

# Automatically Protecting Privacy in Consumer Generated Videos Using Intended Human Object Detector

Yuta Nakashima  
Graduate School of  
Engineering, Osaka University  
2-1 Yamadaoka, Suita, Osaka  
565-0871, Japan  
nakashima@nanase.  
comm.eng.osaka-u.ac.jp

Noboru Babaguchi  
Graduate School of  
Engineering, Osaka University  
2-1 Yamadaoka, Suita, Osaka  
565-0871, Japan  
babaguchi@comm.  
eng.osaka-u.ac.jp

Jianping Fan  
Dept. of Computer Science  
UNC-Charlotte  
Charlotte, NC 28223, USA  
jfan@ncss.edu

## ABSTRACT

The growing popularity of video sharing services such as YouTube enables us to upload and share consumer generated videos (CGVs) easily, resulting in disclosure of the privacy sensitive information (PSI) of persons, i.e., their appearances. Therefore, we need a technique for automatically protecting the privacy in CGVs; however, the main problem is how to determine PSI regions automatically. In this paper, we propose a novel system for automatically protecting the privacy in CGVs. The proposed system tackles the problem of determining PSI regions by using an intended human object detector that detects human objects which the camera person wanted to capture to achieve his/her capture intention. In addition, the proposed system adopts several PSI obscuring methods such as blocking out, blurring and seam carving. We present the results of subjective evaluations of a privacy protected video in terms of the visual quality and acceptability of PSI disclosure, as well as the performance of the intended human object detector.

## Categories and Subject Descriptors

I.4.9 [Computing Methodologies]: Image processing and computer vision—*applications*

## General Terms

Algorithms, Design, Experimentation

## Keywords

Privacy protection, consumer generated videos, intended human object detector

## 1. INTRODUCTION

The growing popularity of video sharing services such as YouTube, Dailymotion, and so on enables us to upload con-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

MM'10, October 25–29, 2010, Firenze, Italy.

Copyright 2010 ACM 978-1-60558-933-6/10/10 ...\$10.00.

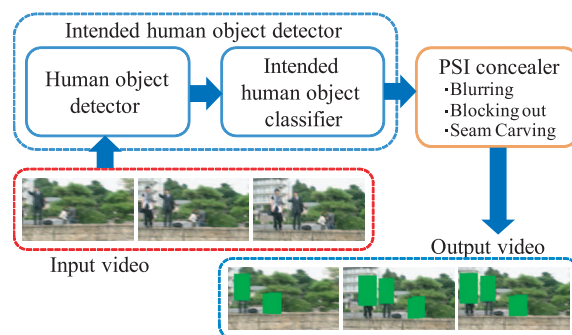


Figure 1: An overview of the proposed system for privacy protection in CGVs.

sumer generated videos (CGVs) very easily. However, this may disclose privacy sensitive information (PSI), i.e., the appearances, of people in these CGVs, and thus, the PSI regions in the CGVs must be obscured to prevent the CGVs from violating the privacy. In conventional techniques, PSI regions are manually specified, and are blurred or blocked out. However, this is a time-consuming task for human operators, and we need a technique which determines and obscures PSI regions automatically.

For static camera applications such as video surveillance and video conference, many techniques for automatically obscuring PSI regions have been proposed. For example, Yu et al. proposed a system called PriSurv [8] which adaptively changes the method for obscuring PSI regions according to the relationships between the viewer and the subjects. Tansuriyavong and Hanaki proposed a system to silhouette people in videos [6]. Their system can display the names of the people using a face recognition technique. Venkatesh et al. proposed a video conference system which obscures PSI in both video and audio [7]. For mobile camera, Kitahara et al. proposed Stealth Vision [2]. Stealth Vision determines PSI regions in a video taken by mobile cameras using static cameras installed in the environment.

In most of such techniques, moving objects are extracted as human objects using background models, and the human objects are identified by, for example, a face recognition technique. Whether each human object is privacy sensitive or not is determined based on the identity of the human object. In the case of the CGVs, however, most of which are taken by camera persons with mobile cameras, constructing background models or human identification are very tough

problems. Furthermore, because the camera persons have capture intentions [3], i.e., what they want to express in their videos, PSI regions should be determined with considering their capture intentions so that the privacy protected videos can maintain them.

In this paper, we focus on that there are intended and unintended human objects in most of CGVs. Intended human objects are defined as human objects which a camera person wants to capture according to his/her capture intention, and therefore, they are essential for maintaining the capture intention. Unintended human objects are all other human objects. Considering this, we propose a system for automatically obscuring PSI regions by applying a PSI obscuring method such as blocking out, blurring, or seam carving. The main contributions of this paper are as follows:

- The intended human object detector is employed for determining PSI regions so that the proposed system can maintain camera persons' capture intentions. To the best of our knowledge, this is the first work to address privacy protection for videos which takes the capture intentions into account.
- We present the result of subjective evaluation of a privacy protected video in terms of visual quality, considering that the visual quality of the privacy protected videos is an important aspect for CGVs.

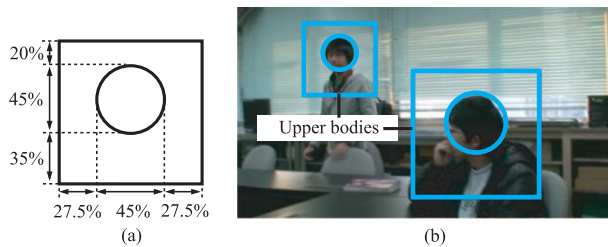
The rest of this paper is organized as follows: The intended human object detector and the PSI obscuring methods are described in Section 2. In Section 3, experimental results are presented. Section 4 concludes this paper.

## 2. A SYSTEM FOR AUTOMATICALLY OBSCURING PSI REGIONS

An overview of the proposed system is illustrated in Figure 1. The proposed system first detects human objects in an input video and classifies them into intended/unintended using the intended human object detector [4]. Finally, a PSI obscuring method is applied to the regions of the unintended human objects, and the intended human objects are left as they are. This is reasonable because of the following two reasons: 1) The intended human objects are essential to present what the camera person wanted to express, and, in many cases, the consent for uploading the video can be obtained from the persons corresponding to the intended human objects. 2) The unintended human objects including persons who accidentally framed in are not important for the capture intention and the consent of such persons is hardly obtained. In the following sections, we describe the intended human object detector and PSI obscuring methods.

### 2.1 Intended Human Object Detector

The intended human object detector first detects human objects by a human object detector, which detects upper bodies defined as in Figure 2, using a support vector machine (SVM) and the histograms of oriented gradients. Then, each of the detected human objects is classified into intended or unintended using SVMs and features related to the camera motion and the visual attention. This is because that whether a human object is intended or not is reflected in the camera motion against the motion of the human object, and that the visual attention which can be modeled by [1] also affects the process for formulating the capture intention.



**Figure 2: The definition of upper body regions. An upper body is defined as a region surrounded by the rectangle in (a) when (a) is placed in the video frame so that the circle in (a) surrounds the head of the human object as in (b).**

For the  $i$ -th human object ( $i = 1, \dots, N_H^t$ ) in the  $t$ -th video frame ( $t = 1, \dots, N_T$ ), the intended human object detector outputs the center position  $(x_i^t, y_i^t)$  and the length  $s_i^t$  of each side of the square detection window together with the label  $l_i^t \in \{-1, +1\}$  where  $-1$  and  $+1$  stand for unintended and intended, respectively.

### 2.2 PSI Obscuring Method

In the proposed system, we employ blocking out, blurring, and seam carving as PSI obscuring methods.

#### 2.2.1 Blocking Out

Blocking out is one of the most simple methods which blocks out the unintended human objects as PSI regions as in Figure 4(a). Although blocking out completely removes the PSI regions, this can cause severe visual artifacts on video frames. A PSI region to be blocked out is an upper body region specified by a square region of which center is at  $(x_i^t, y_i^t)$  and the length of each side is  $s_i^t$ . To obscure the most part of human objects, we also employ blocking out on expanded regions as shown in Figure 4(b). The shape of the expanded region is determined based on the shape prior shown in Figure 3(a).

#### 2.2.2 Blurring

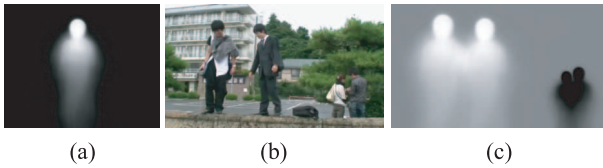
Blurring obscures PSI regions with less visual artifacts by applying the smoothing filter with a  $K_B \times K_B$  uniform kernel. However, colors of the clothes, hair, and skin can be disclosed. We adopt the blurring on both upper body regions and expanded regions as shown in Figure 4(c) and (d), respectively.

#### 2.2.3 Seam Carving

Seam carving for videos proposed by Rubinstein et al. [5] is another option for PSI obscuring. This can remove PSI regions completely with small visual artifacts by removing vertical seams defined as manifolds in the spatio-temporal volume of the video. The seams are the minimum cuts of an appropriately constructed graph obtained by the graph cuts algorithm.

To remove unintended human objects using the seam carving, we first generate a intention map  $IM^t(x, y)$  for each video frame as shown in Figure 3(c) based on outputs of the intended human object detector by

$$IM^t(x, y) = \sum_{i=1}^{N_H^t} l_i^t SP \left( \frac{x - x_i^t}{s_i^t}, \frac{y - y_i^t}{s_i^t} \right), \quad (1)$$



**Figure 3: The shape prior (a). An example frame (b) and its intention map (c). In (c), white and black pixels represent positive and negative values, respectively.**

where  $SP(x, y)$  is the shape prior (Figure 3(a)). To maintain the spatial consistency in the video frame, we also calculate the difference of adjacent pixels  $D^t(x, y)$  by

$$D^t(x, y) = \sum_{c \in \{R, G, B\}} |I_c^t(x, y) - I_c^t(x + 1, y)| + \sum_{c \in \{R, G, B\}} |I_c^t(x, y) - I_c^t(x, y + 1)| \quad (2)$$

where  $I_c^t(x, y)$  is the red ( $c = R$ ), green ( $c = G$ ) or blue ( $c = B$ ) component of the pixel at  $(x, y)$ . A weight  $w^t(x, y)$  is calculated by

$$w^t(x, y) = \frac{\alpha}{1 + e^{-IM^t(x, y)}} + D^t(x, y) \quad (3)$$

where  $\alpha$  is a constant to control the contribution of the intention map. The first term in the left hand side of (3) gives small values around unintended human objects. Therefore, seams tend to concentrate around the unintended human objects, resulting in removal of them. This weight is used as  $E_1$  in [5]. The number of seams to be removed,  $N_R$ , is manually specified by a user.

The seam carving for video [5] imposes constraints on the graph so that seams are connected along the time axis in order to reduce unnatural motions and jerkiness due to removals of temporally inconsistent seams. However, these constraints can lead to removals of inappropriate seams if the camera or an object is in a rapid motion because the seams can move with only one pixel for each video frame. Hence, we divide the video into a sequence of video segments consisting of  $N_S$  video frames each of which overlaps each other by  $N_S - 1$  video frames, and the seam carving described above is applied to each video segments. Then, the middle video frame in each of the video segments is extracted to generate an output video. This can prevent the removals of inappropriate seams if the video segments are sufficiently short since the constraints for temporal consistency are relaxed. At the same time, the unnatural motions and jerkiness are expected to be alleviated by overlapping video segments.

### 3. EXPERIMENTAL RESULT

We objectively evaluated the performance of proposed system. In addition, we also subjectively evaluated the proposed system in terms of visual quality and acceptability of PSI disclosure. The parameters used in our experiments are  $K_B = 20$ ,  $\alpha = 2.5$ ,  $N_R = 107$ , and  $N_S = 5$ .

The video data set used in the objective evaluation contains 20 videos and 31838 frames in total. The frame rate is 30 frames per second. The number of human objects is 55310 where the numbers of intended and unintended human objects are 37544 and 17766, respectively. Our upper

body detector correctly detected 56% of the human objects while giving 1.14 of false positives per video frame. The correctly detected human objects consist of 68% of the intended human objects and 31% of the unintended human objects. Therefore, PSI disclosure rate (PDR) was 69%, which stands for that a PSI obscuring method is not applied to 69% of the unintended human objects. This includes 45% of the unintended human objects which were incorrectly classified as intended. On the other hand, incorrectly obscured regions per frame (IORPF) was 0.63, including 6.9% of the intended human objects. From these results we need to improve the performance of the intended human object detector.

The video used in the subjective evaluation was excerpted from a video in the data set. The size and length of the video are  $427 \times 240$  pixels and 120 frames, respectively. The number of upper bodies is 360 where the numbers of intended and unintended human objects are 120 and 240, respectively, and 52% of the human objects were correctly detected where false positives per video frame was 1.34. PDR was 23% including 17% of the unintended human objects which were incorrectly classified as intended, while IORPF was 0.72 including 0.83% of intended human objects which were incorrectly obscured. PDR for this video was low compare to that for the all videos in the data set. We subjectively evaluated the following two cases: 1) The upper body detector and the intended human object detector were used (AUTO). This represents the actual performance of our implementation of the proposed system. 2) The upper bodies were manually specified and classified into intended/unintended by the camera person who captured the video (MANU). This represents the performance in the case where the upper body detector and the intended human object classifier give the ideal performance.

To subjectively evaluate the visual quality, we generated privacy protected videos using the five PSI obscuring methods, i.e., blocking out (BO), blocking out for expanded regions (EBO), blurring (BL), blurring for expanded regions (EBL), and seam carving (SC). Example frames of the privacy protected videos for AUTO and MANU are shown in Figures 4 and 5, respectively. Then, we asked ten subjects to assign a score ranging from 1 (lowest visual quality) to 5 (highest visual quality) to each privacy protected video. We set the original video as a reference of the highest visual quality. The means and standard deviations of the scores are shown in Figure 6. In the case of AUTO, blurring both on the upper body regions and the expanded regions gave prominent scores while the visual quality of blocking out and seam carving was low. One of the reasons is temporal inconsistency of the intended human object detector. That is, for blocking out, the temporal inconsistency caused frequently intermissive blocking out. For seam carving, the temporal inconsistency prevented from finding appropriate seams, and led to distinct discontinuities as in Figure 4(e).

To evaluate acceptability of PSI disclosure, we also asked the subjects to assign a score ranging from 1 (unacceptable PSI disclosure) to 5 (acceptable PSI disclosure). In this case, we set the original video as a reference of the unacceptable PSI disclosure. The means and standard deviations of the scores are shown in Figure 7. In the case of AUTO, the scores of the five PSI obscuring methods were relatively low and the differences among them were small. This may imply that the impact of the unintended human objects which were incorrectly disclosed was a dominant factor for the subjective

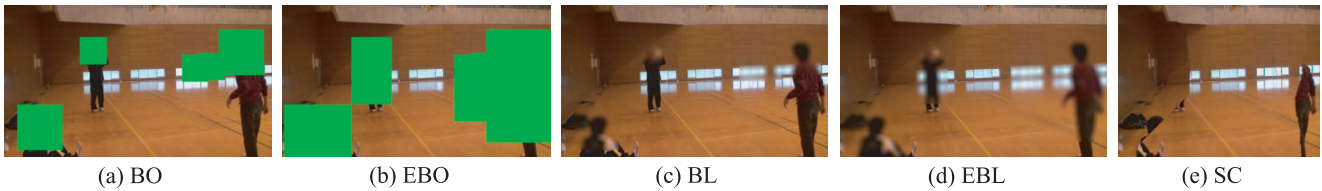


Figure 4: Example frames of a privacy protected video in the case of AUTO.

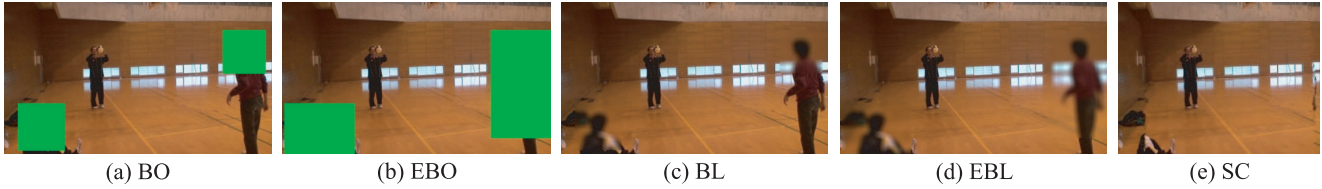


Figure 5: Example frames of a privacy protected video in the case of MANU.

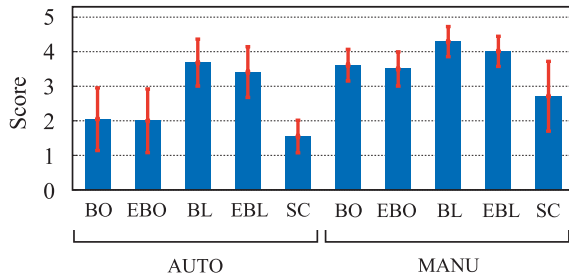


Figure 6: The means and standard deviations of subjective scores for visual quality.

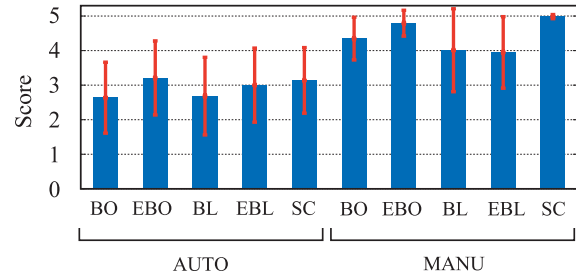


Figure 7: The means and standard deviations of subjective scores for acceptability of PSI disclosure.

scores, and the difference among PSI obscuring methods did not give much influence on the scores. This result also lead to the conclusion that improving the performance of the intended human object detector is crucial for the proposed system. In the case of MANU, the PSI disclosure when seam carving was used was almost completely acceptable. The reason is that seam carving can remove almost all of the pixels in unintended human objects.

#### 4. CONCLUSION

In this paper, we propose a system for automatically obscuring privacy sensitive information (PSI) in consumer generated videos. The proposed system uses an intended human object detector to determine PSI regions so that the camera persons' capture intentions can be maintained in the privacy protected videos. Our objective evaluation indicated that the proposed system obscured 31% of PSI regions while 0.63 regions per frames were incorrectly obscured. A subjective evaluation indicated that blurring gave the least impact on the visual quality. Although the acceptability of PSI disclosure was not sufficient due to the low accuracy of the intended human object detector, we consider that improving the performance of the intended human object detector enables the proposed system to be an alternative for manual privacy protection. This work is partly supported by a Grant-in-Aid for Scientific Research from the Japan Society for the Promotion of Science (JSPS) and a Grant-in-Aid for JSPS fellows.

#### 5. REFERENCES

[1] L. Itti, C. Koch, and E. Niebur. A model of

saliency-based visual attention for rapid scene analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, November 1998.

[2] I. Kitahara, K. Kogure, and N. Hagita. Stealth vision for protecting privacy. In *Proc. of 17th Int'l Conf. on Pattern Recognition (ICPR 2004)*, volume 4, pages 404–407, August 2004.

[3] T. Mei, X.-S. Hua, H.-Q. Zhou, and S. Li. Modeling and mining of users' capture intention for home videos. *IEEE Trans. on Multimedia*, 9(1):66–77, January 2007.

[4] Y. Nakashima, N. Babaguchi, and J. Fan. Detecting intended human objects in human-captured videos. In *Proc. of Conf. on Computer Vision and Pattern Recognition Workshop (CVPRW 2010)*, pages 1–8, 2010.

[5] M. Rubinstein, A. Shamir, and S. Avidan. Improved seam carving for video retargeting. *ACM Trans. on Graphics (SIGGRAPH)*, 27(3):1–9, 2008.

[6] S. Tansuriyavong and S. Hanaki. Privacy protection by concealing persons in circumstantial video image. In *Proc. of the 2001 Workshop on Perceptive User Interfaces*, pages 1–4, November 2001.

[7] M. V. Venkatesh, J. Zhao, L. Profitt, and S. S. Cheun. Audio-visual privacy protection for video conference. In *Proc. of the 2009 IEEE Int'l Conf. on Multimedia and Expo (ICME 2009)*, pages 1574–1575, July 2009.

[8] X. Yu, K. Chinomi, T. Koshimizu, N. Nitta, Y. Ito, and N. Babaguchi. PriSurv: Privacy protecting visual processing for secure video surveillance. In *Proc. of Int'l Conf. on Image Processing (ICIP 2008)*, pages 1672–1675, October 2008.