Robust human motion recognition from wide-angle images for video surveillance in nuclear power plants

Daisuke MIKI^{***}, Shinya ABE^{*}, Shi CHEN^{**} and Kazuyuki DEMACHI^{**}

^{*}Information Technology Group, Tokyo Metropolitan Industrial Technology Research Institute

2-4-10 Aomi, Koto-ku, Tokyo, 135-0064, Japan

E-mail: miki.daisuke@iri-tokyo.jp.or.jp

^{**}Department of Nuclear Engineering and Management, The University of Tokyo

7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

Received: 28 October 2019; Revised: 3 February 2020; Accepted: 24 March 2020

Abstract

Installation of surveillance cameras in nuclear power plants is critical to protecting the facilities against terrorist attacks or monitoring the reactor operator. This has led to large amounts of video surveillance data, creating a demand for automatic detection of anomalies or suspicious movements. Tracking human motion from video sequences is a notable technique used for detecting anomalies in human behavior and is currently achieved with the use of a depth camera. However, depth cameras require a complicated camera system and their field of view is limited. To overcome this problem, there is a need for recognizing human motion in wide-angle images – a view that often causes distortion. In this study, we devised a method for tracking human motion through wide-angle image distortion. The main contribution of this study is a methodology that automatically estimates the transformation parameters needed to improve the accuracy of motion recognition; these parameters are applied to a distorted wide-angle image in every frame. We propose a new multi-layered convolutional neural architecture for estimating the locations of human joints in images and transformation parameters simultaneously. When applied to distorted wide-angle images, the robustness of our method is demonstrated through a quantitative evaluation of human joint location prediction. In addition, we compare our method with a motion tracking system and an infrared-camera-based motion capture system to demonstrate its ability to handle wide-angle and close-range images.

Keywords : Nuclear security, Video surveillance, Human motion recognition, Convolutional neural network, Image processing

1. Introduction

Surveillance cameras in nuclear power plants (NPPs) have various important roles such as protection against terrorist attacks, prevention of radioactive material theft, or monitoring the reactor operator. Therefore, in NPPs, a considerable amount of surveillance video is preserved, creating the demand for automatic detection technique of anomalies or suspicious movements. For video surveillance, various approaches have been proposed, including facial recognition, clothing identification, and object recognition. However, these methods are often ineffective as they cannot detect insider terrorists or unsafe operator conditions; these behaviors can be hidden through various factors such as maintenance behavior or a slightly incorrect operation which results in a dangerous situation. One effective method is to use a motion tracking device that recognizes human motion in a noncontact manner in order to detect malicious or unsafe actions. In a previous study (Chen and Demachi, 2019), this tracking was completed through hand motion time-series data and was used to distinguish malicious action from ordinary actions.

Widely used motion tracking devices recognize human motion using a depth camera such as a stereo or infrared camera. However, the use of a depth camera requires a complicated camera system and has a limited field of view (FOV). In particular, a camera with a narrow FOV is poorly suited for video surveillance because actions can be hidden outside the FOV or at close range. To overcome this problem, there is a need to recognize human motion through wide-angle

surveillance videos. Almost all installed surveillance cameras store monocular RGB videos; thus, many studies have been conducted on human motion recognition using monocular RGB images. These studies are based on the human model (Felzenszwalb and Huttenlocher, 2005), the deformable part model (Yang Yi and Ramanan, 2011, 2012), and a convolutional neural network (CNN) (Toshev, and Szegedy, 2014). Owing to the advance in neural networks and deep learning researches in recent years, a CNN-based human motion recognition method is the most effective and easiest to implement, but these approaches are not designed for robust motion tracking from distorted wide-angle images.

In this study, we devised a method for tracking human motion that is wide-angle image distortion capable. The main contribution of this study is a methodology that automatically estimates the appropriate transformation parameters, which are needed to improve the accuracy of human motion tracking in every frame. The proposed method is based on a CNN using a keypoint estimation (KPE) module and a transformation parameter estimation (TPE) module. The overview of the proposed method is shown in Fig. 1. In this method, 1) the input frame is transformed using the appropriate transformation parameters (except for the initial frame); 2) human motion is estimated as a set of 2D confidence maps generated by the KPE module; then, 3) using the 2D confidence maps obtained from the KPE module, the TPE module estimates the transformation parameters for improving the accuracy of the next frame.

The robustness of our method through distorted wide-angle images is demonstrated through a quantitative evaluation of human joint prediction. In addition, we compare our method with an infrared-camera-based motion tracking system to demonstrate its ability to handle wide-angle and close-range images.

Fig. 1 Overview of the proposed method. This method is comprised of KPE and TPE modules. Human motion in images is expressed as confidence maps which indicates the location of human joints.

2. Human motion recognition

The CNN architecture for human pose estimation is illustrated in Fig. 2. As shown in the figure, the input is an RGB image that includes the target human figure, and the output is a confidence map that indicates the location of the joints. To predict the location of a joint in each image, the image is first given as an input to the CNN. Next, confidence maps are produced by repeating the convolution, pooling, and activation steps. These confidence maps, which are shown in Fig. 2, indicate where there is a high probability that a joint exists. Using this method, 14 confidence maps were produced. The coordinates with the highest score in each confidence map were assumed to be the positions of the joints. Increasing the number of layers in the network is an effective strategy to improve recognition accuracy. However, excessively multilayered networks can cause problems during backpropagation, resulting in the vanishing gradient and CNN training being less successful. Wei proposed a multi-stage CNN architecture (Wei et al., 2016) that enforces intermediate supervision to address the vanishing gradient problem. We adopted this network architecture for this system and made it capable of interpreting wide-angle image distortion. In this experiment, we always selected the last confidence map as the output.

During the training of the CNN, confidence maps $\mathbf{S}_j^*$ were taken from the annotated 2D human joint locations. We used the confidence maps for body part j . The generated ground truth confidence maps $\mathbf{S}_j^*$ for body part j were defined as

$$\mathbf{S}_j^*(\mathbf{p}) = \exp\left(\frac{\|\mathbf{p} - \mathbf{x}_j^*\|^2}{\sigma}\right) \quad (1)$$

where x_j^* is the ground truth location of the body part in the image and $\mathbf{p}$ is location of the pixel in the confidence maps. The cost function we aimed to minimize at the output of each stage at each level is given by

$$E_{KPE} = \sum_j \|\mathbf{S}_j(\mathbf{p}) - \mathbf{S}_j^*(\mathbf{p})\|_2 \quad (2)$$

where $\mathbf{S}_j(\mathbf{p})$ is the output of the CNN and $\mathbf{S}_j^*(\mathbf{p})$ is the ground truth confidence map.

Fig. 2 The CNN architecture in the KPE module. The CNN has a multi-stage structure and it aims to improve recognition accuracy. In each stage, the output confidence and feature maps from the intermediate layer are repeatedly concatenated. Concatenated feature maps are input to subsequent convolution layers.

3. Transformation parameter estimation

3.1 Fisheye image transformation

In this study, we use 180° fisheye images as wide-angle images, making it necessary to devise a method of interpreting distortion. Specifically, there is a method of using an estimator robust enough to interpret distortion without correcting an image, and one for correcting the image distortion. In the method without correction, preparing multiple estimators is necessary because the appearance of a human varies depending on the human's position. This method requires considerable time to train the estimators, creating a disadvantage in development as well as implementation. Therefore, we chose the method with correction.

In the fisheye image, a real-world point (X, Y, Z) is projected onto an image point (x, y) in the two-dimensional plane. Here, (x, y) in the image can be determined by the incidence angle θ and the distance r from the center of image. Therefore, (X, Y, Z) and the coordinates (x, y) in the image plane can be converted using the following equation and (r, θ) . In the method with correction, it is conceivable to undergo conversion from the projection method, which is represented by the following equations

$$r_{SGP}(\theta) = 2f_{SGP} \tan \frac{\theta}{2}, \quad (3)$$

$$r_{PP}(\theta) = f_{PP} \tan \theta. \quad (4)$$

Here, Eq. (3) and (4) denote stereographic projection (SGP) and perspective projection (PP), respectively, where focal lengths are f_{SGP} , f_{PP} , and the incidence angle is θ . This means that the ray is projected to the distance r_{SGP} and r_{PP} from the center of the image. According to Eq. (3), while using a stereographic projection, if an image sensor has radius $2f_{SGP}$, 180° FOV can be obtained. In contrast, an infinite imaging plane is required to obtain 180° FOV in the perspective projection of Eq. (4). Therefore, as shown in Fig. 3, it is necessary to choose the appropriate transformation parameter including azimuth angle, elevation angle and focal length for each distorted image conversion.

Fig. 3 Comparison of projection methods. (a) Elevation angle θ_v , (b) Azimuth angle θ_h , (c) SGP image at $\theta_{FOV} = 90^\circ$, (d) PP image at $\theta_{FOV} = 30^\circ$, (e) PP image at $\theta_{FOV} = 45^\circ$, (f) PP image at $\theta_{FOV} = 60^\circ$, and (g) a PP image at $\theta_{FOV} = 45^\circ$ and $\theta_h = 45^\circ$

3.2 Training the TPE module

To train the TPE module, we prepare synthetic human models in a distorted wide-angle image using a 3D computer graphic (3DCG) software, as shown in Fig. 4. These models are distorted with a given ground truth transformation parameter $\mathbf{T}^*$. Next, the images are given as inputs to the KPE module and confidence maps are returned. The confidence maps are then given as inputs to the TPE module, producing an output $\mathbf{T}$ of four-dimensional transformation parameters $\{f, \theta_h, \theta_v, \theta_r\}$, where focal length is denoted by f , and rotations are denoted by $\theta_h, \theta_v, \theta_r$. Finally, the cost function we aim to minimize is given by

$$E_{TPE} = \|\mathbf{T} - \mathbf{T}^*\| \quad (5)$$

Each transformation parameter is sampled from uniform random variables $U\{\hat{f} - \delta f, \hat{f} + \delta f\}$, $U\{\hat{\theta}_h - \delta\theta_h, \hat{\theta}_h + \delta\theta_h\}$, $U\{\hat{\theta}_v - \delta\theta_v, \hat{\theta}_v + \delta\theta_v\}$ and $U\{\hat{\theta}_r - \delta\theta_r, \hat{\theta}_r + \delta\theta_r\}$. Here, we set $\hat{f} = 1.13 \text{ mm}$, $\hat{\theta}_h = 0^\circ$, $\hat{\theta}_v = 0^\circ$, $\hat{\theta}_r = 0^\circ$, $\delta f = 0.29 \text{ mm}$, $\delta\theta_h = 45^\circ$, $\delta\theta_v = 20^\circ$ and $\delta\theta_r = 30^\circ$. Fig. 5 shows the transformed images obtained using these transformation parameters and the associated confidence maps output from the KPE module.

Fig. 4 Training method of the TPE module. Confidence maps corresponding to synthesized images transformed by ground truth transformation parameter $\mathbf{T}^*$ are estimated by the KPE module. Transformation parameter $\mathbf{T}$ is obtained as output of TPE module. Then, the Euclidean loss between transformation parameter $\mathbf{T}$ and ground-truth transformation parameter $\mathbf{T}^*$ is estimated.

Fig. 5 Human models synthesized with computer graphic software for training of TPE module. Synthesized images are (left) transformed in PP format and (right) output confidence maps are output from KPE module.

4. Experiment and Result

4.1 Implementation

CNN optimization was performed on a workstation with a single NVIDIA GPU (Geforce GTX 1070). For the training of the KPE module, the parameters were updated 100,000 times using 17,000 pairs of MPII dataset (Andriluka et al., 2014) images and the ground truth values of joint locations. The input frames were scaled into 368×368 pixels, randomly rotated between -20° to 20° , and resized back to 368×368 pixels. The variance of the gaussian was set to $\sigma = 1.2$ with the output confidence maps size of 46×46 .

For the training of the TPE module, we prepared 10,000 images, and a 3DCG human model of 50 bodies (25 males and 25 females) was prepared using Adobe Fuse (Adobe Fuse). The color of clothing provided a random RGB value ranging between 0 and 255 for every frame as an output. For pose information, we used the CMU mocap dataset (CMU Graphics Lab Motion Capture Database). The parameters of the CNN were updated 10,000 times.

4.2 Evaluation

In the experiment, qualitative and quantitative evaluations were performed to confirm whether our method could recognize the motion of NPPs workers. In NPPs, it is important to be able to recognize humans wearing special purpose clothing, such as a hard hat or mask. Therefore, in the qualitative evaluation, a human wearing work clothes and protective gear was photographed and used. We also evaluated our method by comparing it to a depth camera-based motion tracking device. The subject had a distance d between 0.8 m and 2.0 m from the camera and the azimuth angle θ_h is varied between -45° and 45° . At each point, the human performed simple exercises, such as swinging their arms, and we observed whether the system could recognize the motion. In quantitative evaluation, we evaluated our method on 1,000 images from the Leeds Sports (LSP) dataset (Johnson and Everingham, 2010). To confirm the distortion robustness, the Probability of Correct Keypoint (PCK) metric (Yang and Ramanan, 2013) was calculated. The PCK metric expresses the accuracy from the comparison between the ground truth location of each joint's location and the estimated location of the joint from the confidence maps, showing the maximum value for each. According to the evaluation protocol for the LSP datasets, we evaluated $\text{PCK}@0.2$. Here, the notation " $\text{PCK}@n$ " is obtained by normalizing the distance between the estimated joint position, x_j , and the ground truth value, x_j^* , with the appearance of the human in the image d_j , where

$$d_j = \frac{\|x_j - x_j^*\|_2}{\|x_{sholder} - x_{hip}\|_2} \quad (6)$$

The ratio was set such that $d_j < n$. Here, $\|x_{sholder} - x_{hip}\|_2$ is the normalization of the size of the human in the image.

4.3 2D Human pose estimation from monocular fisheye images

In order to confirm whether the NPP workers could be recognized by our method, we first evaluated if it was possible to recognize humans wearing general work clothing and protective clothing. Fig. 6 confirms that the human pose was properly recognized in both work clothing and protective clothing.

Fig. 6 Result of human motion recognition with our method. (a) Appearance of subjects, (b) input fisheye image, (c) output confidence map, and (d) detected skeleton.

The comparison between the depth-based motion tracking device and our method is shown in Fig. 7. When a human exists at 2.0 m from the imaging device, it was possible to recognize the human pose using both methods. In particular, it was possible to predict the subject's pose when the subject appears with image distortion using the proposed method. When a human exists at a short distance (0.8 m), the depth-based system shows only a few parts of the human and the pose cannot be recognized. In contrast, using the proposed method, the entire body was in the FOV and the human pose was recognizable. Furthermore, even when the human moved in the horizontal direction and could not fit within the FOV of the depth-based system, the human was clearly within the FOV using the proposed method. Moreover, it was confirmed that location recognition of the human joint is possible despite the distortion of the image.

Fig. 7 Comparison of depth camera-based motion capture device and our method. (a) $d=2.0$ m, $\theta_h=0^\circ$ (b) $d=0.8$ m, $\theta_h=0^\circ$ (c) $d=1.5$ m, $\theta_h=45^\circ$

The quantitative result on PCK metric evaluation is shown in Fig. 8. In the range of θ_h between 0° and 60° , the value of PCK@0.2 was approximately 70%. The recognition accuracy decreased when $\theta_h < 60^\circ$ because part of the recognition target person was outside the FOV. In addition, comparing of PCK@0.2 without the image TPE module, the recognition rate was poor when a person was present near the center of the image. It is possible that this is caused by a human appearing too small when transformed to the stereoscopic projection. Using the TPE module, the proposed method boosts the performance. In the range of θ_h between 0° and 70° , the proposed method demonstrated better results and confirmed that robust recognition against image distortion is possible.

Fig. 8 Comparison of percentage of detected joints on LSP for each body part. (a) head and neck, (b) lower legs, (c) wrists, and (d) average of whole body.

5. Conclusions

In this study, we developed a robust human motion tracking algorithm for video surveillance and proposed a new architecture of CNN for human motion tracking through wide-angle images. Robustness to distortion caused by the standing position of the human to be recognized with respect to the wide-angle camera. In accuracy evaluation, it was confirmed that the proposed method was able to recognize the motion of human in both work and protective clothing. Compared to the depth-based motion tracking system, recognizing the posture of a human existing at a short distance (0.8 m) was possible. It was confirmed that adequate recognition accuracy was obtained in the range of 0° – 70° and human pose recognition at wide angles was possible. As an extension of the current study, we plan to improve the structure of the CNN to improve the recognition accuracy and time-series data analysis algorithm for anomaly detection.

Acknowledgment

This work was supported by Japan Society for the Promotion of Science Grants-in-Aid for Scientific Research under Grant JP19K20310 and JP19K05324

References

- Adobe Fuse, (online), available from < <https://www.adobe.com/jp/products/fuse.html> > (accessed 1 Feb 2020).
- Andriluka, M., Pishchulin, L., Gehler, P., and Schiele, B., 2D human pose estimation: New benchmark and state of the art analysis, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014), pp. 3686–3693.
- Chen, S. Demachi, K., Proposal of an insider sabotage detection method for nuclear security using deep learning, Journal of Nuclear Science and Technology, Vol.56, No.7 (2019), pp.599–607.
- CMU Graphics Lab Motion Capture Database, (online), available from <<http://mocap.cs.cmu.edu/>>, (accessed 1 Feb 2020).
- Felzenszwalb, P. F., and Huttenlocher, D. P., Pictorial structures for object recognition. International journal of computer vision, Vol.61, No.1 (2005), pp.55–79.
- Johnson, S. and Everingham M., Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation. Proceedings of the British Machine Vision Conference (BMVC) (2010), pp. 12.1–12.11.
- Toshev, A., and Szegedy, C., Deeppose: Human pose estimation via deep neural networks. Proceedings of the IEEE

Conference on Computer Vision and Pattern Recognition (CVPR) (2014), pp. 1653–1660.

Wei, S., Ramakrishna, V. Kanade, T. Sheikh, Y., Convolutional Pose Machines, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016), pp.4724–4732.

Yang, Y. and Ramanan, D., Articulated pose estimation with flexible mixtures-of-parts. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2011), pp. 1385–1392.

Yang, Yi and Ramanan, D. Articulated human detection with flexible mixtures of parts. IEEE transactions on pattern analysis and machine intelligence, Vol.35 No.12 (2013), pp.2878-2890.