Automatic De-Identification System Using YOLOv7 and E^{2}FGVI in a Shot

Yeon-Seung Choo
Contents Convergence Research Center
Korea Electronics Technology Institute
Seoul, Korea
piksal@keti.re.kr

Hyun-Sik Kim
Contents Convergence Research Center
Korea Electronics Technology Institute
Seoul, Korea
hskim@keti.re.kr

Yong-Suk Park
Contents Convergence Research Center
Korea Electronics Technology Institute
Seoul, Korea
yspark@keti.re.kr

***Abstract*— In modern society, the demand for various forms of media, such as TV programs, dramas, movies, and YouTube, is ever-increasing. Unintended exposure of information that can be used for personal identification, such as faces, license plates, phone numbers, business names, etc., may occur during content source gathering. This may result in the invasion of privacy for individuals and business operation disruption for companies. As a result, media editing technology for de-identification and anonymization is becoming increasingly prominent. In this paper, a system that can anonymize personal information during media editing is presented. The proposed system first estimates the regional information of the detected targets in a given frame sequence of a shot. After estimation, it tracks target objects and perceives the regional information in the video signal. Finally, for seamless and natural anonymization within frames, video inpainting is performed to replace personal information completely. The proposed method can be applied in various media production processes, enabling convenient media editing.**

Keywords—de-identification, object detection, object tracking, object segmentation, video inpainting

I. Introduction

The advancement in semiconductor hardware has led to transmission performance increase, enabling real-time video streaming. The consumption of media content anytime and anywhere has brought rapid growth in the media industry. A vast amount of media content is being produced, and diverse media editing tools have emerged as well.

However, the growth of the media industry does not only have positive outcomes. It also has side effects, such as unintended exposure of personal information resulting in the invasion of privacy. Sensitive information, such as faces of individuals, license plate numbers, addresses, company brand logos, etc., may be captured from the background during content source acquisition and exposed in the final media output. The release of unconsented information in the media may lead to diverse legal and social problems. Consequently, the demand for the capability of media editing tools to de-identify or anonymize sensitive information is increasing. In this paper, we propose a de-identification system that uses deep learning-based computer vision, object detection and tracking, instance segmentation, and video inpainting to facilitate media editing for content creation.

II. Related Works

The de-identification of specific targets in video sequences may involve the following processes: object detection and segmentation to obtain the segmentation masks of the targets, tracking of targets in multiple frames, and applying video inpainting to the acquired segmentation masks.

The object detection process is required to identify and recognize the target object. Traditionally, object detection has been one of the most challenging tasks in computer vision. Various approaches have been proposed to detect objects, from hand-crafted feature-based approaches, such as Histogram of Gradients (HOG), to deep learning-based approaches, such as Regions with Convolutional Neural Networks features (R-CNN) and Single Shot Detector (SSD) [1-3]. These recent approaches have emerged with the latest developments in deep learning and have attracted much attention. Among the deep learning approaches, *J. Redmon et al.* proposed YOLO, which applied a 1-step object detection network in contrast to the conventional 2-step object detection approaches widely used at the time [4]. The 1-step architecture of YOLO not only enabled real-time processing but also improved object detection performance remarkably in terms of computational efficiency. After the appearance of the YOLO series, *C.-Y. Wang et al*. extended the original YOLO framework with a new network architecture for efficient control of gradient paths and designed model scaling based on network depth to enhance its efficiency, named YOLOv7 [5]. Object segmentation is also a challenging task in traditional computer vision. BlendMask, proposed by *H. Chen et al*., calculates object segmentation regions using extracted information during the detection process. In particular, BlendMask introduces a bi-directional approach that combines top-down from attention regions and bottom-up for score computation, obtaining better performance compared to traditional methods [6]. Additionally, object tracking, which

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. 2021-0-00804, Media Production Technology Using Learning Based Directing Methods).

tracks specific objects in frame sequences, is also one of the most prominent research areas, similar to object detection. Among them, DeepSORT, proposed by *N. Wojke et al.,* stands out for its performance [9]. DeepSORT utilizes continuity with Kalman filters to predict object positions in frames and enhances the traditional Simple Online and Real-time Tracking (SORT) method. SORT matches Kalman filter results with prediction results using the Hungarian algorithm, by incorporating motion and appearance information for comparison [7-8]. For the indistinctive removal of a tracked target object within the sequence signal, video inpainting is employed. Among the many inpainting approaches, *Z. Li et al.* proposed flow-guided video inpainting. This approach departs from the traditional serial, sequential framework by parallelizing the incorporation of local and non-local reference frames. This approach enables faster computation and improves robustness due to seriality [10].

By applying deep learning-based computer vision technologies, targets designated for de-identification can be detected, segmented, and tracked. Various image filters can be applied to the target areas for de-identification or video inpainting can be applied for seamless deletion of the specified targets.

III. Proposed Method

The proposed method detects the target specified using object detection and creates a segmented mask region by outlining the area occupied by the target using instance segmentation. The target is tracked in subsequent frames, and inpainting is applied to the segmented mask regions to seamlessly replace the target from the frames. Fig. 1 presents the overall framework of the proposed system.

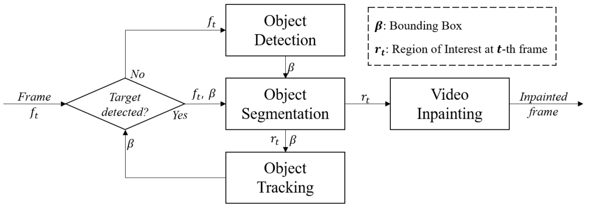
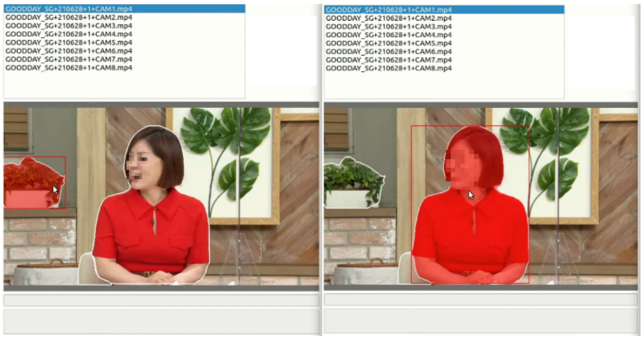


Fig. 1. The framework of the proposed method.

The system initially detects multiple target object within a single shot, $F = \{f_t\}_{t=1}^{T}$, which is composed of multiple frames T. The object detection process utilizes YOLOv7 which has been trained to detect target objects, such as people, license plates, phone numbers, etc. When the media editor selects a specific target to be de-identified from the set of detected objects, its bounding box is designated as β. A sample user interface for de-identification target selection is shown in Fig. 2. Blendmask-based object segmentation is used to obtain the segmentation mask of the de-identification target. Subsequently, a mask region r_t for de-identification is obtained for each t-th frame.

In the given frame sequence, the de-identification target is tracked using the bounding box region β. The system updates and predicts the location of the de-identification target in each frame by using DeepSORT, which tracks objects based on appearance and motion information of the predicted regions.



GOODDAY_SG+210628+1+CAM1.mp4
GOODDAY_SG+210628+1+CAM2.mp4
GOODDAY_SG+210628+1+CAM3.mp4
GOODDAY_SG+210628+1+CAM4.mp4
GOODDAY_SG+210628+1+CAM5.mp4
GOODDAY_SG+210628+1+CAM6.mp4
GOODDAY_SG+210628+1+CAM7.mp4
GOODDAY_SG+210628+1+CAM8.mp4

GOODDAY_SG+210628+1+CAM1.mp4
GOODDAY_SG+210628+1+CAM2.mp4
GOODDAY_SG+210628+1+CAM3.mp4
GOODDAY_SG+210628+1+CAM4.mp4
GOODDAY_SG+210628+1+CAM5.mp4
GOODDAY_SG+210628+1+CAM6.mp4
GOODDAY_SG+210628+1+CAM7.mp4
GOODDAY_SG+210628+1+CAM8.mp4

Fig. 2. Sample user interface for de-identification target selection.

To de-identify acquired target regions in frame n, the proposed system applies the E²FGVI algorithm by selecting local and non-local reference frames from the frame set $F_n = \{f_t\}_{t=n-\theta}^{n+\theta}$ and the regions of interest set $R_n = \{r_t\}_{t=n-\theta}^{n+\theta}$, with pre-defined neighbor-index parameter θ.

As a result, the proposed system enables convenient de-identification within frames of a single shot by obtaining regions of interest through object detection and segmentation processes and preserving regional information for tracking continuously in subsequent frames. Furthermore, by employing video inpainting, the system can seamlessly replace de-identification targets completely, offering a more elaborate de-identification option to the editor.

IV. Experimental Results

A. *Experimental Results*

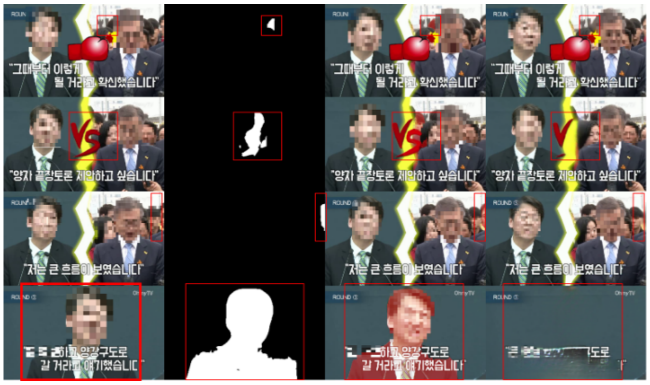

"그때부터 이렇게 될 거라고 확신했습니다"
"양자 끝장토론 제안하고 싶습니다"
"저는 큰 흐름이 보였습니다"

Fig. 3. Experimental result of the proposed system using pre-acquired media data.

To evaluate the performance of the proposed system, we conducted experiments using a server system equipped with an NVIDIA GeForce RTX 3090 GPU, an Intel i9-10900X CPU, and 64 GB RAM. Fig. 3 and Fig. 4 present the experimental results on pre-acquired shots. The experimental results demonstrate that the proposed method successfully accomplishes smooth video de-identification. In addition, we verified that real-time de-identification was feasible in a multi-processing environment. Figs. 5 and 6 illustrate these outcomes. The image columns in the experimental results show, from left to right, the input frames, acquired masks, combined images, and inpainting results, respectively.

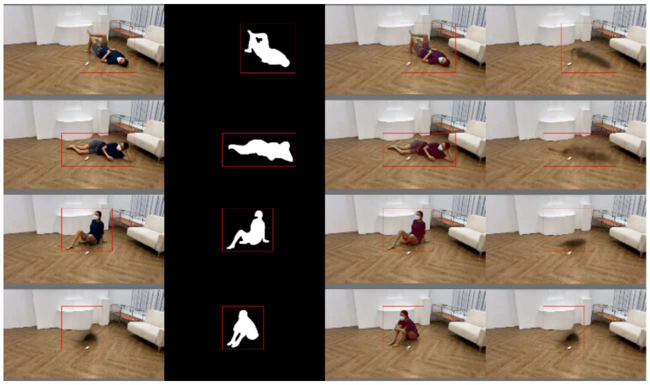


Fig. 4. Experimental result of the proposed system using AI-Hub [11] dataset.

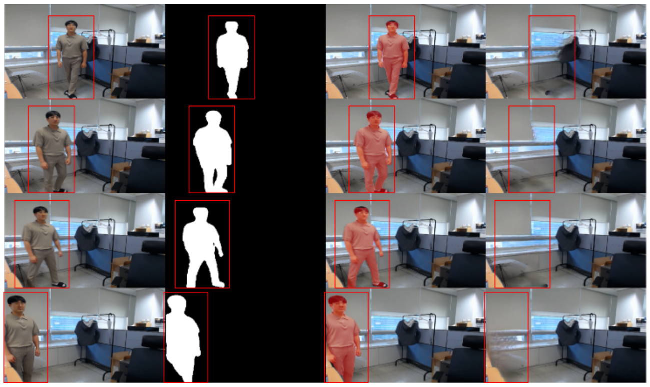


Fig. 5. Experimental result of the proposed system using webcam in-the-wild.

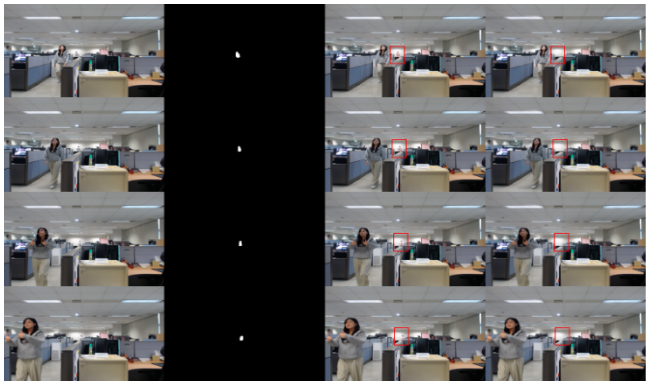


Fig. 6. Experimental result of the proposed system using webcam in the wild. The proposed system performs de-identification through object tracking even if another condidate targets are existed.

V. Conclusions

This paper proposes an easy-to-use de-identification system that can be used in media production environments, designed to prevent unintended exposure of personal identification information. The proposed system performs comprehensive flow-based de-identification of the target objects in frames of a single shot, using object detection and segmentation functions provided by YOLOv7. It then tracks the target de-identification objects in subsequent frames using DeepSORT and applies video inpainting based on E²FGVI algorithm for seamless information removal. Video editing with the proposed system shows effective de-identification results. Employing the proposed system for de-identification can effectively prevent the exposure of personal identification information during media production.

References

[1] N. Dalal and B. Triggs, “Histogram of Oriented Gradients for Human Detection,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 886-893, 2005.

[2] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “SSD: Single Shot MultiBox Detector,” in *European Conference on Computer Vision (ECCV)*, pp. 21-37, 2016.

[3] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation,” *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 580-587, 2014.

[4] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: unified, Real-Time Object Detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3142-3151, 2021.

[5] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, “YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7464-7475, 2023.

[6] H. Chen, K. Sun, Z. Tian, C. Shen, Y. Huang, and Y. Yan, “BlendMask: Top-Down Meets Bottom-Up for Instance Segmentation,” in *Proceedings of the IEEE/CVF Confere nce on Computer Vision and Pattern Recognition (CVPR)*, pp. 8573-8581, 2020.

[7] R. E. Kalman, “A New Approach to Linear Filtering and Prediction Problems,” in *Transactions of the ASME--Journal of Basic Engineering,* vol. 82, pp. 35-45, 1960.

[8] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, “Simple Online and Realtime Tracking,” in *2016 IEEE International Conference on Image Processing (ICIP)*, pp 3464–3468, 2016.

[9] N. Wojke, A. Bewley, and D. Paulus, “Simple Online and Realtime Tracking with a Deep Association Metric, ” in *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 3645-3649, 2017.

[10] Z. Li, C.-Z. Lu, J. Qin, C.-L. Guo, and M.-M. Cheng, “Towards an End-to-End Framework for Flow-Guided Video Inpainting,*” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17562-17571, 2022.

[11] AI Hub: Image data for Inpainting automation, 2023, ‘https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=487’