacked flow image from only one RGB-image is a challenging task.

2.1. Process flow

Figure 2 shows the process flow of dataset preparation, generative model, generation and evaluation. We prepared a pose heatmap and stacked optical flow datasets respectively. The former is generated on the UCF Sports Action Data Set [8] in which the heatmaps are generated with the convolutional pose machine (CPM) [9]. And the latter is generated on the UCF101 dataset [5], we applied Farneback optial flow [1]. As the constraint of image collection, we select image the whole body of estimation target person is shown in the former and the motion of a human remarkably appears in the latter. Images of human pose heatmap are learned by pix2pix, one of derivation methods of GAN, to create a model generating the pose heatmap corresponding to the newly-appeared pose. Concerning the optical flow image, the same procedures are performed. Each dataset includes 600 images; 400 images for learning, 100 images for network evaluation, and 100 images for testing.

2.2. Evaluation methods

We compare a generated heatmap or flow-image with processed image with each CPM and DenseFlow. The evaluation methods of the generated image are shown below:

- Zero-mean Normalized Cross Correlation (ZNCC)
- Root Mean Squared Error (RMSE)
- Peak Signal-to-Noise Ratio (PSNR) [3]
- Structural Similarity Index Measure (SSIM) [10]

ZNCC is one of basic similarity indices of template matching. The calculation formula of ZNCC is shown in equation (1).

$$ZNCC = \frac{\sum_{i=1}^{N}\sum_{i=1}^{M}(g(i,j)-\mu_g)(f(i,j)-\mu_f)}{\sqrt{\sum_{i=1}^{N-1}\sum_{i=1}^{M-1}(g(i,j)-\mu_g)^2}\sqrt{\sum_{i=1}^{N-1}\sum_{i=1}^{M-1}(f(i,j)-\mu_f)^2}} \quad (1)$$

The possible areas for ZNCC are -1 to 1. In ZNCC, we assume the density value distribution to f and g in the image area and calculate statisticaly the cross correlation coefficient. μ_f and μ_g are averege values of f and g. The higher the number is, the higher the similarity.

We calculated RMSE by equation (2).

$$RMSE = \sqrt{\frac{1}{MN}\sum_{i=1}^{N-1}\sum_{i=1}^{M-1}(g(i,j)-f(i,j))^2} \quad (2)$$

PSNR represents the ratio of the maximum power that can be taken by the signal that affects the reproducibility of image quality and the noise that causes deterioration. In other words, PSNR is an index to judge image quality by the average change amount of pixel values before and after encoding. PSNR is obtained by the following equation.

$$PSNR = 20\log_{10}\frac{MAX}{RMSE} \quad (3)$$

The standard value of PSNR in irreversible image and video compression is 30 to 50 dB, and the higher the numerical value, the better the image quality.

SSIM evaluates the image quality by the amount of change before and after encoding of the following three elements.

- Pixel value (luminance value)
- Contrast

Table 1. Evaluation of our pose heatmap generation

Method	Mean	Standard deviation
ZNCC	0.950	0.0269
RMSE	9.20	2.00
PSNR	29.1	1.94
SSIM	0.909	0.0212

Table 2. Evaluation of our stacked optical flow generation

Method	Mean	Standard deviation
ZNCC	0.0179	0.121
RMSE	2.99	3.35
PSNR	40.6	5.06
SSIM	0.962	0.0388

- Structure

SSIM is obtained by the following equation.

$$SSIM = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (4)$$

$$C_1 = (K_1 L)^2 \quad (5)$$

$$C_2 = (K_2 L)^2 \quad (6)$$

μ is the covariance and σ is the standard deviation. We set the parameter as $K_1 = 0.01$, $K_2 = 0.03$, $L = 255$. In the SSIM, the larger the number, the better the image quality.

3. Results

We describe the results of generated pose heatmaps and flow images by our generative model. In Table 1 and 2, the evaluation results of the generated images. Also, generated image are shown in Figure 3(a) and 3(b).

The evaluation methods ZNCC and SSIM show high score in generating heat map images (Table 1). This indicates that the generated image has a high similarity with respect to the correct image. In addition, the RMSE shows a low numerical value of 9.20. This indicates that the error between the correct image and the generated image is low. Therefore it can be considered that the generation of the heatmap can be performed with high accuracy. In addition, as shown in Figure 3(a), a plurality of persons are present in one image, not only a person at the image center but also a person appearing in the background and an adjacent are estimated. From this, it is possible to confirm the flexibility of generating the heatmap by pix2pix.

Table 2 indicates that SSIM is high and RMSE is low score in the generation of optical flow image. However, the ZNCC score is very low. The score shows that although the similarity of the generated image is high, there are very few matching points. In other words, it is considered that a different image is generated compared with the correct image in the generated image. As a cause of high SSIM and RMSE scores, it is conceivable that the brightness value, contrast and structure peculiar to the optical flow image have a bad influence. However, we have confirmed that the generated image is not wrong. An example is shown in Figure 3(b). The original input image has large movement and shake by the camera and the pure motion flow of human is not extracted by Dense flow. In generative optical flow, it can be seen that a generated flow image is generated focusing only on the flow of the human. The generation of optical flow by the generative model suggested the possibility to capture the pure motion of human.

4. Conclusion

In this paper, we proposed a generated motion maps which is direct human-modality (e.g. pose heatmap, stacked optical flow) generation from an input image. A recent generative model allows us to convert from RGB-image to an accurate pose heatmap and stacked optical flow. The approach is straight-forward, however, the evaluation of human-specific generative model is quite important for human behavior analysis. In the creation of pose heatmap, we have created accurate images in terms of both quantitative and qualitative evaluation. Using the pose heatmap, 2D human joints are easily predicted in an image sequence. Especially in the stacked optical flow generation, we can get 10-frame stacked flow fron an RGB-image. We believe that the process is one of the motion prediction. Moreover, the proposed generative model predicted human-specific pure motions without background areas.

In the future, we try to understand human actions or estimate object-focused optical flow with generated motion maps.

References

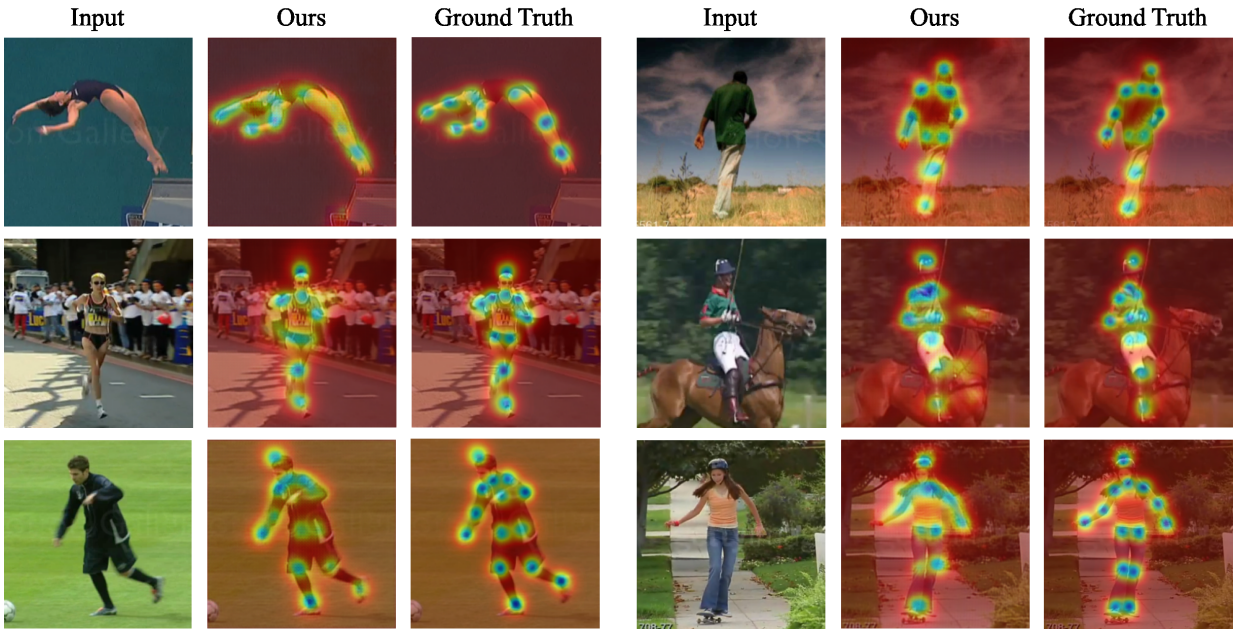
[1] G. Farneback. Two-frame motion estimation based on polynomial expansion. SCIA'03 Proceedings of the 13th Scandinavian conference on Image analysis, 2003.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. Neural Information Processing Systems (NIPS), 2014.

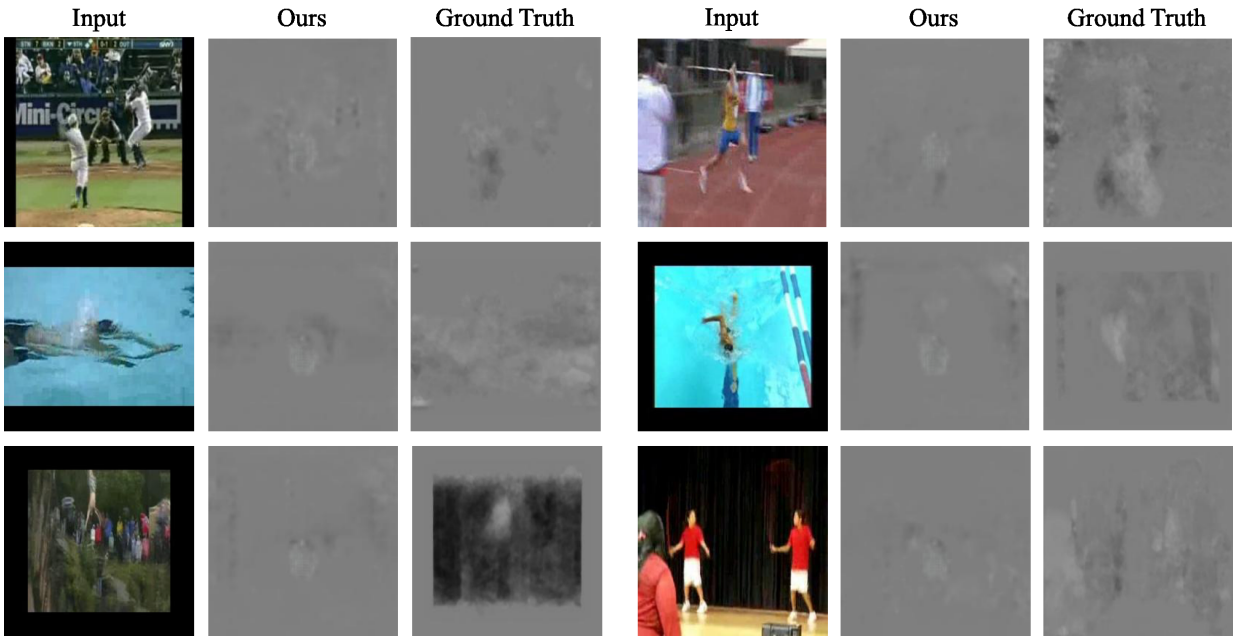
[3] Q. Huynh-Thu and M. Ghanbari. Scope of validity of psnr in image/video quality assessment. Electronics Letters, vol.44, issue.13, 2008.

[4] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial nets. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[5] A. R. Z. Khurram Soomro and M. Shah. Ucf101: A dataset of 101 human action classes from videos in the wild. CRCV-TR-12-01, 2012.



(a) Pose heatmap generation. From an RGB-image to pose heatmap.



(b) Stacked optical flow generation. From an RGB-image to 10-frame stacked optical flow.

Figure 3. Results of our proposal on human action videos.

[6] T. B. Moeslund, A. Hilton, V. Kruger, and S. L. Visual analysis of humans: Looking at people. Springer, 2011.

[7] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. International Conference on Learning Representation (ICLR), 2016.

[8] K. Soomro and A. R. Zamir. Action recognition in realistic sports videos. Computer Vision in Sports. Springer International Publishing, 2014.

[9] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[10] H. S. Zhou Wang, A.C. Bovik and E. Simoncelli. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing, vol.13, issue 4, 2004.