Head Pose Estimation for an Omnidirectional Camera using a Convolutional Neural Network

Yusuke Yamaura, Yukihiro Tsuboshita, and Takeshi Onishi
Research & Technology Group
Fuji Xerox Co., Ltd.
6-1 Minatomirai Nishi-ku, Yokohama-shi, Kanagawa 220-8668, Japan
{yamaura.yusuke, Yukihiro.Tsuboshita, Takeshi.Onishi}@fujixerox.co.jp

Abstract—**The human head pose provides insights on the activities or intentions of a given person. Head pose estimation techniques are thus often employed in intelligent surveillance camera systems for marketing analysis or security monitoring. Nowadays, omnidirectional cameras have become widely used in surveillance systems owing to their unique property of wide-range coverage. However, this property causes significant changes in visual appearance and distortions inside the image, and general approaches using a head image may fail in estimation. In this paper, we thus propose a method for head pose estimation using omnidirectional camera images. The proposed model employs both a head image and full body image for cases in which a face is self-occluded and the head image is thus almost useless. In addition, image attribute data are integrated into the network to learn the relation between the changes in appearance or distortion and locations inside the whole image. Experiments are conducted to compare the accuracy of the presented approach with those of ordinary methods. It is verified that the proposed method improves the accuracy by more than 19% over the baseline method.**

Keywords—convolutional neural network; head pose estimation; omnidirectional camera

I. INTRODUCTION

Head pose estimation in surveillance video images is an important task in computer vision because it provides insight on human behavioral intentions. Therefore, it is useful for marketing analysis in retail stores and security monitoring in public spaces [1,2,3,4,5]. Recently, the omnidirectional camera has become increasingly used in surveillance systems on account of its significant advantage of wide-range coverage. Specifically, it can capture a 360-degree field of view in the horizontal plane and thus helps reduce the number of cameras needed and overall costs [6].

However, to the best of our knowledge, few studies have focused on the application of head pose estimation using an omnidirectional camera. There are two major technical difficulties due to the key characteristic of the omnidirectional camera: large changes in visual appearance, and strong distortions in the radial direction. In particular, the former property causes self-occlusion of a face when appearing under the camera. Therefore, general approaches that rely on extracted features from a head image may fail.

In this paper, we propose a head pose estimation technique for omnidirectional images. The primary contributions of this research are outlined as follows. (i) We solve the self-occlusion problem by using information of both a head image and full body image based on the concept of a dual-source convolutional neural network (DS-CNN) proposed by Fan et al. [7]. Our model architecture enables inference of the head pose from body appearance, even when the face is occluded. (ii) We incorporate the location dependence information into the network as additional features to enable the network to consider the relation between the changes of appearance or distortion and projected locations inside a whole image. The effectiveness of the proposed method is shown with our original omnidirectional image dataset collected from an actual retail store. The proposed approach achieves an accuracy improvement of greater than 19% compared with ordinary CNN model.

II. RELATED WORK

Various head pose estimation techniques using a monocular camera have been proposed in the last few decades [8,9]. More recently, the convolutional neural network (CNN) has succeeded in many image recognition tasks in computer vision and head pose estimation. Here, we review only the CNN-based methods because we likewise adopt the CNN architecture. Most of these existing methods can be categorized into a classification or regression problem, and a few studies have combined them.

Classification: In the classification problem, the head pose is labeled in accordance with a discrete head angle at certain intervals, which are basically three types of rotations: yaw, pitch, and roll. Studies that addressed the classification problem in this context include those that follow. Patacchiola et al. [10] employed a CNN architecture with adaptive gradient methods for head pose classification as a real-time application. Tran et al. [11] constructed three CNNs corresponding to each component. They used a transfer learning technique to train each model. Raza et al. [12] proposed a system that estimates head pose and body orientation automatically and displays then simultaneously. However, in their method, the head pose and body orientation classifiers are separately trained with different CNNs and they do not at all combine features of head pose and

978-1-5386-0951-4/18/$31.00 ©2018 IEEE

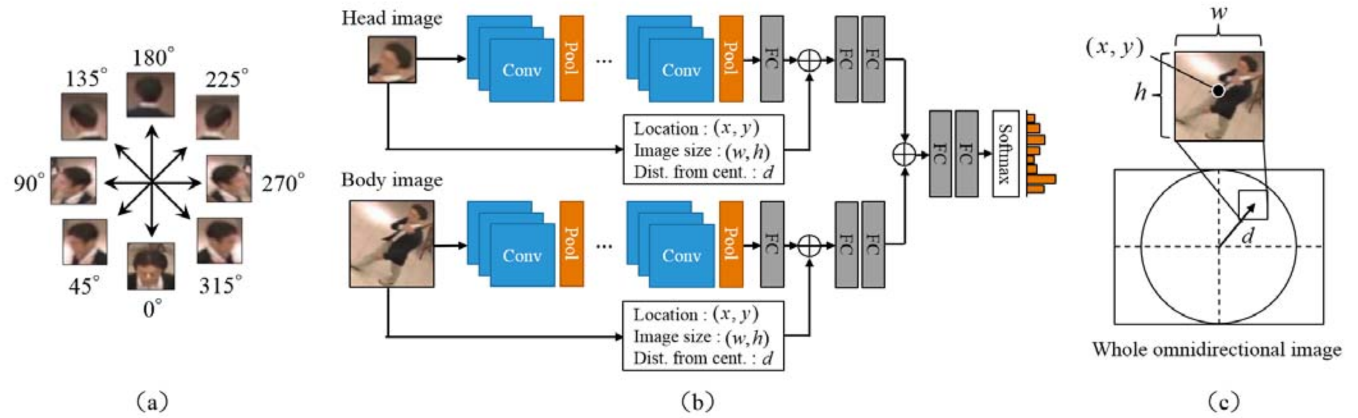

Fig. 1. Summary of the proposed method. (a) Eight head pose classes in the yaw rotation. (b) Architecture of the proposed network, which takes two input images, the head and body images, and their attribute data. The words, "Conv," "Pool," and "FC" refer to the respective convolutional, pooling, and fully connected layers. (c) Description of the attribute data: the x, y coordinates and the width, height, and distance from the center of the whole omnidirectional image.

body orientation for head pose estimation as we do in the present study.

Regression: In the regression problem, on the other hand, the head pose angles are directly estimated as continuous values. Ahn et al. [13] built a single CNN that outputs continuous head angles in three components for real-time applications. For head pose regression, Beyer et al. [14] proposed biternion nets, which can be trained with coarse head angles. This network enables sharing the dataset in classification and regression approaches. Liu et al. [15] generated synthetic head images by using three-dimensional (3D) head models for head pose regression. They evaluated the performance with both synthetic and real data. Xu et al. [16] combined local and global features to estimate head poses and to localize landmarks using CNN with regression.

Combination: Several researchers combined both classification and regression along with a single deep neural network model. Mukherjee et al. [17] combined two models, each of which was trained in classification and regression approaches. Ruiz et al. [18] introduced a multi-loss approach for head pose estimation, whereby the loss combines classification and regression to improve learning.

The above methods succeed in general-perspective camera images captured from an approximately frontal view. However, they are not assumed to be used in omnidirectional images. When appearing under an omnidirectional camera, the face is frequently self-occluded and the head image is almost fully comprised of the top of the head. In this case, it is almost impossible to estimate the head pose from the head image, even by the human eye. In addition to this problem, image distortions gradually increase toward the edge of the omnidirectional image in a radial direction. This causes slight differences among training data in the same class and leads to degraded performance.

III. Network Architecture

In this section, we describe details of our proposed method for head pose estimation. Following to previous works on head pose estimation for surveillance cameras, we employ the classification strategy and target only the yaw component. We define eight discrete head poses as classification labels, which are aligned at a certain interval in the yaw rotation, as shown in Fig. 1(a). Our model takes two types of input data, namely images of both the head patch and the full body patch, as well as their attribute data, and we classify them into eight head pose classes, as shown in Fig. 1(b). We assume that the head and body regions are detected inside a whole omnidirectional image using some effective object detection technique, such as Faster R-CNN [19].

A. Combination of Head and Body Image Features

The image visual appearance gradually changes in the radial direction in omnidirectional images, as previously mentioned. Owing to this unique property, the face is occluded under the camera and a cropped head image is almost fully comprised of the top of head, as noted above. Fig. 2(a) shows cases in which persons are present under a camera, and their faces are completely obscured. In these cases, it is nearly impossible to judge their head poses, even by a human. Nevertheless, this issue is not the case with respect to the body appearance. The body part is larger than the face part. Hence, it is difficult to be fully hidden under the camera. The body appearance should help in estimating the head pose. This is because the body orientation is highly correlated with the head pose and an inter-relational constraint exists between them, i.e., the head pose cannot be directed opposite of the body orientation. To incorporate the information of body appearance, we designed our network to employ a full body image as well as a head image, as shown in Fig. 1(b). This architecture is similar to the DS-CNN proposed by Fan et al. [7]. In their work, the DS-CNN is designed to take two kinds of input—the local part appearance, and the holistic view of the local part in the full body—for more accurate human pose estimation. We believe that this strategy is more effective in omnidirectional images than in general-perspective camera images because head images frequently become useless owing to the face self-occlusion in omnidirectional images, as described above. The difference is that our input images are fixed with respect to the head and body. The visual features of the head image and full body image are separately extracted through each CNN and then combined prior to the latter fully connected and softmax layers.

B. Integration of Attribute Features

We integrate the image location-dependent attribute data into our network for each head image and full body image to learn the relation between the degree of changes in visual appearance or distortion and the location in an omnidirectional image. In pre-processing, the head and body regions are assumed to be detected inside a whole omnidirectional camera image. As shown in Fig. 1(c), we define the attribute data as being comprised of three types of data (five continuous values): the x, y coordinates; the width and height of the detected region inside the whole image; and the distance from the center of the whole image. These five continuous values are vectorized and normalized, where each element in the vector is divided by the sum of all elements. That approach was shown in our pre-experiment to be the most effective for normalization compared with several major normalization methods.

This normalized attribute feature vector and image feature vector extracted through a CNN are simply concatenated and fed into the latter fully connected layers to learn the relation between them. Consequently, we obtain two feature vectors from each CNN architecture of the respective head and body. Then, they are integrated in the fully connected layers to provide a higher expressive performance.

IV. Experimentation

A. Implementation Details

In our experiment, the CNNs for the respective head and body images basically followed the architectures of two state-of-the-art CNNs, AlexNet [20] and VGGNet [21]. We adopted these two CNNs for the implementation to prove the effectiveness of our proposed method and to compare their performances. The most significant difference between our architecture and those of the two CNNs is that ours incorporates the attribute vector into the network. We concatenate the five-dimensional vector with 4,096 outputs of the first fully connected layer. Then, the concatenated 4,011-dimensional vector is fed into the two latter fully connected layers. We initially use parameters pre-trained on ImageNet [22] for CNNs, and then train the entire network end-to-end with our original omnidirectional image dataset.

B. Dataset Description

To the extent that we investigated, no omnidirectional image dataset for head pose estimation exists. Therefore, we uniquely collected a dataset using a commercially available omnidirectional camera. The data collection was conducted at a real convenience store located in the business area in Tokyo, Japan. We fully installed three cameras on the ceiling of the store and recorded a video over the course of a day with the cooperation of store managers. In this store, over ten perspective cameras are required to cover a whole area inside the store. We randomly selected frame images from the video data to reduce biases on the dataset, as many variations of people, clothes, and locations were included in the dataset. In addition, we created an annotation tool with which an annotator can "box" the area of the head and body to crop the images of them. The attribute data of the boxed rectangle, that is, the coordinates, size, and distance from the center of the whole image, were automatically recorded in a CSV file. Consequently, our original dataset contained data of 43,990 persons. All person images in the dataset were rotated and blurred by the Gaussian filter for privacy in pre-processing.

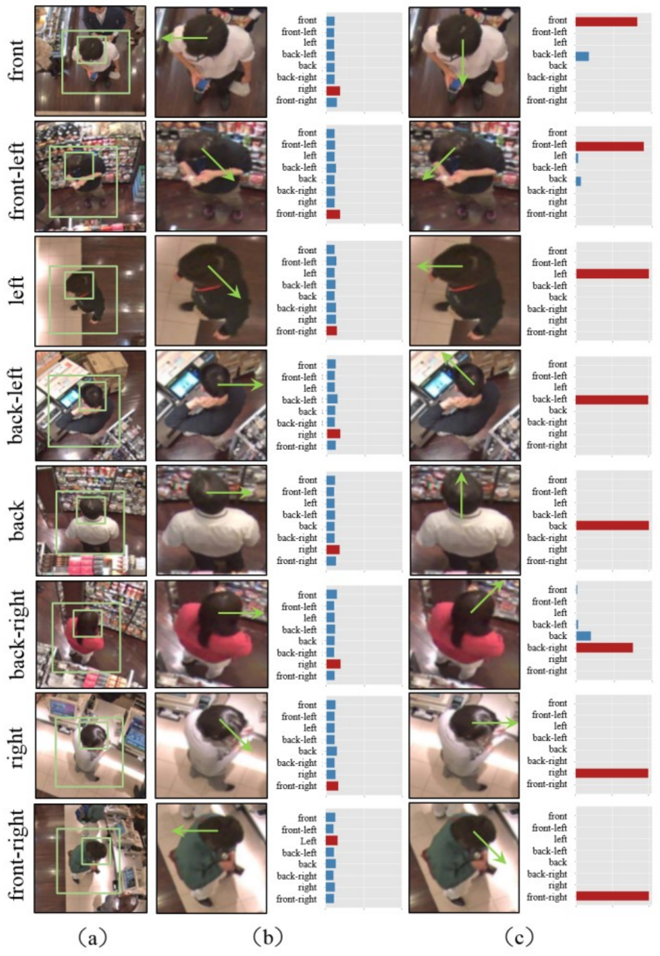

Fig. 2. Examples of head pose estimation results. (a) Detected regions of the head and body in pre-processing. (b)(c) Results of the model that employs only a head image and a body image as input. Left illustrates head orientation and right is classification probabilities that range 0 to 1.

C. Evaluation Results

We evaluated the classification accuracies using our original dataset divided into 75% training data and 25% test data. We had a total of six patterns of input data in our architecture—(i) a head image only, (ii) a body image only, and (iii) both the head and body images—with or without the attribute data. We tested these several patterns on AlexNet and VGGNet architectures. In addition, to demonstrate the performance of CNNs, we compared them with a conventional machine learning based method proposed by [23]. In their work, the Kullback-Leibler divergence is adopted to extract features and SVM classifiers are trained with one-against-rest strategy. We call their method as KL+SVM in this paper.

Table 1 and 2 shows the evaluation results of classification accuracy, KL+SVM and our proposed method, respectively. Obviously, the accuracy of KL+SVM is low compared with CNN methods and unsuitable for the omnidirectional images. As for Table 2, first, significant improvements are evident by combining the body image. The accuracies in both AlexNet and VGGNet are improved by more than 10% and 19%, respectively. Second, it is apparent that several increases are achieved by integrating the attribute data. It is interesting that a more significant improvement is achieved in the case in which only a body image is used instead of only a head image, whereby the difference is 2.5%. We infer that the appearance of body changes more than that of head because the body part largely appears inside an image and this increase the importance of attribute data.

TABLE I. CLASSIFICATION ACCURACY OF KL+SVM MODEL (%)

	KL+SVM
(i) head	27.5
(ii) body	20.5
(iii) head+body	27.6

TABLE II. CLASSIFICATION ACCURACY OF EACH CNN MODEL (%)

	AlexNet w/o att.	**AlexNet w/ att.**	**VGGNet w/o att.**	**VGGNet w/ att.**
(i) head	67.9	68.2	71.3	71.3
(ii) body	64.8	65.6	72.8	75.3
(iii) head+body	78.3	78.6	90.5	90.5

V. DISCUSSION

A. Effectiveness of Combining Head and Body Image Features

In this subsection, we discuss the effectiveness of combining the full body image features in our network. As mentioned previously, a head image is useless in the case in which the person is situated under the camera because the head image is comprised of the top of head and the face is self-occluded. We select eight examples of this case in Fig. 2(a). We can observe that it is almost impossible to estimate the persons' head poses from the head image because the faces are completely hidden by their heads. Fig. 2(b) and (c) illustrate the estimation results of the one-source model and dual-source model, respectively, where the head orientations are visualized by green arrows. We can observe that, although the one-source model fails in estimation, the dual-source model succeeds on all examples. These results demonstrate the fact that general image-based approaches using a head image may fail to estimate the head pose in the omnidirectional images, whereas combining the body appearance into the network is very effective, especially in the omnidirectional images.

B. Effectiveness of Integration of Attribute Features

We incorporated the attribute features into our network to learn the relation between the changes of appearance or distortion and the projected locations inside a whole image. From the evaluation results shown in Table 1, it is evident that the accuracy is slightly improved on AlexNet and VggNet. In addition, the degree of the improvement is not as significant compared with that of the combined body appearance. These results show that the changes of visual appearance in the same class are not a crucial problem in classification and that the most significant problem of face self-occlusion is mostly solved by incorporating the body appearance. However, attribute data are added to a body image, the accuracy is increased by a few percentages from the baseline. These phenomena were observed with both AlexNet and VGGNet; thus, the relation between the changes of appearance or distortion and the locations inside the whole image were considered to have been properly learned in the network.

VI. CONCLUSION

In this paper, we introduced an effective method of head pose estimation for omnidirectional camera images. Our model follows the DS-CNN and both head and body images are fed into the network, thereby enabling estimation of the head pose from the body appearance. Furthermore, we integrate attribute data of the input image into the network to learn the relation between the changes of appearance or distortions and the locations inside the whole image. The performance of our proposed method was evaluated using original dataset collected in a real store. We achieved more than 19% improvement of accuracy compared to the general state-of-the-art CNN. Because our proposed architecture employs two types of data—images and their attribute data—it can be widely applied in computer vision and various other fields.

References

[1] G. Medioni, I. Cohen, F. Brémond, S. Hongeng, and R. Nevatia, “Event detection and analysis from video streams,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 8, pp. 873 - 889, 2001.

[2] M. Valera and S.A. Velastin, “Intelligent distributed surveillance systems: A review,” Proc. Inst. Elect. Eng.-Vision, Image Signal Process., vol. 152, no. 2, pp. 192–204, Apr. 2005.

[3] B. T. Morris and M. M. Trivedi, “A survey of vision-based trajectory learning and analysis for surveillance,” IEEE Trans. Circuits Syst. Video Technol., vol. 18, no. 8, pp. 1114–1127, Aug. 2008.

[4] X. Wang, “Intelligent multi-camera video surveillance: A review, ” Pattern Recognition Letters, vol. 34, pp. 3–19, 2013.

[5] M. Zabłocki, K. Gościewska, D. Frejlichowski, and R. Hofman, “Intelligent Video Surveillance Systems for Public Spaces - A Survey,” Journal of Theoretical and Applied Computer Science, 2014.

[6] M.L. Wang, C.C. Huang, and H.Y. Lin, “An intelligent surveillance system based on an omnidirectional vision sensor,” Proc. IEEE Conference on Cybernetics and Intelligent Systems, pp.1-6, 2006.

[7] X. Fan, K. Zheng, Y. Lin, and S. Wang, “Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1347-1355, 2015.

[8] E. Murphy-Chutorian and M. M. Trivedi, “Head pose estimation in computer vision: A survey,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 4, pp. 607–626, April 2009.

[9] T. Siriteerakul, “Advance in head pose estimation from low resolution images: A review,” International Journal of Computer Science Issues (IJCSI), val. 9-2, no.3, pp. 442-449, 2012.

[10] M. Patacchiola, A. Cangelosi, “Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods,” Pattern Recognition, vol. 71, pp. 132–143, 2017.

[11] B.H. Tran and Y.G. Kim, “Deep head pose estimation for faces in the wild and its transfer learning,” Seventh International Conference on Information Science and Technology (ICIST), pp. 1-10, 2017.

[12] M. Raza, Z. Chen, S.U. Rehman, P. Wang, and P. Bao, “Appearance based pedestrian’s head pose and body orientation estimation using deep learning,” Neurocomputing, vol. 272, pp. 647-659, 2018.

[13] B. Ahn, J. Park, and I.S. Kweon, “Real-time head orientation from a monocular camera using deep neural network,” in Asian Conference on Computer Vision (ACCV), pp. 82–96, 2015.

[14] L. Beyer, A. Hermans, and B. Leibe, “Biternion nets: Continuous head pose regression from discrete training labels,” GCPR, 2015

[15] X. Liu, W. Liang, Y. Wang, S. Li, and M. Pei, “3D head pose estimation with convolutional neural network trained on synthetic images,” In Proc. of IEEE International Conference on Image Processing, pp. 1289–1293, 2016.

[16] X. Xu and I. A. Kakadiaris, “Joint head pose estimation and face alignment framework using global and local CNN features,” In Proc. of IEEE Conference on Automatic Face and Gesture Recognition, 2017.

[17] S.S. Mukherjee and N. M. Robertson, “Deep head pose: Gaze-direction estimation in multimodal video,” IEEE Transactions on Multimedia, vol. 17, no. 11, pp. 2094–2107, 2015.

[18] N. Ruiz, E. Chong, and J.M. Rehg, “Fine-Grained Head Pose Estimation Without Keypoints,” arXiv preprint arXiv:1710.00925, 2017

[19] S. Ren, K. He, R. Girshick, and J. Sun. “Faster R-CNN: Towards real-time object detection with region proposal networks,” In NIPS, 2015.

[20] A. Krizhevsky, I. Sutskever, and G.E. Hinton. “ImageNet classification with deep convolutional neural networks,” In NIPS, 2012.

[21] K. Simonyan and A. Zisserman. “Very deep convolutional networks for large-scale image recognition,” In ICLR, 2015.

[22] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, “ImageNet large scale visual recognition challenge,” IJCV, 2015.

[23] J. Orozco, S. Gong, and T. Xiang, “Head pose classification in crowded scenes”, In Proc. BMVC, vol. 1, 2009, p. 3.