Invisible-to-Visible: Privacy-Aware Human Segmentation using Airborne Ultrasound via Collaborative Learning Probabilistic U-Net

Risako Tanigawa[1], Yasunori Ishii[1], Kazuki Kozuka[1], and Takayoshi Yamashita[2]

[1] Panasonic Holdings Corporation, 3-1-1 Yagumo-naka-machi Moriguchi City Osaka 570-8501, Japan
{tanigawa.risako,ishii.yasunori,kozuka.kazuki}@jp.panasonic.com
[2] Chubu University, 1200 Matsumoto-cho Kasugai-shi Aichi 487-8501, Japan
takayoshi@isc.chubu.ac.jp

Abstract. Color image enables highly accurate recognition such as segmentation, but it is difficult to protect privacy. We propose a new task for human segmentation from invisible information, especially airborne ultrasound. To generate human segmentation from ultrasound, we first convert ultrasound waves to reflected ultrasound images to perform segmentation from invisible information. Although an ultrasound image can roughly identify a person's location, the shape of the person in the image is ambiguous. To address this problem, we propose a collaborative learning probabilistic U-Net. This method is applicable to contour ambiguity. Experiments showed the effectiveness of the proposed method.

Keywords: Segmentation, Probabilistic U-Net, Ultrasound

1 Introduction

Segmentation has attracted wide applications [7, 1, 10]. Although camera-based segmentation has been widely investigated, camera images do not preserve privacy for human segmentation. Since the audio signal is invisible, it is effective for privacy protection [4]. However, the conventional method only deals with objects that emit sound. Therefore we propose segmentation using active sensing of ultrasound, which is the human inaudible range. Since this method performs active sensing, there is no need for a person to emit sound. In Figure 1, the region with strong intensity in the ultrasound image and the segmentation image are located close to each other, but the difference in contour is large. Therefore we propose a collaborative learning probabilistic U-Net (CLPU-Net), which is based on probabilistic U-Net (PU-Net) [6]. The CLPU-Net uses mean squared error (MSE) to minimize the distance between latent distributions for the optimization suitable for ultrasound images. Experiments showed that segmentation images can be generated from ultrasound images. To the best of our knowledge, this is the first work to estimate segmentation images from airborne ultrasound.

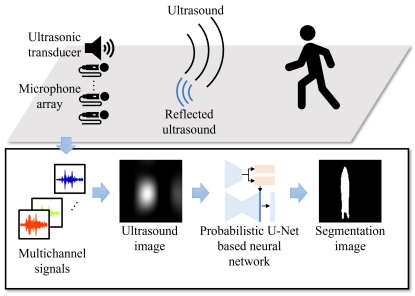

Fig. 1. The figure shows the experimental environment and our methodology.

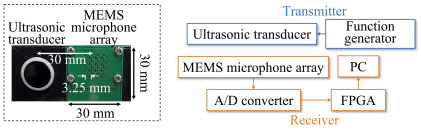

Fig. 2. Hardware setup of the ultrasound sensing system.

2 Related work

Privacy-preserved human segmentation methods have been proposed. A method of segmentation using wifi has been proposed [9]. However, the Wi-Fi signals are highly affected by the surrounding environment due to the multipath effect. Irie et al. [4] proposed a method that generates segmentation images from sounds. Their method can estimate human and environmental objects only from sounds. However, estimating segmentation images for non-sounding people is difficult in principle because this method analyzes the sound emitted from objects.

3 Proposed ultrasound sensing and data preprocessing

We develop hardware to detect people using ultrasounds as shown in Figure 2. The ultrasonic transducer was driven with burst waves of 20 cycles and 50 ms intervals at 62 kHz. A 4×4 grid MEMS microphone array, whose microphones were mounted on a 30 mm^2 substrate at 3.25 mm intervals, was used. The distance between the microphone array and ultrasonic transducer was set to 30 mm. Data preprocessing is as follows. First, a band-pass filter was used for signals captured by the 16 microphones. The filtered ultrasound signals were divided into blocks including single pair of direct and reflected waves. Following that, we produced ultrasound images from reflected ultrasounds via a delay-and-sum (DAS) [8] beamforming. We subtracted a reference map, which was the data without people, from the reflected directional heat maps to reduce the noise from reflected waves from objects other than people. In addition, the subtracted heat map was normalized from 0 to 1 when it was converted to ultrasound images X_{us}.

4 Human segmentation via ultrasound

Although the appearance of ultrasound and ground truth segmentation images differ significantly at the edges, the appearance of the other parts, particularly positions, is relatively similar (Input and output images of Figure 3). Therefore, it is important to learn latent space by focusing on the difference in edges. In PU-Net, prior distribution P and posterior distribution Q are penalized using a KLD. Because the ultrasound and segmentation images are roughly matched

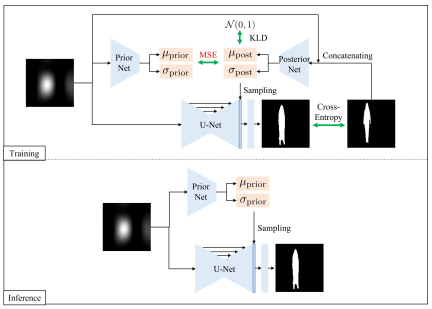

Fig. 3. Network architecture of CLPU-Net. The top row illustrates the training network and the bottom row illustrates the inference network.

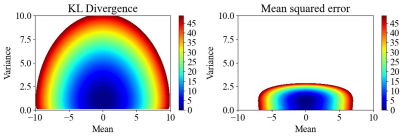

Fig. 4. Errors with standard normal distribution by KLD and MSE.

Table 1. IoU, accuracy, precision, recall, F1-score of conventional and proposed method

Model	IoU	Accuracy	Precision	Recall	F1-score
VAE	0.265	0.880	0.351	0.526	0.403
Joint-VAE	0.278	0.889	0.376	0.507	0.392
PU-Net	0.329	0.912	0.485	0.490	0.456
CLPU-Net	**0.388**	**0.921**	**0.536**	**0.546**	**0.519**

other than edges, reducing the distance between the distributions to estimate edges with high accuracy by comparing the distributions, which were obtained after the spatial dimensions decreased at the prior/posterior network, is difficult. Thus, we propose a method to use MSE of means and variances. The losses calculated by KLD and MSE are shown in Figure 4. The MSE loss was calculated as $\text{MSE} = (\mu - \mu_0)^2 + (\sigma - \sigma_0)^2$, where $\mu_0 = 0$ and $\sigma_0 = 1$ are the mean and variance. Since the value around the error of 0 for the MSE loss changes more rapidly than that of the KLD, the MSE loss is more sensitive than the KLD. The proposed network is illustrated in Fig. 3. The loss function L of the proposed method is $L = \alpha L_{\text{VAE}} + (1-\alpha)L_{\text{MSE}}$, where α is the weight adjusting the scale. The L_{VAE} and L_{MSE} are

$$L_{\text{VAE}} = \mathbb{E}_{z\sim Q(\cdot|X_{\text{seg}},X_{\text{us}})}\left[-\log P(X_{\text{seg}}|X_{\text{out}}(X_{\text{us}},z))\right] + \beta D_{\text{KL}}(P_0(z) \parallel Q(z|X_{\text{seg}},X_{\text{us}})), \quad (1)$$

$$L_{\text{MSE}} = \frac{1}{N}\left(\sum_{n=1}^{N}(\mu_{\text{prior},n} - \mu_{\text{post},n})^2 + \sum_{n=1}^{N}(\sigma_{\text{prior},n} - \sigma_{\text{post},n})^2\right), \quad (2)$$

where X_{seg} and X_{us} are the segmentation and ultrasound images, respectively, $X_{\text{out}}(\cdot)$ outputs the estimated segmentation image, z is the latent variables, β is the weight parameter, and $P_0(\cdot)$ is the probability density function of the standard normal distribution. N is the dimension of the latent vector.

5 Experiments

Datasets We created a dataset because no datasets have previously used airborne ultrasound to detect humans. For 10 s, we captured the ultrasounds from 16 channel microphones and videos at 30 frames per second (fps) from the RGB camera, which was located 35 mm under the microphone array. The resolution was 180×120 pixels, and the videos were used for creating ground truth. We used Mask R-CNN [3] for automatical annotation. We used the dataset that people, who were located from 1 to 3 m away from the sensing devices, performed

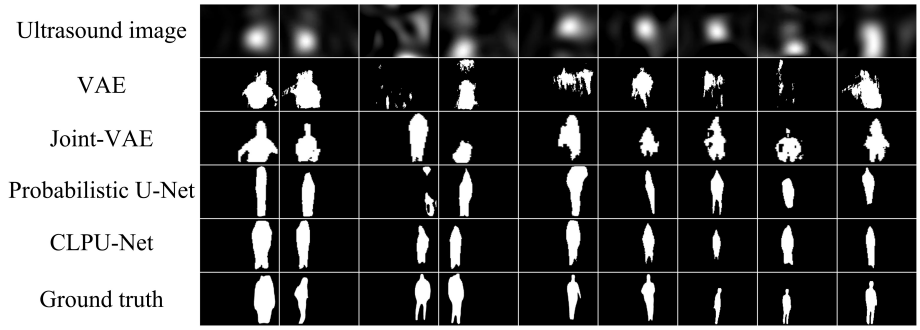

Fig. 5. Quantitative results. The top row is ultrasound images, the second to fourth rows are estimated images by conventional methods, the fifth row is estimated images by CLPU-Net, and the bottom row is ground truth.

motions such as standing, sitting, walking, and running. There were six participants and they performed in four different rooms. Images were captured about 7,700 images at each participant and the total number was 46,494.

Evaluation We evaluated the proposed method using k-fold cross-validation. To confirm the robustness of the unknown person data, the dataset was divided based on the participants. Therefore, k was set to six, and all six patterns were trained and evaluated. The performance of the model was evaluated using an intersection-over-union (IoU), accuracy, precision, recall, and F1-score. The structures and parameters of models are the same as the conventional method.

Experimental results We compared our model with PU-Net, variational auto-encoder (VAE) [5], Joint-VAE [2]. The VAE and Joint-VAE were trained using segmentation images and inferred using ultrasound images. The CLPU-Net and PU-Net were trained using ultrasound and segmentation images and inferred using ultrasound images. Table 1 illustrates the qualitative results. The CLPU-Net marked the highest performance in all metrics of the four models.

The quantitative result is shown in Figure 5. In the VAE and Joint-VAE, the shapes are not properly estimated. In these methods, the information on ultrasound images was not used during the training phase. Therefore, using the information in the segmentation images during the training phase affects the estimation. The PU-Net fails the estimation where the input and segmentation images have a large discrepancy. The estimated images of PU-Net tend to swell or shrink. In contrast, those images of CLPU-Net are closer to the ground truth.

6 Conclusions

We proposed privacy-aware human segmentation from airborne ultrasound using CLPU-Net. Our method used the MSE of the means and variances, which are the output of the prior and posterior networks in PU-Net. This enables optimization suitable for the ultrasound image obtained by our proposed device. This method can be used to detect human actions in situations where privacy is required, such as home surveillance.

References

1. Desouza, G., Kak, A.: Vision for mobile robot navigation: a survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **24**(2), 237–267 (2002)
2. Dupont, E.: Learning disentangled joint continuous and discrete representations. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems (NeurIPS). vol. 31. Curran Associates, Inc. (2018)
3. He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 2961–2969 (2017)
4. Irie, G., Ostrek, M., Wang, H., Kameoka, H., Kimura, A., Kawanishi, T., Kashino, K.: Seeing through sounds: Predicting visual semantic segmentation results from multichannel audio signals. In: International Conference on Acoustics, Speech, and Signal Processing (ICASSP). pp. 3961–3964 (2019)
5. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2014)
6. Kohl, S., Romera-Paredes, B., Meyer, C., De Fauw, J., Ledsam, J.R., Maier-Hein, K., Eslami, S.M.A., Jimenez Rezende, D., Ronneberger, O.: A probabilistic U-Net for segmentation of ambiguous images. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems (NeurIPS). vol. 31. Curran Associates, Inc. (2018)
7. Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., Terzopoulos, D.: Image Segmentation Using Deep Learning: A Survey. arXiv e-prints arXiv:2001.05566 (2020)
8. Perrot, V., Polichetti, M., Varray, F., Garcia, D.: So you think you can DAS? A viewpoint on delay-and-sum beamforming. Ultrasonics **111**, 106309 (2021)
9. Wang, F., Zhou, S., Panev, S., Han, J., Huang, D.: Person-in-WiFi: Fine-grained person perception using WiFi. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5451–5460 (2019)
10. Zhang, H.B., Zhang, Y.X., Zhong, B., Lei, Q., Yang, L., Du, J.X., Chen, D.S.: A comprehensive survey of vision-based human action recognition methods. Sensors **19**(5) (2019)