PAPER

Privacy Protection for Social Video via Background Estimation and CRF-Based Videographer's Intention Modeling

Yuta NAKASHIMA[†a)], ***Member***, **Noboru BABAGUCHI**[††], ***Fellow***, ***and*** **Jianping FAN**[†††], ***Nonmember***

SUMMARY The recent popularization of social network services (SNSs), such as YouTube, Dailymotion, and Facebook, enables people to easily publish their personal videos taken with mobile cameras. However, at the same time, such popularity has raised a new problem: video privacy. In such social videos, the privacy of people, *i.e.*, their appearances, must be protected, but naively obscuring all people might spoil the video content. To address this problem, we focus on videographers' capture intentions. In a social video, some persons are usually essential for the video content. They are intentionally captured by the videographers, called intentionally captured persons (ICPs), and the others are accidentally framed-in (non-ICPs). Videos containing the appearances of the non-ICPs might violate their privacy. In this paper, we developed a system called BEPS, which adopts a novel conditional random field (CRF)-based method for ICP detection, as well as a novel approach to obscure non-ICPs and preserve ICPs using background estimation. BEPS reduces the burden of manually obscuring the appearances of the non-ICPs before uploading the video to SNSs. Compared with conventional systems, the following are the main advantages of BEPS: (i) it maintains the video content, and (ii) it is immune to the failure of person detection; false positives in person detection do not violate privacy. Our experimental results successfully validated these two advantages.

key words: *intentionally captured person, conditional random field, background estimation, privacy protection, social video*

1. Introduction

Recently, many personal videos taken with mobile cameras are being uploaded on such social network services (SNSs) as YouTube and Facebook, where they are widely navigated and viewed. Generally, such social videos contain the appearances of persons, which are crucial privacy sensitive information. Since disclosing such privacy sensitive information may seriously violate the privacy rights of others, systems are required that obscure the appearances of persons.

Many systems for video privacy protection have been proposed for applications with specific tasks, such as video surveillance [1]–[4], Google Street View [5], [6], and lifelogging [7]. These systems basically detect all or a predetermined set of persons and obscure them using such appearance obscuration methods as blurring and blocking out. The fundamental idea underlying these systems is that whether the persons captured in the videos are shown to viewers is solely determined based on whether they grant permission. Some systems strictly adhere to this idea and allow the individuals to choose whether to be shown [4], [8], [9], but this requires special infrastructure to identify persons.

Systems solely based on such an idea can suffice for various tasks, but they are not suitable for protecting privacy in social videos for two reasons: (i) Social videos are usually taken by videographers who have a strong intention for capturing the videos, *e.g.*, recording a child's play or a friend's wedding. Such intentions determine persons who are essential for the videos, and videographers intentionally capture them. We call them intentionally captured persons (ICPs). Obscuring all or a predetermined set of persons without considering the ICPs significantly spoils the video content. (ii) In social videos, visual quality is an essential factor for enhancing the viewers' experience; how persons are obscured is thus crucial for privacy protection.

Considering these differences, we summarize the requirements for a privacy protection system for social videos, besides completely obscuring the appearances of persons without permission, as follows:

1. It must present ICPs in privacy protected videos.
2. It must provide various methods for obscuring the appearances of persons.

Requirement 1) confirms that privacy protected videos retain the content of their original. This requirement might cause conflict between videographer intentions and ICPs' privacy because ICPs are determined solely based on videographer intentions but the fundamental idea underlying privacy protection systems does not allow this. Automatically resolving it is impossible without negotiation between the videographer and the ICPs. Even if the conflict is resolved, we still need a technique to locate ICPs or non-ICPs in the videos for automatic privacy protection. Requirement 2) allows the users to choose a suitable appearance obscuration method to maintain the visual quality. This requirement is demanding because flexible obscuration in, *e.g.* [4], requires background images, which are usually not available for videos taken by mobile cameras.

In our previous work [10], [11], we developed a system that fulfills requirement 1) by selectively obscuring non-ICPs under the assumption that no conflict exists between videographer intentions and the ICPs' privacy. For selective obscuration, our previous system adopts non-ICP

Manuscript received September 19, 2015.
Manuscript revised December 12, 2015.
Manuscript publicized January 13, 2016.

[†]The author is with Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma-shi, 630–0192 Japan.
[††]The author is with Graduate School of Engineering, Osaka University, Suita-shi, 565–0871 Japan.
[†††]The author is with Department of Information Science, University of North Carolina, USA.
a) E-mail: n-yuta@is.naist.jp

DOI: 10.1587/transinf.2015EDP7378

detection [12] that detects all of the persons and classifies them into ICPs or non-ICPs to obscure only the latter group (Fig. 1 (a)). Unfortunately, this system still suffers from a limited number of obscuration methods; it can only provide blocking out, blurring, and seam carving, all of which can be applied without using background images. In addition, experimental results revealed the difficulty in detecting non-ICPs because they are not essential for the video, and videographers do not pay much attention to them. This results in too small or partially captured non-ICPs. In our previous system, such detection failure directly leads to the disclosure of appearances, making it inadequate for video privacy protection.

In this paper, we propose a system for social video privacy protection, called BEPS (Background Estimation-based Privacy protection system for Social videos), which is an extension of our previous system [13]. BEPS takes a novel approach for video privacy protection that estimates the background image of each frame, finds ICPs to be presented, and superimposes them onto the background image to maintain the video content (Fig. 1 (b)). Superimposition of ICPs enables BEPS to present them for requirement 1), and background estimation offers various obscuration methods, as in Fig. 2, for requirement 2). BEPS also overcomes the problem of sensitivity to detection failure in our previous system. Background estimation generates background images that contain no person; therefore, explicit detection of non-ICPs is no longer necessary. This makes BEPS's privacy protection capability immune to non-ICP detection failure.

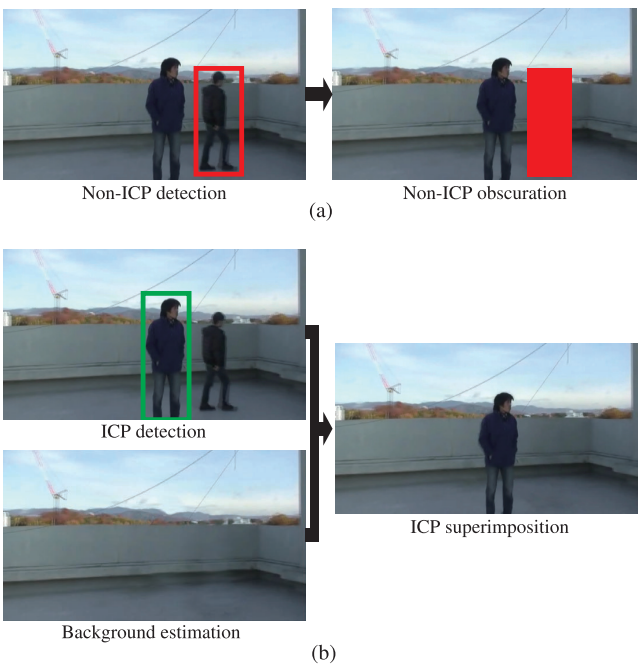


Fig. 1 Differences between (a) our previous system [10], [11] and (b) BEPS in their approaches to privacy protection.

The following are our contributions:

- We develop a novel system for social video privacy protection called BEPS assuming that the conflict between the videographer and the ICPs is resolved by negotiation. By using ICP detection and background estimation, BEPS automatically generates a privacy protected video without any additional burden on videographers. This is a new research direction toward video privacy protection.
- We introduce new preprocessing and frame selection stages for background estimation to make BEPS applicable to longer videos than our previous system [13]. Our experimental results demonstrate that these stages actually achieve this without sacrificing BEPS's performance.
- We propose a new ICP detection method using the conditional random field (CRF). Compared to our previous method [10], [12], our CRF-based model encodes the spatial relationship among the persons. Our experimental results indicate performance gain over the previous methods.

The rest of this paper is organized as follows: The next section introduces related work. Section 3 presents an overview of BEPS and justifies the assumption that no conflict exists between videographer intentions and the privacy of ICPs. The details of the methods used in BEPS are described in Sects. 4–6. In Sect. 7, our experimental results are illustrated. Section 8 gives discussion, and we conclude this paper in Sect. 9.

2. Related Work

Video privacy protection has been extensively studied for fixed camera applications, especially for video surveillance. Tansuriyavong and Hanaki [1] silhouetted persons, whose names were displayed based on a face recognition technique. Chen *et al.*'s system [2] obscures a person's appearance while preserving his/her shape and motion for surveillance purposes. Dufaux and Ebrahimi [3] provided scrambled video for unauthorized viewers, but for authorized

Fig. 2 Some Example of appearance obscuration. From left to right: original, see-through, dot, edge, and avatar, some of which are from [4].

viewers, the original video was restored from the scrambled one. Chinomi *et al.*'s system [4], called PriSurv, adaptively applies various appearance obscuration, such as blurring, blocking out, and complete removal, based on the relationship between the person and the viewer. Mitsugami *et al.* [14] proposed to replace persons to icons, which may not cover the persons' entire silhouette. For this, they synthesized a background image for a fixed camera and superimposed the icons on it. To handle different lighting conditions, their background images are synthesized from long-term observation of the environment.

The privacy issue of Google Street View has also been addressed [5], [6]. Frome *et al.* [5] proposed a technique to increase the recall rate of face detection to avoid disclosing faces. Flores and Belongie [6] replaced persons with corresponding regions in other Google Street View images captured at different positions.

For videos taken with mobile cameras, Kitahara *et al.* proposed a system called Stealth Vision [15], which applies pixelization to persons. To locate persons in a mobile camera's frame, their system uses fixed cameras installed in the target environment. Brassil's system [8] obscures individuals with a special device for locating them, and he argued that his system was applicable to mobile cameras. It projects the person's position in the target environment onto the view of a camera. Chaudhari *et al.* [7] blocked out faces in videos by a wearable life-log system. YouTube obscures persons in uploaded videos based on person detection [16].

Another interesting direction for video privacy protection research is knowledge discovery while preserving privacy. Fan *et al.* [17] and Peng *et al.* [18] developed methods for statistical inference from video collections distributed among multiple parties without disclosing the original videos to the other parties.

Privacy protection systems offer various types of privacy obscuration for fixed cameras, *e.g.*, blurring, showing a dot at the person's position, or even complete removal [19], because the background images are available. For mobile cameras, however, the difficulty in obtaining the background prevents privacy protection systems from using such obscuration as complete removal. Thus, BEPS adopts background estimation for mobile cameras to gain more flexibility.

In addition, most existing systems for video privacy protection preliminarily determine those to be obscured. Some obscure all persons (*e.g.* [5]–[7]), and others use a database and such devices as RFID readers and tags for identification (*e.g.* [4], [8], [15]); this approach completely ignores videographer intentions. Using such devices makes these systems impractical, and ignoring intentions can detract from the video content (Fig. 3). In contrast, BEPS uses ICP detection [12] for determining the individuals to be presented, which makes BEPS practical and suitable for social videos.

The existing systems for video privacy protection detect persons to be obscured and apply appearance obscuration methods. With this approach, the detection failure immediately discloses their appearances. By defining a criterion to measure privacy loss in videos, Saini *et al.* experimentally demonstrated that obscuring persons based on their detection is not reliable due to detection failure [20]. Therefore, research has addressed the reduction of detection failure [5]. Since BEPS takes a very different approach that superimposes those to be presented on estimated background images, the generated privacy protected videos are immune to detection failure. In other words, BEPS can be viewed as a system that makes a privacy protected background image of the entire video frame regardless of the persons' presence and selectively reveals the ICPs by superimposing them on the estimated background. This approach basically shares the same idea as the conclusion of [20], where global obscuration is preferable to obscuration based on person detection in terms of privacy loss.

Fig. 3 Examples of original frame (left) and blocked out video frames without considering videographer's intention (right).

BEPS can be deemed an extension of our previous system [10], [11], as it provides various privacy obscuration (Fig. 2) as well as complete removal. BEPS is also an extension of the system [13]. To improve ICP detection accuracy, we developed an ICP model, which leverages ICPs' temporal and spatial relationship using the CRF. This new model is similar to one in previous work [21] for real-time applications, but we introduce richer features and a model at the cost of computational complexity to improve the accuracy. In addition, the system [13] is not applicable to long video because of its severe memory consumption during background estimation. Our preprocessing and frame selection stages relax this problem by reducing the number of frames used in background estimation.

Another extension of the system [10], [11] is real-time privacy protection [22], which is crucial for live streaming from mobile cameras. To achieve real-time processing, it is constrained from various aspects, including applicable features for distinguishing ICPs from non-ICPs and privacy obscuration methods. In contrast, since BEPS's target is videos to be uploaded to SNSs, the real-time requirement is not imposed on BEPS.

3. System Overview

Most existing systems for video privacy protection (*e.g.*, [4], [8]) obscure all persons in a video or provide each person in it an opportunity to consent to be presented in the privacy protected video based on the fundamental idea for privacy protection mentioned in Sect. 1. This idea potentially conflicts with videographer intention. However, especially regarding videos uploaded to SNSs, we make the following

observations:

- In most cases, ICPs are the videographer's friends or family. For them, the videographer can easily obtain permission to capture and upload the video.
- Since non-ICPs are generally passers-by, obtaining their permission is practically impossible.

This means the potential conflict between videographer intention and the fundamental idea for privacy protection can be resolved easily by negotiation between videographers and ICPs, where there is no conflict between videographers and non-ICPs because non-ICPs are obscured. BEPS generates privacy protected videos based on ICP detection, assuming no conflict.

Thus, BEPS is not designed to be a fully automatic system that strictly follows the idea behind privacy protection. Rather, it is a tool for supporting videographers, who are compliant to that idea, by automatically generating privacy protected videos. The following shows the flow that videographers go through to upload their video: (i) Obtain permission to upload it to SNSs from the ICPs who appear in it. (ii) Obscure all non-ICPs (i.e., people irrelevant to the video content) using BEPS so that they can upload the video without permission from the non-ICPs. (iii) Upload the video only when all ICPs in the video allow them to do so.

Figure 4 shows an overview of BEPS. First, BEPS estimates the background of frames in an input video using the method [23]. Since it is designed for a small number of images and computationally expensive, we perform preprocessing and frame selection stages to extract the frames suitable for background estimation. For each frame in the input video, we synthesize its background image from the estimated background for selected frames. After detecting the ICPs, BEPS extracts the ICPs using the graph cuts algorithm [24]. Finally, a privacy protected video is generated by superimposing the ICPs onto the synthesized background image. The following provides their details.

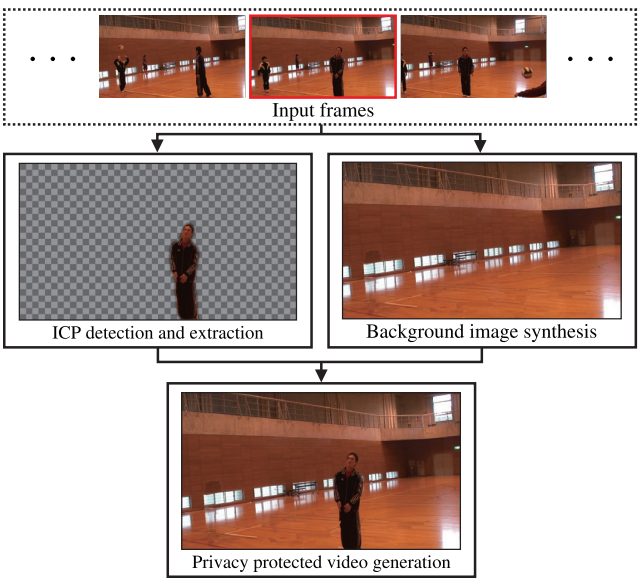


Fig. 4 Overview of BEPS. Privacy protected video frame is generated for the frame in input frames enclosed by red.

4. Background Estimation

In BEPS, we adopt Kim *et al.*'s method for its applicability to the images from a moving camera; however, their method is computationally infeasible when it is applied to long videos. Thus, observing that successive frames in a video resemble each other and hardly contribute to the background estimation, we introduce preprocessing and frame selection stages to reduce the number of frames used for background estimation.

Figure 5 shows an overview of our background estimation. The preprocessing stage clusters the frames and extracts a representative frame for each cluster. Then, given a target frame for which the background is estimated, the frame selection stage further extracts a subset of representative frames that are suitable for it. In the estimation stage, the algorithm based on [23] estimates the target frame's background using that subset. We further reduce the computational burden by applying background estimation only to representative frames and synthesizing the other frames from them.

Preprocessing stage. Although there are many clustering algorithms, *e.g.*, k-means [25] and affinity propagation [26], we develop a simple clustering algorithm to confirm that each resulting cluster consists of consecutive frames to simplify the synthesis of non-representative frames.

Our clustering algorithm is provided in Algorithm 1. Given set of frames in a video $\{F_t | t = 1, 2, \ldots, N_{\text{Frame}}\}$, the algorithm outputs set of clusters $S = \{C_k | k = 1, 2, \ldots, N_{\text{Clusters}}\}$, where C_k is the k-th cluster. N_{Frame} and N_{Cluster} are the numbers of frames in the video and clusters, respectively. The basic idea is building a cluster that consists of consecutive frames that resemble the representative frame. Let C_k and t be a current set of consecutive frames

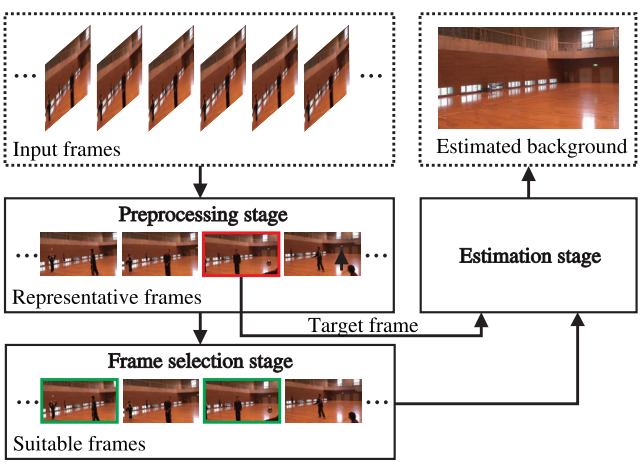


Fig. 5 Overview of our background estimation. Target frame and frames selected for it are indicated by red and green rectangles, respectively.

Algorithm 1. Clustering algorithm.

input $\{F_t | t = 1, 2, \ldots, N_{\text{Frame}}\}$
$S \leftarrow \emptyset, t \leftarrow 1, k \leftarrow 1$
while $t < N_{\text{Frame}}$ **do**
  $C_k \leftarrow \emptyset$
  repeat
    $C_k \leftarrow C_k \cup \{F_t\}$
    Calculate criterion *crit* for set $C'_k = C_k \cup \{F_{t+1}\}$
    Increment t
  until $crit > TH_{\text{crit}}$ and $|C_k| < TH_{\text{N}}$
  $S \leftarrow S \cup \{C_k\}$
  Increment k
end while
return S

and its largest time index, respectively. The algorithm calculates criterion *crit* for $C'_k = C_k \cup \{F_{t+1}\}$ by

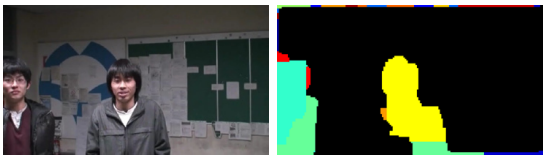
$$crit(C'_k) = \max_{F_\tau} \prod_{F_{t'} \in C'_k / F_\tau} s(F_{t'}, F_\tau)^{\frac{1}{|C_k|}}, \quad (1)$$

where C'_k / F_τ is C'_k but excludes F_τ and $|C_k|$ is the number of frames in C_k. $s(F_{t'}, F_\tau)$ is the similarity between $F_{t'}$ and F_τ based on the number of matched feature points, of which detail is given in our previous paper [13]. The maximization is calculated over $C'_k = C_k \cup \{F_{t+1}\}$. This finds a set C_k of frames whose geometric mean of the similarities over C_k / F_τ is high. If $crit \geq TH_{\text{crit}}$, the algorithm adds F_{t+1} to C_k and repeats this process until $crit < TH_{\text{crit}}$ holds true. Representative frame F^*_k for C_k is defined as F_τ that maximizes *crit*. In addition, the number of frames in a cluster is limited to at most TH_{N} because too many frames in a few clusters excessively suppress their contribution to background estimation.

Threshold TH_{crit} controls the balance between the computational cost and the accuracy of background estimation with TH_{N}. A large value results in many small clusters and raises the computational cost because the number of representative frames increases. In contrast, a small value results in large clusters that decrease the accuracy of the estimated background of non-representative frames, because the synthesized non-representative frames become different from the actual ones. We used 100 for TH_{crit}, which gave a good balance ($|C_k|$ averages 7.5 frames). The value of TH_{N} was empirically set to 20.

Frame selection stage. Since the number of representative frames remains large for the graph cuts-based background estimation and may include frames that do not share a field of view, we further extract frames that are suitable for the background estimation of a given target frame. Observing that the suitable frames uniformly overlap with a major portion of the target frame, *i.e.*, the number of frames that cover each pixel in the target frame is uniform, we extract them as follows.

Let $R = \{F^*_k | k = 1, 2, \ldots, N_{\text{Cluster}}\}$ be the set of all representative frames. As we estimate their backgrounds and those of other frames are synthesized, target frame F^*_{T} is an element of R. We reinterpret the above observation as the following maximization problem:



(a) (b)

Fig. 6 (a) Original frame and (b) estimated background labels l_n.

$$R_{\text{T}} = \arg\max_{R' \subset R} \sum_{F^*_k \in R'} Z_{\text{T}}(F^*_k) + \gamma \sum_{F^*_k, F^*_{k'} \in R'} D(F^*_k, F^*_{k'}), \quad (2)$$

where $Z_{\text{T}}(F^*_k)$ is the proportion of the area of F^*_{T} covered by F^*_k to the area of F^*_{T} and $D(F^*_k, F^*_{k'})$ is the distance between the frame centers of F^*_k and $F^*_{k'}$ projected to F^*_{T}. Parameter γ determines the contribution of each term. In this equation, the first term rewards frames with large overlapping areas, and the second term rewards frames that give large distances from the other frames in R'. The second term confirms that suitable frames are uniformly distributed over the target frame. Since this criterion monotonically increases as $|R'|$ increases, we set the maximum number of frames in R' to 20, considering the computational burden of the graph cuts algorithm. This maximization problem is computationally expensive; therefore, a greedy algorithm finds a suboptimal solution by adding one frame to R' at a time, where the added frame gives the maximum value of the criterion.

Estimation stage. To construct the background of F^*_{T} using R_{T}, we employ a method used in our previous system's background estimation [13], which is originally based on [23]. The method assigns to each pixel of the target frame F^*_{T} a label, l_n, indicating the frame whose corresponding pixel's color is most likely to be the background. The likelihood of being background color is measured based on the assumptions that persons in video frames move while capturing and that the pixel's background color appears more frequently than the moving persons (see [13] for more detail).

Figure 6 (b) shows example labels obtained for the frame shown in Fig. 6 (a). Black represents the regions where $l_n = L_{\text{T}}$, and the other colors represent $l_n \neq L_{\text{T}}$, where L_{T} represents the target frame, *i.e.*, colors of the pixels with $l_n = L_{\text{T}}$ remain the same. The resulting labels are referred to as background labels.

5. ICP Detection and Extraction

To detect ICPs, we focus on our observation that a videographer pays much attention to them but less to the non-ICPs. This difference is reflected in a videographer's behavior, namely, how she moves her camera. Based on this observation, we detect all the persons and classify them as ICPs or non-ICPs using features that encode the videographer's behavior.

Assuming that videographers sufficiently capture the ICPs so that at least their upper bodies are visible in the frame, we first detect and track the upper bodies of all the persons. There are many methods for detection and tracking

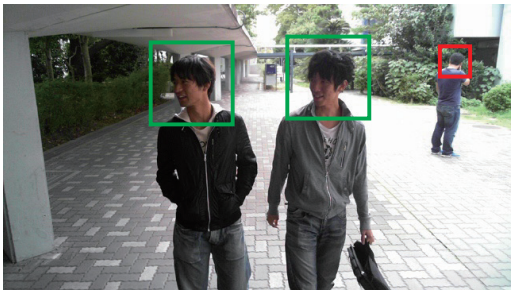


Fig. 7 Vertical positions u_t^{Y} of ICPs (surrounded by green rectangles) in the same frame tend to have similar values. Their sizes (u_t^{S}) are also similar. Those values of ICPs and non-ICP (surrounded by red rectangle) are different.

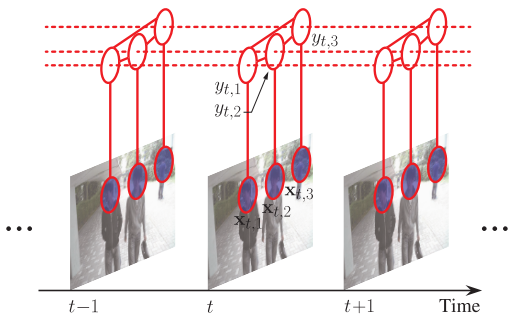


Fig. 8 Graphical representation of our CRF-based model. Lines connecting $\mathbf{x}_{t,k}$ and $y_{t,k}$ represent local classifier terms f. Dashed lines are temporal consistency terms g, which depend on $y_{t,k}$ and $y_{t-1,k}$. Spatial terms h hare represented by lines connecting $y_{t,k}$ and $y_{t,k'}$.

of persons, *e.g.*, [27]–[29], but we assume that the detection and tracking results are given so we focus on our goal of video privacy protection.

Our ICP classification uses as features the trajectories

$$\mathbf{x} = \{(u_\tau^{\mathrm{X}}, u_\tau^{\mathrm{Y}}, u_\tau^{\mathrm{S}}) | \tau = 1, \ldots, N_{\mathrm{T}}\} \tag{3}$$

of tracked person in N_{T} consecutive frames, where u_τ^{X} and u_τ^{Y} are the position of the detection window, and u_τ^{S} is its size, which can describe the videographer's behavior with respect to that person. For a person in the t-th frame, $\mathbf{x}$ consists of his trajectory centered at that frame. Each feature is normalized so that its mean and variance over the training dataset can be 0 and 1, respectively.

For classifying each person as an ICP or a non-ICP, we use a statistical model. To make viewers understand what the videographers want to present, the videographers tend to capture the same ICPs for a while. Also, some features from a pair of ICPs in a single frame have correlation. For example, the vertical position u_t^{Y}'s of ICPs in the t-th frame tend to have similar values as shown in Fig. 7. Taking these into account, we adopt a CRF-based model. Let $\mathbf{x}_{t,k}$ denote the features extracted from the k-th person in the t-th frame, and $\mathbf{X}$ the set of features from all persons in all frames. The problem is to estimate label $y_{t,k} \in \{-1, 1\}$ for $\mathbf{x}_{t,k}$, where $y_{t,k} = -1$ means the non-ICP and $y_{t,k} = 1$ means the ICP. We denote a vector of all labels by $\mathbf{y}$. Figure 8 shows a graphical representation of our CRF-based model. Under this model, the probability of $\mathbf{y}$ given $\mathbf{X}$ is obtained by

$$p(\mathbf{y}|\mathbf{X}) = \frac{1}{Z} \exp\left[E_{\mathrm{ICP}}(\mathbf{X}, \mathbf{y})\right], \tag{4}$$

where Z is a normalization term, and E_{ICP} is the energy function defined as

$$E_{\mathrm{ICP}}(\mathbf{X}, \mathbf{y}) = \sum_{t,k} f(\mathbf{x}_{t,k}, y_{t,k}) + \sum_{t,k,k'} h_{t,k,k'}(y_{t,k}, y_{t,k'}) + \sum_{t,k} g(y_{t,k}, y_{t-1,k}), \tag{5}$$

consisting of local classifier terms f, spatial relationship terms h, and temporal consistency terms g.

A local classifier term indicates how likely the person with $\mathbf{x}_{t,k}$ is an ICP. Basically, our CRF-based ICP model is linear, *i.e.*, all terms are linear combinations of features; therefore, to improve the discrimination power, we employ a two-step approach. First, we learn a classifier $\tilde{f}(\mathbf{x}_{t,k})$ for which we employ a support vector machine (SVM) with all features in $\mathbf{x}_{t,k}$. Second, as f, we use a linear function with a bias term that takes the decision value $\tilde{f}(\mathbf{x}_{t,k})$ as well as $u_{t,k}^{\mathrm{X}}$, $u_{t,k}^{\mathrm{Y}}$ and $u_{t,k}^{\mathrm{S}}$ as a feature. To summarize, our local classifier term is

$$f(\mathbf{x}_{t,k}, y_{t,k}) = y_{t,k} \mathbf{w}^\top \mathbf{u}_{t,k}^{\mathrm{F}}, \tag{6}$$

where the feature vector $\mathbf{u}_{t,k}^{\mathrm{F}} = (1, u_{t,k}^{\mathrm{X}}, u_{t,k}^{\mathrm{Y}}, u_{t,k}^{\mathrm{S}}, \tilde{f}(\mathbf{x}_{t,k}))^\top$ and the parameter vector $\mathbf{w} = (w_0, w_1, w_2, w_3, w_4)^\top$. This local classifier term uses the position and size of the t-th frame's detection window multiple times to emphasize their importance in classification.

A temporal consistency term describes the relationship between the labels $y_{t,k}$ and $y_{t-1,k}$, defined as

$$g(y_{t,k}, y_{t-1,k}) = \begin{cases} a & \text{if } y_{t,k} = y_{t-1,k} \\ 0 & \text{otherwise} \end{cases}. \tag{7}$$

Negative a rewards consistent labels while positive a penalizes them.

Our spatial relationship term makes use of the correlation among ICPs. We design six features that encode such correlation between a pair of persons in a frame, *i.e.*, $|u_{t,k}^m - u_{t,k'}^m|$ and $|u_{t,k}^m + u_{t,k'}^m|$ ($m \in \{\mathrm{X}, \mathrm{Y}, \mathrm{S}\}$). We denote a vector containing them together with 1 for a bias parameter by $\mathbf{u}_{t,k,k'}^{\mathrm{H}}$. The spatial relationship term is

$$h_{t,k,k'}(y_{t,k}, y_{t,k'}) = -y_{t,k} y_{t,k'} \mathbf{v}^\top \mathbf{u}_{t,k,k'}^{\mathrm{H}}, \tag{8}$$

where $\mathbf{v} = (v_0, v_1, v_2, v_3, v_4, v_5, v_6)$.

All parameters $\mathbf{w}$, $\mathbf{v}$, and a can be determined via learning. However, we learn $\mathbf{w}$ and $\mathbf{v}$ but not a to demonstrate the effect of temporal consistency term. For learning, we employ contrastive divergence learning [30], which is repeatedly applied for various values of a. The feature vectors $\mathbf{u}_{t,k}^{\mathrm{H}}$ and $\mathbf{u}_{t,k}^{\mathrm{F}}$ are normalized so that the mean and variance of each element over the training dataset can be 0 and 1, respectively.

To present ICPs in privacy protected videos, they are

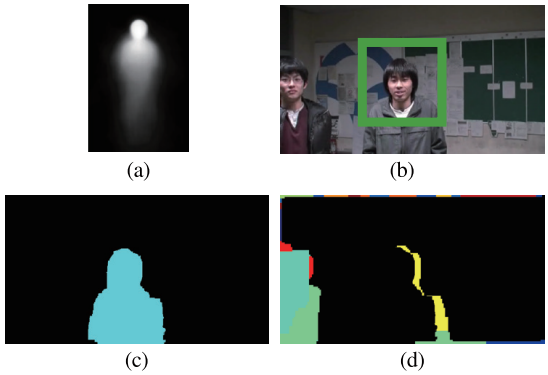


Fig. 9 (a) Shape prior constructed using a dataset of manually labeled images. (b) ICP detection result. (c) Obtained ICP labels $\bar{l}_n$. (d) Combined labels $\tilde{l}_n$.

extracted from the original frames again using the graph cuts algorithm as in [13]. More specifically, it infers ICP labels $\bar{l}_i \in \{0, 1\}$, where $\bar{l}_i = 0$ represents that the i-th pixel belongs to an ICP and $\bar{l}_i = 1$ otherwise. To maintain the detail of ICPs' shape, the algorithm works on pixels in all frames, not grids consisting of 5×5 pixels in representative frames as in background estimation. Figure 9 (c) shows an example of the obtained labels. The blue region represents $\bar{l}_i = 0$, and the black region represents $\bar{l}_i = 1$. In this example, the ICP is accurately extracted. Our previous paper [13] provides more detail.

6. Privacy Protected Video Generation

Privacy protected video frames are generated by superimposing the extracted ICPs onto the estimated background image. Although BEPS potentially generates videos shown in Fig. 2, we implement only non-ICP removal. Other types of obscuration can easily be implemented after non-ICP removal.

Since BEPS only applies background estimation to the representative frames, it synthesizes the background of the non-representative frames based on the representative frames. First, $\bar{l}_i$ in F_t is projected to the two nearest representative frames. Combined label $\tilde{l}_i$ for each representative frame is then generated:

$$\tilde{l}_i = \begin{cases} L_{\mathrm{T}} & \text{if } \bar{l}_i = 1 \\ l_n & \text{otherwise} \end{cases}, \tag{9}$$

where l_n is the label for Ω_n that contains the i-th pixel. An example of $\tilde{l}_n$ is shown in Fig. 9 (d). The regions with $\tilde{l}_n \neq L_{\mathrm{T}}$ are replaced by the pixels from the representative frames corresponding to $\tilde{l}_n$. Since pixel values from different frames may differ due to illumination change, for example, to prevent visual artifacts, we use Poisson blending [31]. After reconstructing two frames from $\tilde{l}_n$ and two representative frames, we again project them to target frame F_t and add them with a weight. That is, by letting $H_{k,t}(B'_k)$ and $H_{k',t}(B'_{k'})$ be the projection of the two reconstructed frames, B'_k and $B'_{k'}$, we generate privacy protected video frame B_t by

$$B_t = \omega_{k,k'}(t) H_{k,t}(B_k) + [1 - \omega_{k,k'}(t)] H_{k',t}(B_{k'}), \tag{10}$$

where $\omega_{k,k'}(t) = (t - t_k)/(t_{k'} - t_k)$ is a weight. t_k and $t_{k'}$ are the indices of the k-th and k'-th representative frames.

7. Experimental Results

We first evaluate the performance of ICP detection using our dataset, comparing it with several baselines including [10]. Then we demonstrate the advantage of BEPS over our previous system [10] using several example videos. We also show the impact of the frame selection stage on the risk of privacy disclosure by comparing the performance of BEPS with and without it.

7.1 Performance Evaluation of ICP Detection

To evaluate ICP detection, we used a video dataset consisting of 20 videos. We could not use publicly available video datasets because they did not contain ground truth labels of ICPs and non-ICPs, which should be specified by the videographers who captured them. Instead, we collected the dataset ourselves[†]. We employed three videographers, and they captured the videos from various situations including both indoor and outdoor scenes. The main subjects were moving and static people. The videographers then manually assigned to their video a label identifying each person as either ICP or non-ICP in each frame. The frame size of these videos was 854×480 pixels, and the frame rate was 30 frames per second. The total number of frames in the video dataset was 32,725. As mentioned in Sect. 5, we did not apply person detection but instead manually specified them. After manually specifying the individuals in the video dataset, videographers assigned to each person a label representing ICP or non-ICP. The cumulative total number of persons was 56,067, where the numbers of ICPs and non-ICPs were 38,122 and 17,945, respectively. We manually tracked the persons in the videos, which resulted in 361 tracked persons, where tracking was terminated if the person's upper body was disappear or occluded. Among them, 251 persons were tracked more than 30 frames, 151 persons were labeled as ICP in more than one frame. We adopted five-fold cross-validation, where 16 videos were used for SVM training and CRF model parameter learning, and five others for evaluation. We employed false positive rate (FPR) and true positive rate (TRP) as performance measures.

Figure 10 (left) shows the AUC values for various a. To generate ROC curves from $\mathbf{y}^*$, we calculated the probability of $y_t = 1$ given X and $\mathbf{y}^*$ by

$$p(y_t = 1 | X, \mathbf{y}^*_{\setminus t}) = \frac{p(y_t = 1, X, \mathbf{y}^*_{\setminus t})}{\sum_{y_t \in 0,1} p(y_t, X, \mathbf{y}^*_{\setminus t})}, \tag{11}$$

where $\mathbf{y}^*_{\setminus t}$ is $\mathbf{y}^*$ excluding the t-th element y^*_t, and we applied thresholding to this probability. The result indicated

[†]We will provide our dataset with ground truth labels upon request. Please contact the first author.

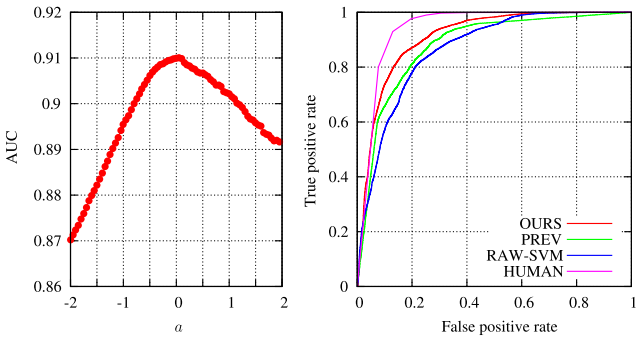


Fig. 10 Left: AUC values for various a. Right: ROCs for our proposed ICP detection (OURS) and various baselines.

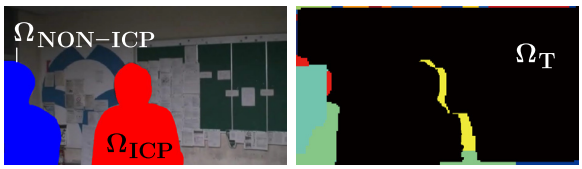
that $a = 0.05$ gave the best performance, which means that our temporal consistency term is not very significant in our model. This can be explained by our new features fed to the SVM: They are a trajectory of detection window's position and size, and is temporally redundant. Therefore, temporal consistency is already counted in the features.

To show the superiority of our ICP detection (OURS), we compared it with three baselines: our previous ICP detection method used in [10] (PREV), the raw SVM decision value-based method [12] (RAW-SVM), and human annotators (HUMAN). In RAW-SVM, we directly applied thresholding to decision values. PREV improved the detection accuracy using ICPs' temporal consistency. We used the same parameters as [10] for PREV. To evaluate the performance of the human annotators, six human annotators[†] individually inferred whether a person in the videos was an ICP. For this, we assigned specific IDs to all the persons in the videos. For each ID, each human annotator specified the frames in which she/he inferred that the corresponding person was an ICP. A person in a frame was judged to be an ICP if four or more human annotators agreed. Figure 10 (right) shows the ROC curves. The human annotators outperformed the others. OURS and PREV worked better than RAW-SVM, indicating that ICPs' temporal consistency improved the performance. Our ICP detection outperformed PREV.

7.2 Comparison with Our Previous System

We evaluated BEPS's capability for privacy protection and video content preservation by comparing it with our previous system (BASELINE) [10]. The parameter γ for BEPS was empirically set to 1. An example frame of BASELINE is shown in Fig. 1 (a). To show the potential performance of BEPS and BASELINE, we used the ground truths of ICPs and non-ICPs specified by the videographers of our video dataset instead of person detection and the ICP classification results. In some videos, persons were not manually specified because they were too small or were only partially captured. For fair comparison, we clarified whether such non-ICPs were included in each video. In addition, for demonstrating BEPS's immunity to the failure of person detection and ICP classification, we randomly dropped the manually specified persons at the rate of ϵ to simulate failure, where ϵ was set to either 0 or 0.1.

[†]The human annotators were students in our laboratory, excluding the authors and videographers who captured the videos in our dataset.



(a) (b)

Fig. 11 Definitions: (a) Ω_{T} consisting of black pixels and (b) Ω_{ICP} and $\Omega_{\text{NON-ICP}}$.

As evaluation measures for BEPS, we defined removal rate (RR) and preservation rate (PR) as follows:

$$\text{RR} = \frac{|\bar{\Omega}_{\text{T}} \cap \Omega_{\text{NON-ICP}}|}{|\Omega_{\text{NON-ICP}}|}, \quad \text{PR} = \frac{|\Omega_{\text{T}} \cap \Omega_{\text{ICP}}|}{|\Omega_{\text{ICP}}|}, \tag{12}$$

where Ω_{T} is the set of pixels with $\bar{l}_n = L_{\text{T}}$, as in Fig. 11 (a), and $\bar{\Omega}_{\text{T}}$ is its complementary set. Ω_{ICP} and $\Omega_{\text{NON-ICP}}$ are the sets of the pixels in ICPs and non-ICPs, respectively, and were manually labeled for all representative frames (Fig. 11 (b)). The RR and PR stand for how accurately BEPS removes the pixels belonging to non-ICP and how accurately BEPS retains the pixels belonging to ICPs, respectively. For BASELINE, we deemed the blocked out regions to be $\bar{\Omega}_{\text{T}}$ and used the same definitions as BEPS. We empirically determined the regions to be blocked out based on the manually specified detection windows so that most of a non-ICP region was blocked. Some ICP regions may also be partially blocked out if their positions are close to the non-ICPs. Since the performances of the background estimation and ICP extraction largely depend on the video content, such as camera motion and person motions, we show RR and PR for each representative frame instead of the statistics of our video dataset. In addition, we applied BEPS to four videos in our dataset, but the evaluation measures are presented for their excerpts with textual descriptions of their content.

Figure 12 shows RR (left) and PR (right) for each representative frame of VID1–VID4, where the black dots on the horizontal axes are the representative frames. Figures 13–16 are examples of the original and generated frames in VID1–VID4 for $\epsilon = 0$. In these figures, the red rectangles indicate ICPs. Brief textual descriptions of each excerpt (about 10 s) with the length of the original video follow:

- VID1 (64.9 s, Fig. 13): The videographer first captured a sitting person as an ICP until around the 100-th frame. Then two persons were captured as non-ICPs, corresponding to the interval without RR in the first row of Fig. 12. Finally, the person who was an ICP at the beginning became a non-ICP, and the other person was captured as an ICP.
- VID2 (63.9 s, Fig. 14): This video contained two persons. First, one person was an ICP and the other was a

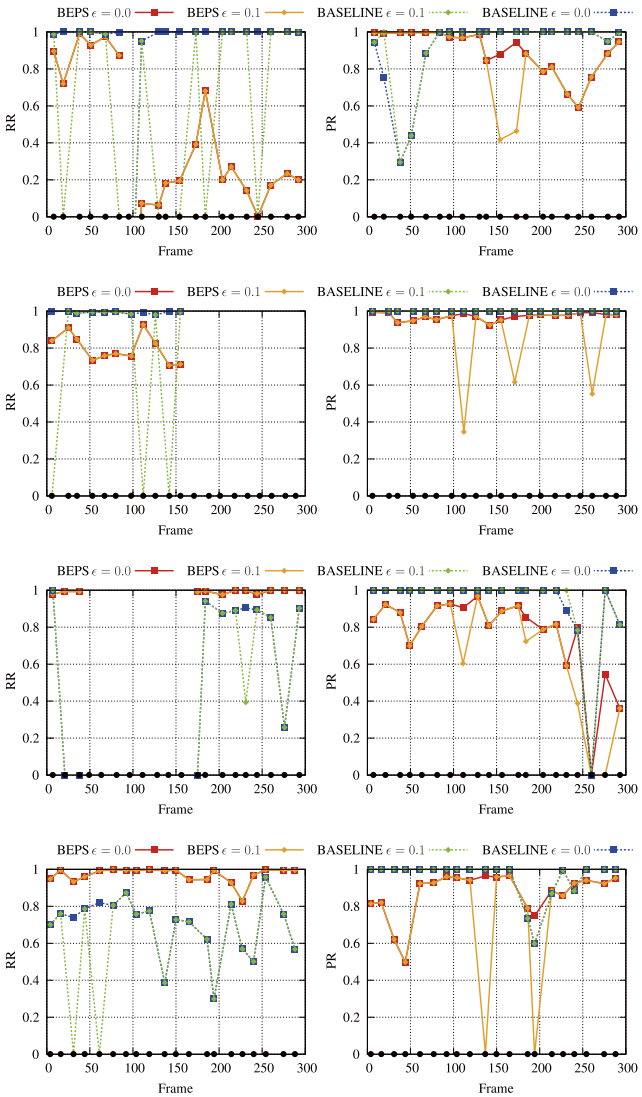


Fig. 12 Comparision: RR (left) and PR (right) between BEPS and BASELINE for VID1–VID4 (from top to bottom).

non-ICP who was partially out of the frame. Then the videographer captured both as ICPs.

- VID3 (9.93 s, Fig. 15): VID3 kept capturing two ICPs. After the 200-th frame, some non-ICPs occluded the two ICPs. Since the non-ICPs intersected the camera's field of view, they were partially captured when they were framed in and out.
- VID4 (40.9 s, Fig. 16): There were four persons, one of whom was an ICP. One of the non-ICPs in this video was too small, and another was only partially captured.

For $\epsilon = 0$, the RR values gave comparable results for BEPS and BASELINE in most videos. However, in some videos, BEPS's RR values dropped, which was significant after the 100-th frame in VID1. This RR degradation was caused by background estimation failure. Our background estimation, which is analogous to median-based methods, basically fails to estimate the background when non-ICPs do not move, as in the third to fifth frames in Fig. 13. The slight degradation in the RRs of VID2, VID3, and VID4 was also caused by the failure in the background estimation.

The PR of BEPS was mostly lower than BASELINE. This was caused by our inflexible shape prior. To present ICPs in privacy protected videos, BEPS extracts them from the original video frames and superimposes them on an estimated background. In this process, our shape prior for ICP extraction gave a weak likelihood of being included in an ICP region when its actual shape was different from our shape prior shown in Fig. 9 (a). This resulted in the failed extraction of the exact shapes of ICPs, especially for feet and arms. The inflexibility of our shape prior also led to blurring-like visual artifacts due to Poisson blending (Fig. 16). Employing a human pose estimation technique (*e.g.*, [28]) can relax this inflexibility by deforming the shape prior based on a person's pose, although its implementation is beyond the scope of this paper.

For $\epsilon = 0.1$, BASELINE's RR values significantly degraded in some frames because some manually detected persons were randomly dropped. This result implies that BASELINE and all other conventional systems that detect persons to be obscured are vulnerable to detection failure of non-ICPs, resulting in their complete disclosure. On the other hand, the RR values did not change for BEPS, and thus, it was immune to the detection failure. Such insensitivity is desirable for privacy protection systems because failure in person detection does not lead the disclosure of the privacy sensitive information, although it in turn depends on the performance of background estimation. In contrast, BEPS's PR values fell due to randomly dropped persons, and BASELINE was insensitive to detection failure of ICPs.

Table 1 Timing results.

	VID1	VID2	VID3	VID4
# frames	1947	1917	298	1228
# representative frames	196	308	14	63
Background estimation	5289 s	6488 s	1067 s	4863 s
ICP detection & extraction	757 s	780 s	109 s	461 s
Privacy protected video gen.	1552 s	1001 s	303 s	1632 s
Total	7598 s	8270 s	1479 s	6957 s

Table 1 shows the timing results on Windows 7 PC with Intel Core i7 CPU at 3.4 GHz and 16 GB memory. We implemented ICP detection in Python and all other components in C++. Background estimation for representative frames took processing time the most, although it was applied only to representative frames. It took roughly 5 s per frame on average. The graph cuts algorithm, which is used in all processes, was the bottleneck.

7.3 Validation of Frame Selection Stage

To investigate the influence of the frame selection stage on BEPS's capability in privacy protection and video content preservation, we compared BEPS with and without it, where BEPS without the frame selection stage simulates our system in [13]. We used the same videos and parameters as in Sect. 7.2, and ϵ was set to 0. As mentioned above, handling

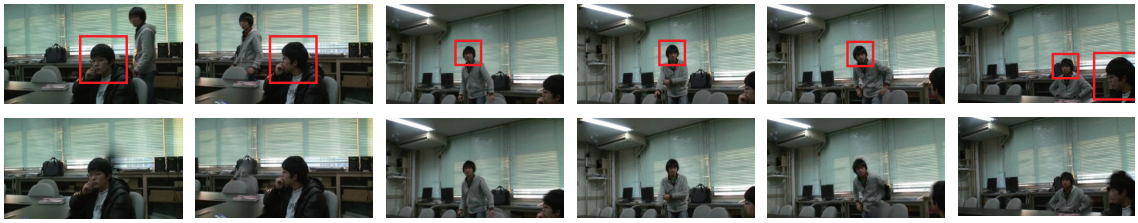


Fig. 13 Examples of original frames (top) and resulting frames (bottom) for VID1. ICPs are indicated by red rectangles.

Fig. 14 Examples of original frames (top) and resulting frames (bottom) for VID2. ICPs are indicated by red rectangles.

Fig. 15 Examples of original frames (top) and resulting frames (bottom) for VID3. ICPs are indicated by red rectangles.

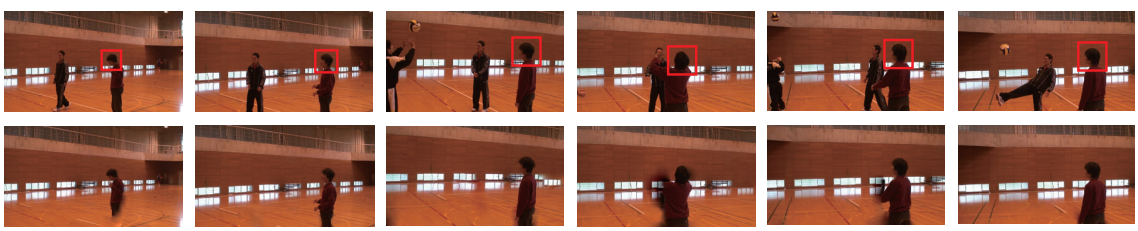


Fig. 16 Examples of original frames (top) and resulting frames (bottom) for VID4. ICPs are indicated by red rectangles.

long videos is practically infeasible without the frame selection stage because such videos have many representative frames and require a large amount of memory for processing. Therefore, we extracted an excerpt from each video and applied BEPS without the frame selection stage to it, while we applied BEPS with the frame selection stage to the entire video and extracted the evaluation result that corresponds to the excerpt. In this way, we show that the frame selection stage can give suitable representative frames for a target frame from an entire video.

Figure 17 shows the evaluation results. The results for BEPS with frame selection stage were almost identical to those when $\epsilon = 0$ in the previous section. Without the frame selection stage, BEPS's background estimation failed in VID2 and gave low RR values. This is because the excerpt did not contain the background pixels, while BEPS with frame selection stage selected suitable frames from the other part of the original video. The slight difference in RR and PR was caused by the difference in the frames used for the background estimation. These results indicate that the frame selection stage enabled BEPS to handle long video sequences without significantly the RR and PR performances.

8. Discussion

This section discusses (i) how BEPS fulfills the requirements in Sect. 1 and (ii) the limitations of BEPS.

(i) Fulfillment of requirements. A system for social

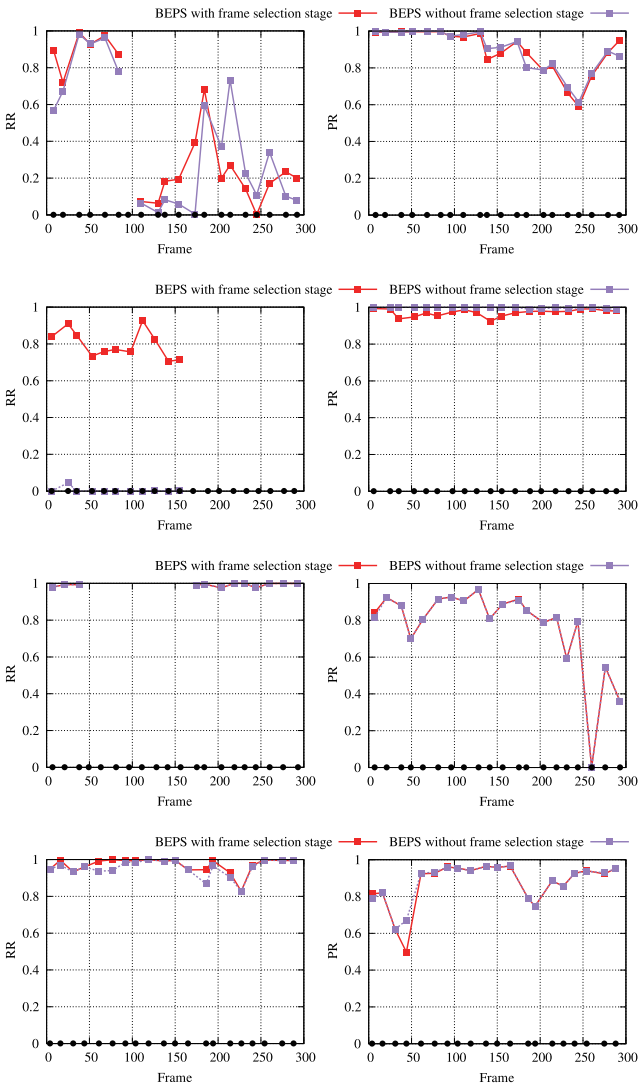


Fig. 17 Comparisons: RR (left) and PR (right) between BEPS with and without frame selection stage for VID1–VID4 (from top to left).

video privacy protection must fulfill the requirements mentioned in Sect. 1. Similar to our previous system [10], [11], BEPS successfully satisfied requirement 1) in most cases using ICP detection. For requirement 2), BEPS can provide various appearance obscuration methods, as in Fig. 2, based on the estimated background. Although we did not experimentally demonstrate various appearance obscuration methods except the complete removal of the appearance, implementation is trivial for most of them.

(ii) Limitations. From the experimental results, BEPS worked well for video in which non-ICPs move. Compared with such conventional systems as [10], which first detect individuals to be obscured and then apply a privacy obscuration method, BEPS was immune to the failure of person detection. This immunity is preferable for privacy protection. In conventional systems, the failure of person detection directly resulted in privacy disclosure. In contrast, it resulted in missing ICPs in BEPS but did not disclose non-ICPs. However, our results also exposed two limitations.

First, in principle, background estimation fails if non-ICPs are stationary, which leads to privacy disclosure. VID1 clearly showed this problem. From the third to sixth frames in Fig. 13, the person on the left side was a non-ICP, but the current implementation of BEPS failed to remove him. To address this problem, we need to use the output of a person detection algorithm so that BEPS can identify the regions that potentially contain non-ICPs. If the background for these regions is not estimated (*i.e.*, $l_n = L_{\mathrm{T}}$), such a technique as image inpainting removes non-ICPs. However, this approach spoils one of the advantages of BEPS: insensitivity to the low detection accuracy of non-ICPs. Another possible approach is incorporating [32], which automatically detects regions in which the background estimation failed and applies image inpainting to them.

The second problem is the inflexibility of the shape prior. Currently, we use a shape prior of persons that does not change regardless of their pose. In addition, since we built it as an averaged region of a body, our ICP extraction fails to extract body parts that are invisible in many videos in our dataset, such as feet. As mentioned above, human pose estimation techniques (*e.g.*, [28]) can solve this problem by modifying the shape prior based on estimated poses.

9. Conclusion

We developed a system called BEPS to protect privacy in social videos. Privacy protection for videos in SNSs is a new class of problems that requires privacy obscuration and video content preservation. For this problem, BEPS first estimates the background of the video frames and superimposes the detected ICPs on it. The advantage of BEPS is that it does not require accurate non-ICP detection, which is usually more difficult than accurate ICP detection. We experimentally showed that our ICP detection achieved an AUC value of 0.91, and our ICP extraction and background estimation worked well if non-ICPs move. Also, we demonstrated that BEPS was preferable to conventional systems due to its immunity to detection failure. Our experimental results also implied that background estimation and ICP extraction might fail, but BEPS's advantage over conventional systems makes it a beneficial alternative for social video privacy protection. Future work includes implementing a more sophisticated shape prior using, *e.g.*, human pose estimation [28] and various privacy obscuration as shown in Fig. 2.

Acknowledgements

This work was partly supported by the Ministry of Internal Affairs and Communications SCOPE No. 152107001, JSPS KAKENHI Nos. 24240031, 25730115, 15H01686, and NICT No. 178B0504.

References

[1] S. Tansuriyavong and S. Hanaki, "Privacy protection by concealing

persons in circumstantial video image," Proc. 2001 Workshop on Perceptive User Interfaces, 4 pages, 2001.
[2] D. Chen, Y. Chang, R. Yan, and J. Yang, "Tools for protecting the privacy of specific individuals in video," EURASIP Journal on Applied Signal Processing, vol.2007, no.1, 9 pages, 2007.
[3] F. Dufaux and T. Ebrahimi, "Scrambling for privacy protection in video surveillance systems," IEEE Trans. Circuits and Systems for Video Technology, vol.18, no.8, pp.1168–1174, 2008.
[4] K. Chinomi, N. Nitta, Y. Ito, and N. Babaguchi, "PriSurv: privacy protected video surveillance system using adaptive visual abstraction," Advances in Multimedia Modeling, Lecture Notes in Computer Science, vol.4903, pp.144–154, 2008.
[5] A. Frome, G. Cheung, A. Abdulkader, M. Zennaro, B. Wu, A. Bissacco, H. Adam, H. Neven, and L. Vincent, "Large-scale privacy protection in google street view," Proc. Int'l Conf. Computer Vision, pp.2373–2380, 2009.
[6] A. Flores and S. Belongie, "Removing pedestrians from google street view images," Proc. Int'l Workshop on Mobile Vision, pp.53–58, 2010.
[7] J. Chaudhari, S.S. Cheung, and M.V. Venkatesh, "Privacy protection for life-log video," Proc. Workshop on Signal Processing Applications for Public Security and Forensics, pp.1–5, 2007.
[8] J. Brassil, "Technical challenges in location-aware video surveillance privacy," in Protecting Privacy in Video Surveillance, pp.91–113, Springer Verlag, 2009.
[9] J.A. Halderman, B. Waters, and E.W. Felten, "Privacy management for portable recording devices," Proc. ACM Workshop on Privacy in the Electronic Society, pp.16–24, 2004.
[10] Y. Nakashima, N. Babaguchi, and J. Fan, "Intended human object detection for automatically protecting privacy in mobile video surveillance," Multimedia Systems, vol.18, no.2, pp.157–173, 2012.
[11] Y. Nakashima, N. Babaguchi, and J. Fan, "Automatically protecting privacy in consumer generated videos using intended human object detector," Proc. Int'l Conf. Multimedia, pp.1135–1138, 2010.
[12] Y. Nakashima, N. Babaguchi, and J. Fan, "Detecting intended human objects in human-captured videos," Proc. Conf. Computer Vision and Pattern Recognition Workshop, 8 pages, 2010.
[13] Y. Nakashima, N. Babaguchi, and J. Fan, "Automatic generation of privacy-protected videos using background estimation," Proc. Int'l Conf. Multimedia and Expo, 6 pages, 2011.
[14] I. Mitsugami, M. Mukunoki, Y. Kawanishi, H. Hattori, and M. Minoh, "Privacy-protected camera for the sensing web," Proc. Information Processing and Management of Uncertainty in Knowledge-Based Systems. Applications, pp.622–631, 2010.
[15] I. Kitahara, K. Kogure, and N. Hagita, "Stealth vision for protecting privacy," Proc. Int'l Conf. Pattern Recognition, pp.404–407, 2004.
[16] "Face blurring: when footage requires anonymity," July 2012.
[17] J. Fan, H. Luo, M.-S. Hacid, and E. Bertino, "A novel approach for privacy-preserving video sharing," Conf. Information and Knowledge Management, pp.609–616, 2005.
[18] J. Peng, N. Babaguchi, H. Luo, Y. Gao, and J. Fan, "Constructing distributed hippocratic video databases for privacy-preserving online patient training and counseling," IEEE Trans. Information Technology in Biomedicine, vol.14, no.4, pp.1014–1026, 2010.
[19] N. Babaguchi, T. Koshimizu, I. Umata, and T. Toriyama, "Psychological study for designing privacy protected video surveillance system: PriSurv," in Protecting Privacy in Video Surveillance, pp.147–164, Springer Verlag, 2009.
[20] M.K. Saini, P.K. Atrey, S. Mehrotra, and M.S. Kankanhalli, "Privacy aware publication of surveillance video," International Journal of Trust Management in Computing and Communications, vol.1, no.1, pp.23–51, 2013.
[21] T. Koyama, Y. Nakashima, and N. Babaguchi, "Markov random field-based real-time detection of intentionally-captured persons," Proc. Int'l Conf. Image Processing, pp.1377–1380, 2012.
[22] T. Koyama, Y. Nakashima, and N. Babaguchi, "Real-time privacy protection system for social videos using intentionally-captured persons detection," Proc. Int'l Conf. Multimedia and Expo, 6 pages, 2013.
[23] D.-W. Kim and K.-S. Hong, "Practical background estimation for mosaic blending with patch-based Markov random fields," Pattern Recognition, vol.41, no.7, pp.2145–2155, 2008.
[24] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," IEEE Trans. Pattern Analysis and Machine Intelligence, vol.26, no.9, pp.1124–1137, 2004.
[25] J.B. MacQueen, "Some methods for classification and analysis of multivariate observations," Proc. 5th Berkeley Symposium on Mathematical Statistics and Probability, pp.281–297, 1967.
[26] B.J. Frey and D. Dueck, "Clustering by passing messages between data points," Science, vol.315, pp.972–976, 2007.
[27] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," Proc. Conf. Computer Vision and Pattern Recognition, pp.886–893, 2005.
[28] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp.1014–1021, 2009.
[29] S. Walk, N. Majer, K. Schindler, and B. Schiele, "New features and insights for pedestrian detection," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp.1030–1037, 2010.
[30] G.E. Hinton, "Training products of experts by minimizing contrastive divergence," Neural Computation, pp.1771–1800, 2002.
[31] P. Pérez, M. Gangnet, and A. Blake, "Poisson image editing," ACM Trans. Graphics (Proc. ACM SIGGRAPH 2003), vol.22, no.3, pp.313–318, 2003.
[32] X. Chen, Y. Shen, and Y.H. Yang, "Background estimation using graph cuts and inpainting," Proc. Graphics Interface, pp.97–103, 2010.

Yuta Nakashima received the B.E. and M.E. degrees in communication engineering from Osaka University, Osaka, Japan in 2006 and 2008, respectively, and the Ph.D. degree in engineering from Osaka University, Osaka, Japan, in 2012. He is currently an assistant professor at Graduate School of Information Science, Nara Institute of Science and Technology. He was a research fellow of the Japan Society for the Promotion of Science (JSPS) from 2010 to 2012, and was a visiting scholar at the University of North Carolina at Charlotte in 2012. His research interests include video content analysis using probabilistic and statistical approaches. He is a member of the IEEE, the ACM, the IEICE, and the IPSJ.

Noboru Babaguchi is currently a Professor of the Department of Communication Engineering, Osaka University. He has published over 200 journal and conference papers and several textbooks. He is a fellow of the IEICE, a senior member of the IEEE, and a member of the ACM, the IPSJ, the ITE and the JSA.

Jianping Fan received the M.S. degree in theory physics from Northwest University, Xian, China,in 1994, and the Ph.D. degree in optical storage and computer science from Shanghai Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Shanghai, China, in 1997. He was a Postdoctoral Researcher with Fudan University, Shanghai, China, during 1998. From 1998 to 1999, he was a Researcher with the Japan Society of Promotion of Science (JSPS), Department of Information System Engineering, Osaka University, Osaka, Japan. From September 1999 to 2001, he was a Postdoctoral Researcher with the Department of Computer Science, Purdue University, West Lafayette, IN. In 2001, he joined the Department of Computer Science, University of North Carolina, Charlotte, as an Assistant Professor and became Associate Professor and Professor. His research interests include image/video analysis, semantic image/video classification, personalized image/video recommendation, surveillance videos, and statistical machine learning.