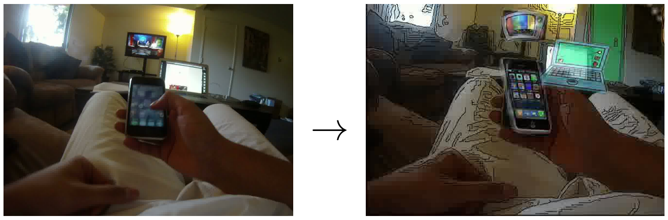
Cartooning for Enhanced Privacy in Lifelogging and Streaming Videos

Eman T. Hassan Rakibul Hasan Patrick Shaffer David Crandall Apu Kapadia
School of Informatics and Computing
Indiana University Bloomington
{emhassan, rakhasan, patshaff, djcran, kapadia}@indiana.edu

Abstract

We describe an object replacement approach whereby privacy-sensitive objects in videos are replaced by abstract cartoons taken from clip art. Our approach uses a combination of computer vision, deep learning, and image processing techniques to detect objects, abstract details, and replace them with cartoon clip art. We conducted a user study (N=85) to discern the utility and effectiveness of our cartoon replacement technique. The results suggest that our object replacement approach preserves a video's semantic content while improving its privacy by obscuring details of objects.



(a) (b)

Figure 1: Streaming and first-person video often contains private information, such as digital device displays. We propose a "cartooning transformation" that automatically replaces objects with clip art representations and abstracts background regions, to obscure private details while preserving overall semantics of the scene.

1. Introduction

Sharing videos has become very popular: YouTube alone receives 300 hours of new footage every *minute* [44]. Meanwhile, live video-sharing services like YouStream and Periscope let people take and broadcast videos to other people in real-time. This new live-streaming technology is similar to traditional peer-to-peer video conferencing services like Skype and Google Hangouts, but encourages people to broadcast video from their mobile phones to many others at a time. Not only do people use videos to simply share their day-to-day lives with others, video sharing has become a powerful tool for exposing fraud and improving accountability of public officials [23].

However, video sharing also introduces significant risks to privacy because it can capture a huge amount of incidental information about the activities, interactions, and environment around the camera. For instance, a user might wear a GoPro camera to share a video of his or her work life, but the video will very likely capture private details like emails on computer screens, instant messages on smartphones, financial details on documents, and the identities of other people [20]. A simple Skype conversation between a student and her parents might take an embarrassing turn when her roommate enters the room in only a towel.

Recent work has considered how to protect privacy in video from several different perspectives. For instance, Scanner Darkly [22] tries to prevent sensitive image data from being released to "perceptual applications" by transforming raw images into high-level metadata that abstracts away details while maintaining enough information for the applications. That solution, however, does not address situations in which other *people* are the consumers of the video, where abstract representations are unsuitable. Other work has taken the opposite approach of sharing most imagery but automatically detecting and censoring certain objects and scenes, including faces [2, 17, 26], computer monitors [25], private rooms [37], and specially marked regions of scenes (like portions of whiteboards) [35]. These approaches may work when a small set of possible privacy concerns needs to be considered, but may not be able to ever cover the full spectrum of possible scenarios.

We propose automatic 'cartooning' transforms to enhance privacy in live-streamed and first-person imagery (Figure 1). Much in the way animated movies abstract away the details of the real world to convey only the most important semantic elements, cartooning transformations can obscure private details of videos while still retaining the overall 'story.' Parameters of the algorithms can be adjusted to control the aggressiveness of the transformations.

As a first step, we develop an initial automatic algorithm for transforming videos into cartoon-like representations, applying several types of image processing and computer vision techniques. The algorithm has two major components. The first is to apply image processing to abstract out visual details of the whole scene in an object-independent way. The second detects certain objects and replaces them with clip art images that convey general attributes of the object but not the fine-grained details. We address the significant challenge of how to automatically select, align, and integrate the clip art into the scene in an aesthetically pleasing way. The combination of these two components has several advantages over using either one individually: (1) background details are removed while the presence (but not details) of certain sensitive objects are highlighted through clip art, and (2) some degree of privacy preservation is ensured by the image processing transform even when the system fails to replace a sensitive object properly.

We test our techniques on three real-world video collections from three scenarios: first-person video capturing activities of a single person at home, first-person video captured in public at an amusement park, and Skype videos with people and objects in the background. To evaluate how well the transformations preserve privacy while retaining semantics, we conducted a user study on Amazon Mechanical Turk [1] to measure observers' (in)ability to recognize properties of the transformed scenes.

2. Related Work

Preserving privacy in image and video data has been studied from several different perspectives. A large body of literature focuses on surveillance scenarios [31], where the goal is to protect people's privacy when they are captured by static cameras. Here we consider consumer video from devices like smartphones, which is significantly more challenging because of the highly dynamic camera motion and rapidly changing and uncontrolled scenes.

Simple approaches to protect privacy in images and videos include filters such as blurring, pixelating, and masking to obscure privacy-sensitive regions [22]. However, Neustaedter *et al.* [30] and Gross *et al.* [16] showed that these methods often either do not obscure enough detail to provide adequate privacy, or obscure so much that they destroy the utility of the video. Techniques like face morphing [26], face de-identification [17], face swap [2], and image melding [29] can anonymize faces, but do not attempt to block information that may be revealed by other objects.

Boyle *et al.* [4] studied blurring and pixelating videos in online collaboration scenarios, such as an employee at home communicating with other employees via video chat, and concluded that blurring effectively reduces privacy risks and retains utility of visual data. Our proposed system allows for selective obscuring, so that the person using the chat application remains fully visible, while the background and other objects, including people, are abstracted away by cartooning.

PuPPIeS [19] and P3 [34] address security of images stored or shared online by encrypting all or part of the shared images with secret keys, so that only authorized users can reconstruct the original. POP [45] employs a similar approach to protect sensitive regions in a photo, and provides a framework for privacy preserving photo sharing and searching in the cloud.

Outside of the privacy domain, work in image processing and computer graphics has studied how to create visual abstractions of real imagery, typically using low-level operations like segmentation and posterization [27,42]. For example, Bousseau *et al.* [3] and Hays and Essa [18] present techniques for creating artistic "painted" versions of images. In the video domain, Winnemöller *et al.* [42] propose an extended nonlinear diffusion filter to blur small discontinuities and sharpen edges, and then detect edges and quantize colors. While our approach uses similar low-level, content-independent image processing to create abstract representations for the background, we also detect certain objects and replace them with clip art. Hwang *et al.* [21] and Wang *et al.* [41] create comic "narratives" that also take semantic content into account, but require movie scripts and other metadata; for example, Wang *et al.* [40] rely on user input to mark semantic regions in key frames, whereas we propose a fully automatic approach.

Many cartoon abstraction techniques focus in particular on how to represent faces [7, 28, 36, 43, 46]. Most of this work assumes faces are seen in frontal views, while we need to handle arbitrary poses in unconstrained environments. Moreover, most of these papers try to produce photo-realistic cartoon faces that preserve identity, whereas our objective is to obscure facial identity while still representing general properties of the face like gender and emotion. Similar to Rhee and Lee [36], we use cartoons to represent facial features (eyes, mouth, nose), but use blurring and random noise to obscure the person's identity.

Perhaps the most similar work to ours is that of Erdélyi *et al.* [10, 11], who (like us) use cartooning to preserve privacy, but their study is restricted to surveillance videos. To our knowledge, we are the first to study the effect of cartooning on mobile and first person video. Moreover, they applied cartoon effects globally and uniformly, whereas our system also performs object recognition to replace specific objects with clip art. Their later work [12] can perform local cartooning, but requires annotated data including locations of sensitive regions in the frames. As a final step, they applied pixelation to faces to achieve greater privacy, while our proposed system overlays real faces with cartoon faces, which is potentially not only more visually appealing but also can preserve some semantic information about the person (e.g., gender, facial expression, etc.). We generalize this idea to the ability to detect a broader set of objects besides faces, replacing them with clip art that obscures private details while preserving information about the attributes of the object (e.g. position, color, shape, size). This ability to recognize objects could give users of our system finer-grained control to adjust the aggressiveness of the privacy filtering, allowing them to strike a their own balance between privacy and realism of the video.

3. Video cartooning

We propose an initial prototype system for creating these cartoon representations automatically. We use a combination of two broad classes of techniques. The first is to apply global image-level, object-independent transformations to remove incidental private information, especially in the background of scenes. The second is to use object recognition to apply local-level, object-centric transformations that replace specific objects with clip art images. The goal is to create an abstract, cartoon-like representation that preserves semantic scene information while obscuring details.

3.1. Global transformation

Our first step is to apply a global transformation to each video frame I that tries to preserve major edges in the image in order to capture the overall 'sharpness' of the scene, while obscuring details within the edges. To do this, we first apply the Felzenszwalb and Huttenlocher [14] segmentation algorithm to partition the image into contiguous regions based on similarity of color and texture features. We set these parameters to fixed values based on hand tuning ($\sigma = 0.5$, threshold $k = 50$, minimum component size 50) in our implementation, although in practice these could be tuned by the user to trade-off between privacy and realism; intuitively, the more segments, the more the transformed image faithfully reproduces the original, whereas fewer segments means more image details are obscured. Figures 2(a) and (b) show an example of an image before and after segmentation.

To weakly preserve some local scene details, for each pixel $p \in I$ we take a weighted average of the original image and the segmented image,

$$I'(p) = \alpha I(p) + (1 - \alpha) I_s(p),$$

where I_s is the segmented image. The parameter α is also tunable, again controlling trade-off between realism and privacy. The result of this averaging is illustrated in Figure 2(c). Finally, to highlight the major edges of an image (similar to the black boundary lines characteristic of cartoons), we apply Canny edge detection [6] to the original image, and color each major edge pixel in I' in black (Figure 2(d)). This global image transformation technique works well in our experience in that it is fast to compute, obscures private details while keeping overall information about the scene, and (in our subjective opinion) produces images that are generally aesthetically pleasing. Our user study (below) quantitatively measured how well the transformations obscure private details while preserving important semantics, and also solicited some feedback on the aesthetic quality.

3.2. Local image transformation

The second phase of the cartooning process is to identify specific objects and replace them with clip art, in order to obscure object details that may be private (e.g., a laptop display or the title of a book). This requires addressing several challenges, including accurately identifying and localizing objects, selecting suitable clip art based on the properties of the specific object instance, aligning the clip art to the image in terms of orientation, scale, location, and perspective, and then blending it in an aesthetically-pleasing way.

3.2.1 Selecting clip art

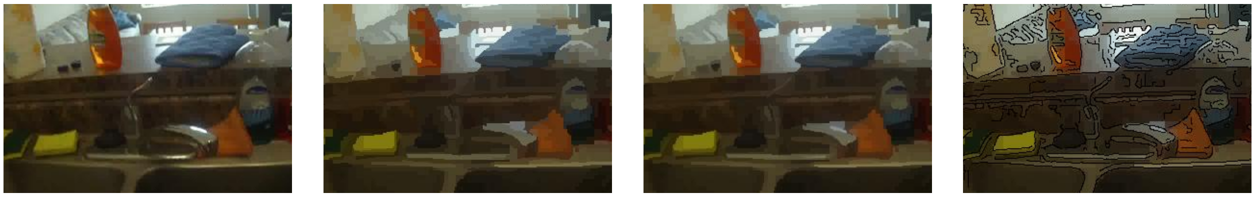
We use Region-based Convolutional Neural Networks (R-CNNs) [15] to detect and localize objects in each video frame.

Then, we choose suitable replacement clip art imagery. We also need to transform the clip art to align it with the actual object in the image, so that it appears in the right position, scale, and viewpoint. For example, having a bounding box around the instance of a car is not sufficient to insert a reasonable clip art representation, because cars are 3D objects whose appearance differs dramatically from different viewpoints.

We assembled a library of 2D clip art images categorized into different types of objects by manually searching the web and collecting several clip art images for each of the 200 classes supported by the public R-CNN model. For example, Figure 3 shows the six clip art images collected for the "pan" class. We also manually edited the clip art to remove any background information. For each instance of a detected object, we randomly choose a clip art image from the same category randomly; in future work we could choose these based on some other criteria (e.g., user preference based on certain colors, styles, or to reflect certain moods).

3.2.2 Aligning and rendering clip art

We view the problem of finding a fine-grained alignment of the clip art to the image as a visual feature matching problem, in which we want to find a transformation of the clip art such that its visual appearance matches the appearance of the actual object as much as possible. This transformation step is necessary because a clip art image's original scale



(a) (b) (c) (d)

Figure 2: *Illustration of steps in the global-level cartooning process:* (a) an input video framed (zoomed in to show detail), (b) result of image segmentation, (c) result after blending segmentation with original image, and (d) result after highlighting strong edges. (Best viewed in color.)

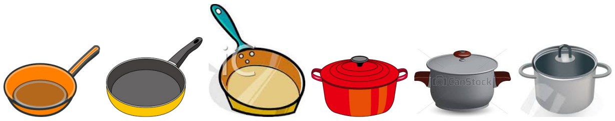
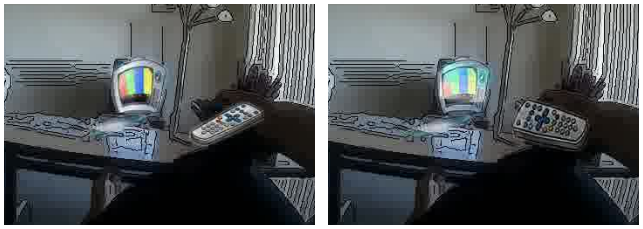


Figure 3: *Images in our clip art library for object "pan."*



(a) (b)

Figure 4: *Illustration of clip art alignment.* (a) Without transforming clip art to align with the configuration of objects in a scene, some objects (e.g. the TV) appear properly but most will have incorrect scale or orientation (e.g., the remote control). (b) After aligning and transforming, the remote control clip art better fits the scene.

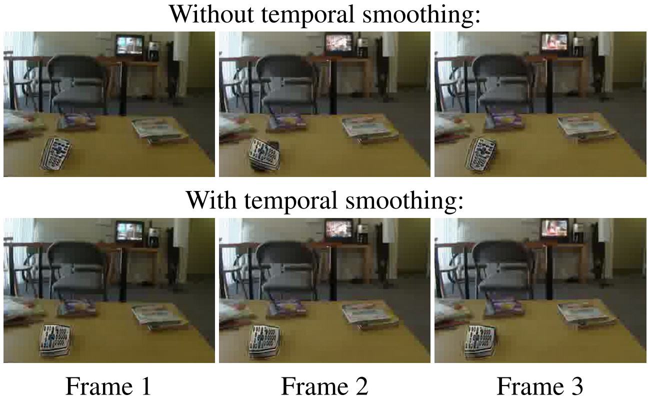


Figure 5: *Illustration of temporal object pose smoothing.* Top: If clip art is aligned with each frame individually, the inferred pose may vary, such as the remote control whose orientation oscillates back and forth here. Bottom: With temporal smoothing using a Markov Random Field, a consistent pose is chosen across time.

and orientation is unlikely to match that of the object in the real image, as in the example in Figure 4. To do this we estimate an ideal "pose" – i.e., a position, scale, and orientation of the clip art in the image. Formally, a pose $p = (o, s, \theta)$ consists of the location of the object center $o \in \mathbb{R}^2$ (a 2D coordinate), its scale $s \in \mathbb{R}^2$ (height and width), and its orientation $\theta \in [0, 2\pi)$ (in-plane rotation angle).

We first consider how to do this for a single object having bounding box b in a single image I, and then generalize this method for video. We compute the Histogram of Oriented Gradients (HOG) [9] feature representation of $I(b)$, the image region corresponding to the detected bounding box. For the clip art, we generate candidate renderings with many different poses, and calculate the HOG features for each of these renderings. We then choose the candidate whose HOG features best match those of the actual object bounding box b,

$$p^* = \operatorname*{argmin}_{p=(o,s,\theta)} \phi(I(b), p),$$

where ϕ measures the similarity between the image bounding box and the rendered clip art with a given pose p,

$$\phi(I, p) = ||\mathcal{H}(I(b)) - \mathcal{H}(\mathcal{T}_p(C))||,$$

C is the clip art image, $\mathcal{H} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^n$ denotes the HOG feature extractor, and $\mathcal{T}_p$ is a transformation that applies pose p (rotation, scaling, and translation) to an image.

To generalize this idea to video, we could simply apply the above operation on a per-frame basis, but found that this gave poor results in practice because a rendered object's appearance can vary dramatically from frame to frame, as illustrated in Figure 5. To impose temporal smoothness, we solve a joint optimization that tries to find the best transformation to match the visual appearance of the object while also avoiding major pose changes from frame to frame. More formally, let $\mathcal{I} = (I_1, I_2, ..., I_m)$ be a sequence of contiguous video frames in which an object is detected with corresponding bounding boxes $b_1, b_2, ..., b_m$. Let $P = (p_1, ..., p_m)$ be the unknown ideal pose parameters of the clip art in each frame. Then solving for all of the poses across time is an optimization problem,

$$P^* = \underset{P}{\operatorname{argmin}} \sum_{i\in[0,m]} \phi(I_i, p_i) + \sum_{i\in[0,m-1)} \psi(p_i, p_{i+1}),$$

where ψ is a distance function that penalizes sudden changes in object pose between frames,

$$\psi(p_i, p_j) = \beta_1||o_i - o_j|| + \beta_2||s_i - s_j|| + \beta_3||\theta_i - \theta_j||,$$

and the β values are constants. This optimization is a chain-structured Markov Random Field model which can be solved by the Viterbi algorithm in linear time [24]. Nevertheless, computing HOG features for many possible scales and orientations of a clip art is still computationally expensive, so we simplify the problem in several ways. First, we assume that an object's bounding box location is accurate enough that the center of the bounding box is a good estimate for the center of the clip art, and thus fix $o_i^* = \hat{b_i}$, the bounding box center. Second, we break apart the optimization of the rotations and scales by solving for scale within each possible rotation. Third, we discretize the set of possible poses into 8 scales and 19 orientations (in 10 degree increments between -90 and 90 degrees).

3.2.3 Inserting the clip art

Once the ideal pose for each object has been found in each image, the final step is to insert the clip art. One subtle issue is how to decide the order in which objects are rendered, which is important when objects occlude one another. If we had scene depth data we could do this exactly, but here we use a simple pre-defined precedence of objects based on typical scene characteristics. For example, we render "doors" before "tv remotes" since the latter might occlude the former in real life, but rarely vice-versa.

We experimented with various approaches for blending the rendered clip art into the cartoonized image in an aesthetically pleasing way, including linear alpha blending (which tended to either preserve too many details of the real image or obscure it too much) and Poisson-based blending [32] (which was very slow). In the end, we settled for a compromise technique inspired by Brown and Lowe [5] that applies different linear blending techniques at different image scales. The idea is to merge the low-frequency components of the real scene with the high-frequency components of the clip art, which makes the clip art appear "sharp" while obscuring details of the real image. We implemented this efficiently using Wavelet transforms.

3.2.4 Face Cartooning

Faces are particularly sensitive and common in streaming videos, so we handle them separately. To locate faces, we use a combination of upright person (pedestrian) detection and face detection. We first detect people using R-CNNs [15], identifying candidates by thresholding on high-confidence detections (above 0.8). To detect faces, we applied the face detector of Zhu and Ramanan [47], which is specifically designed for 'in-the-wild,' unconstrained datasets, to all of the candidate pedestrian regions. This technique not only identifies faces but also estimates the location and configuration of specific facial features.

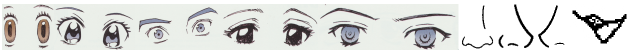
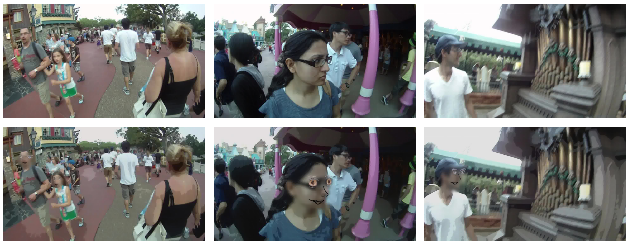


Figure 6: *Sample cartoon facial features.*



(a) (b) (c)

Figure 7: *Sample face cartooning results,* showing original images (top) and corresponding face cartoons (bottom).

Once faces are identified, we wish to replace them with cartoon representations that hide distinguishing facial features of particular individuals while reflecting their general properties (e.g. pose), while also assigning each face a distinct avatar so that different faces can be visually distinguished. To find recurring instances of the same individual across time, we extract Eigenface features [39] from all faces that are looking at the camera, and then cluster these features using Mean Shift [8]. We assign a distinct identity to each cluster, and then assign non-frontal-facing faces to the closest centroid in Eigenface space.

For each face cluster, we choose a distinct eye appearance, as illustrated in Figure 6. To produce the final cartoon output we blur the image using Bilateral filtering [38], blend eyes based on the identity label, and then render the nose and mouth based on the facial orientation estimated from the detected facial feature points. Some examples of face cartooning are shown in Figure 7. As the figure shows, the face cartooning system generally works well, with most failures caused by failing to detect faces, as with the people in the backgrounds of images (c).

4. Experiments

To evaluate our cartooning algorithm, including testing whether it was effective in obscuring private details while preserving important semantic-level details of a video, we applied the technique on video from three diverse datasets. We then conducted a user study that tested if participants could identify important properties of the cartooned videos, as well as their (in)ability to recognize private details.

4.1. Scenarios and datasets

We used three video datasets that reflect three real-world use cases for a privacy-preserving transformation system. *Indoor first-person* reflects the scenario in which someone wears a first-person video camera while inside their home, and may be concerned that private details like specific objects and information are collected. We used the publicly-available Activities of Daily Living (ADL) dataset [33] for this scenario, which consists of about 550 minutes of GoPro video in a simulated home environment. *Outdoor first-person* reflects a scenario in which someone is wearing a first-person camera in a public, outdoor space, and the main privacy concern is about capturing faces on video. For this scenario, we use the publicly-available First-Person Social Interactions dataset [13], which consists of more than 40 hours of video taken by GoPro cameras at amusement parks. *Video conferencing* reflects a scenario where a user is communicating with others using a fixed camera, as with Skype. In this scenario, the user wants themselves to be visible, but wants to obscure objects and people in the background of the scene. For this scenario, three of the authors recorded about 40 minutes of simulated Skype sessions in three different environments (home, office, and a public cafe) having busy, dynamic backgrounds.

4.2. User study design

We conducted a user study in which we showed people a selection of cartooned videos and collected their feedback. After asking background demographic information, the survey gave a series of videos which had been subjected to our cartooning technique. After viewing each video, participants were asked to answer questions about the videos and particular objects within them. The survey employed 7-level Likert scales.

Participants reviewed seven videos (presented in random order) taken from the three datasets described above, selected to represent several different use cases with different privacy concerns, and were chosen before viewing the output of the cartooning. Figure 8 presents still shots taken from the videos in our survey. The questions for each image asked participants to identify the objects identified by the green bounding box or to identify details in the scene such as the content of the television or the brand on a soda or shampoo bottle. We included several questions of this type to better understand how well our object replacement retained scene context but removed detail. We also included a free-form question for comments on the survey or videos.

We implemented the survey using Qualtrics and deployed it on Amazon Mechanical Turk. We required Mechanical Turk participants to be over age 18, live in the U.S., and have a lifetime approval rating of at least 95%. We received a total of 93 responses. Following standard practice, we included several 'attention check' questions on the survey with trivial answers to identify participants who were not answering questions carefully. Of the 93 responses, we removed 8 who either did not answer the attention questions correctly or were unable to view the videos due to technical problems. Our final sample thus contained 85 participants. The user study was reviewed and approved by the relevant ethics board at our institution.

4.3. Results

4.3.1 Participants

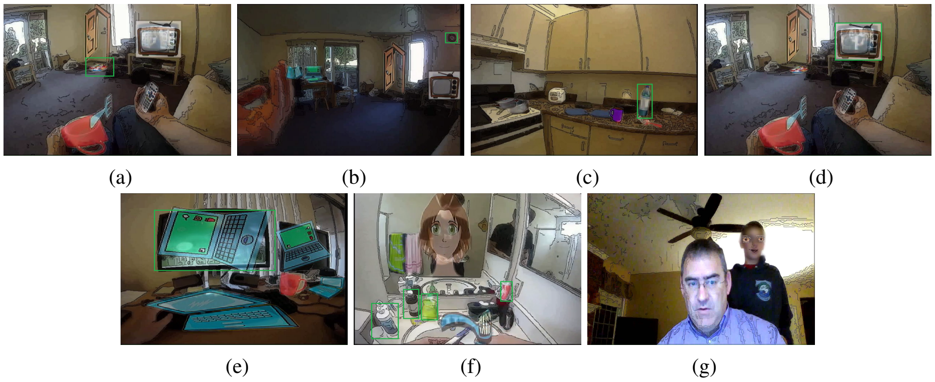
Of the 85 participants, 58% were male (n=49), 41% were female (n=35), and 85% (n=72) were aged 18–49. In terms of technology use, 95% (n=81) indicated that they used social networking applications or websites, 64% (n=52) shared videos on social networking apps or websites between a few times a month and several times a day, and 31% (n=25) shared videos a few times a week and 11% (n=9) shared videos between once and several times a day. In terms of education, 66% (n=56) had a bachelor's degree or higher.

4.3.2 Activity recognition

To understand the effect of object replacement on the video's semantic meaning, we asked participants to identify the activity occurring in four selected videos featuring four activities ("watching TV," "working on a PC," "brushing teeth," and "making tea"). Across participants, the accuracies on these four videos were 100 (n=85), 95.3% (n=81), 97.6% (n=83), and 90.6% (n=77), respectively. These results suggest that cartooning had little effect on most people's ability to interpret high-level video semantics.

We also asked participants to indicate how easy it was to identify the activities. Figure 10 summarizes these responses. For three of the videos, an overwhelming majority (88.3%) of participants felt it was slightly, moderately, or extremely easy to determine the activity in the video. The exception was Video 4, in which the camera wearer was making tea; here approximately half of participants indicated at least slight difficulty (although a majority, 90.6%, were able to do so). This difficulty is likely because it is only at the very end of the clip when a tea bag becomes visible.

Utility and privacy. To determine how well our system strikes the balance between abstraction and utility, we asked participants whether they could recognize the object category (to measure utility) as well as the object instance or content of a specific area of an image (to measure privacy protection). For utility, we asked participants whether



(a) (b) (c) (d)

(e) (f) (g)

Figure 8: *Images used to test participants' ability to recognize object category and instance detail.* For (a), (b), and (c), we asked about the category of object bounded by the green rectangle. For (c) we asked brand of the beverage, and for (d) and (e) we asked about content on the TV screen and computer monitor. For (f) we tested identification capability for both object category and instance. For the Skype video frame in (g), we asked whether users could identify the person in the background.

they could identify the general category of an object (such as shoe, clock, or beverage bottle) when the object was replaced by clip art. The three objects shown in Figures 8(a), (b), and (c) were correctly identified 95%, 96%, and 100% of the time, respectively. To measure the privacy, we asked the participants if they could identify the brand of the beverage or discern the content of the TV screen or computer monitor clearly (Figures 8(c), (d), and (e)). None of the participants could select the correct option for beverage brand or content for the TV screen, and 72% participants disagreed (from somewhat to strongly) that the content in the computer monitor was clearly visible. For Figure 8(f), 91% of participants agreed (35% strongly) with the statement "I can recognize the generic type of these objects (i.e. that they are some bottles), but I cannot identify the specific type (e.g. shampoo) or brand of any of the objects." This demonstrates that cartoon replacement enhances video privacy by abstracting detail with cartooned objects.

Finally, to understand the effect of the replacement clip art on individual objects and the semantic meaning of the videos, we asked the participants how much they agreed with the statement, "the video hides individual object details, but the overall activity and presence of objects in the scene is retained." Figure 9(a) shows mostly positive responses to this question.

Selective abstraction and streaming scenario. To measure the effectiveness of selective abstraction, we showed the participants a video where one person was using Skype, and another person was walking behind him (see Figure 8(g)). We added the cartoon effect to all the other objects except the primary user. All participants correctly identified the number of people in this video. 95% of them agreed that they would recognize the person in the foreground if he/she were familiar, but only 27% said the same about the person in the background. While 86% of the participants could recognize the activity in that video after watching only once, only 9% of the participants agreed (from somewhat to strongly) that "After watching the video once, I would not be able to recognize the person in the video, even if they were a friend." However, 62% of them agreed (from somewhat to strongly) with "If I wanted to share such events by live streaming, I believe the cartooned version will reduce privacy risks for both me and the surrounding people." This is strong evidence of usability of our system in such scenarios.

4.3.3 Usefulness of cartooning approach

Finally, we sought to understand the general usefulness of the cartooning approach. We asked participants whether they agreed with the statement: "If this were my video and I wanted to share it in a social networking site or to show it to other people, I would prefer to use the cartooned version instead of the original version." Figure 9(b) shows that our survey participants were more likely *not* to use the cartooned videos. We attribute this to several factors: the choice of the clip art used as replacement, the placement of the clip art in the scene and the semantics of the scene itself (limited by the data set). We believe that enhancements to the our initial prototype to address the first two factors will ameliorate these responses.

Over 36% (n=32) of the total survey participants gave responses in the free response portion of the survey, and these gave us additional insight into their thoughts concerning our object replacement approach, its application to privacy in videos, and how well the approach worked. Of the 32 respondents, 34% (n=11) used the word "interesting" to describe the idea of cartooning, and 37% (n=12) gave re-

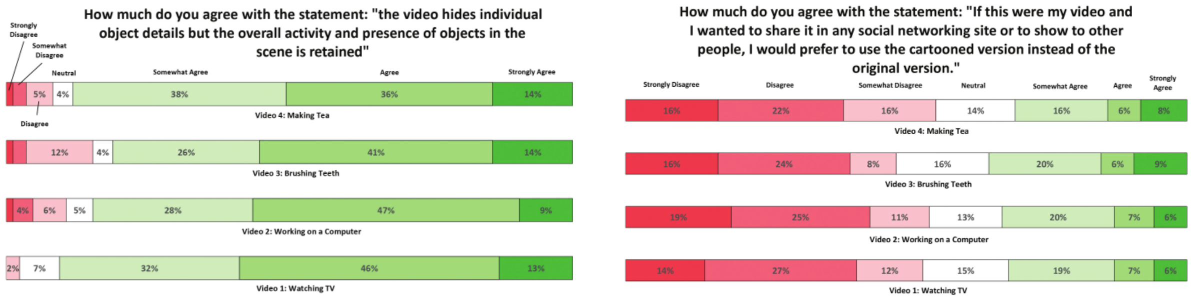


Figure 9: *Mechanical Turk Survey Results:* (a) shows how well we were able to maintain semantics of the video after object replacement (b) shows our participant's willingness to use cartoonized videos.

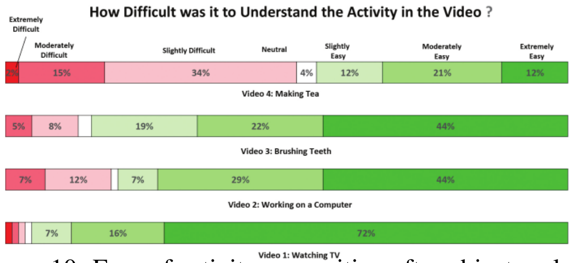


Figure 10: Ease of activity recognition after object replacement, according to our study participants.

sponses that were clearly supportive of the approach. Seven (22%) of the participants indicated that they would use a program that replaces objects with clip art.

Several participants specifically commented on the privacy implications of our approach:"Extremely interesting. I would be likelier (sic) to use social media if functionality like this were available," "I think this kind of video obscuring would be the most relevant to live streaming events in which not everyone has given consent to be filmed.", "It was kind of neat how the clip art was placed on the video. Would really help with privacy if that was someone's big concern."

On the other hand, 28% (n=9) of participants who responded in the free response section gave critical comments. Several did not like the idea of using cartoon clip art to replace objects: "The cartoon like quality was off putting and annoying to me. I would want a video's content to be clear and not have to guess what's going on.", "It looks creepy.", "I get the point of cartooning but I would not use it. It seems silly.", and "Cartooning just seems silly to me." One wondered why someone would share a video if they had privacy concerns, saying "The cartooning didn't really obscure a lot in my opinion, especially when it came to the faces. I think that if I had a problem with privacy issues, I just wouldn't share the video rather than try to make it unrecognizable." Another participant criticized the quality of the automatically-generated cartoons: "I felt the cartoonish images had very odd color profiles and that made me feel uneasy and uncomfortable." Another pointed out a potential unintended effect of our object replacement: "The presence of the cartoony faces sometimes compelled me to look even more keenly at features that would allow me to get a sense of what they might actually look like – even if I'd gloss over those non-distorted features otherwise."

In summary, the results of the survey suggest that the cartooning transformation concept has promise in preserving privacy while maintaining the semantics of the video. While some participants disliked the quality of cartooning in our prototype implementation, nearly a third responded positively to the potential of the approach, suggesting that cartooning will only become more practical and useful as technology in computer vision and graphics improves.

5. Conclusion and Future Work

We proposed enhancing privacy in videos in which objects are replaced by cartoon representations taken from clip art. We applied this approach to videos in several realistic use cases ranging from first-person video to video conferencing. Our user study suggests that people are open to the idea of cartooning as a privacy enhancing measure but are hesitant to use it until the aesthetic quality is improved. Despite the aesthetic issues, we demonstrated a good balance between retaining semantic meaning of the videos while removing potentially sensitive detail. The survey responses were mostly positive, and the critical comments provide useful feedback to inform future work. For the use of cartooning for enhanced privacy to be realized in practice, continued research should focus on improving the aesthetic qualities of object replacement and extending this work to address the privacy of activities involving multiple objects.

6. Acknowledgements

This material is based upon work supported in part by the National Science Foundation (under grants CNS-1408730 and IIS-1253549), Google, and NVidia.

References

[1] *Amazon Mechanical Turk.* `https://www.mturk.com/mturk/welcome` (accessed Mar 31, 2016).

[2] D. Bitouk, N. Kumar, S. Dhillon, P. Belhumeur, and S. K. Nayar. Face swapping: Automatically replacing faces in photographs. *ACM Transions on Graphics*, 27(3):39:1–39:8, Aug. 2008.

[3] A. Bousseau, F. Neyret, J. Thollot, and D. Salesin. Video watercolorization using bidirectional texture advection. *ACM Transactions On Graphics*, 26(3), july 2007.

[4] M. Boyle, C. Edwards, and S. Greenberg. The effects of filtered video on awareness and privacy. In *CSCW*, pages 1–10, 2000.

[5] M. Brown and D. G. Lowe. Automatic panoramic image stitching using invariant features. *International Journal of Computer Vision*, 74(1):59–73, 2007.

[6] J. Canny. A computational approach to edge detection. *TPAMI*, 8(6):679–698, jun 1986.

[7] H. Chen, N.-N. Zheng, L. Liang, Y. Li, Y.-Q. Xu, and H.-Y. Shum. Pictoon: A personalized image-based cartoon system. In *ACM International Conference on Multimedia*, pages 171–178, 2002.

[8] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *TPAMI*, 24(5):603–619, 2002.

[9] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, volume 1, pages 886–893, 2005.

[10] A. Erdélyi, T. Barát, P. Valet, T. Winkler, and B. Rinner. Adaptive cartooning for privacy protection in camera networks. In *AVSS*, volume 6, pages 44–49, 2014.

[11] A. Erdélyi, T. Winkler, and B. Rinner. Serious fun: Cartooning for privacy protection. In *International MediaEval Benchmarking Initiative for Multimedia Evaluation (MediaEval), Workshop, Barcelona*, page pp.2, 2013.

[12] Á. Erdélyi, T. Winkler, and B. Rinner. Multi-level cartooning for context-aware privacy protection in visual sensor networks, 2014.

[13] A. Fathi, J. K. Hodgins, and J. M. Rehg. Social interactions: A first-person perspective. In *CVPR*, pages 1226–1233, 2012.

[14] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2):167–181, 2004.

[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pages 580–587, June 2014.

[16] R. Gross, E. Airoldi, B. Malin, and L. Sweeney. Integrating utility into face de-identification. In *PETs*, pages 227–242, 2006.

[17] R. Gross, L. Sweeney, J. Cohn, F. Torre, and S. Baker. *Protecting Privacy in Video Surveillance*, chapter Face De-identification, pages 129–146. Springer London, 2009.

[18] J. Hays and I. Essa. Image and video based painterly animation. In *NPAR*, pages 113–120, 2004.

[19] J. He, B. Liu, D. Kong, X. Bao, N. Wang, H. Jin, and G. Kesidis. Puppies: Transformation-supported personalized privacy preserving partial image sharing. In *IEEE International Conference on Dependable Systems and Networks*, 2014.

[20] R. Hoyle, R. Templeman, S. Armes, D. Anthony, D. Crandall, and A. Kapadia. Privacy behaviors of lifeloggers using wearable cameras. In *UbiComp*, pages 571–582. ACM, 2014.

[21] W.-i. Hwang, P.-j. Lee, B.-k. Chun, D.-s. Ryu, and H.-g. Cho. Cinema comics: Cartoon generation from video stream. In *GRAPP*, page 299304, 2006.

[22] S. Jana, A. Narayanan, and V. Shmatikov. A Scanner Darkly: Protecting User Privacy from Perceptual Applications. In *IEEE Symposium on Security and Privacy*, pages 349–363, 2013.

[23] B. Knobel and J. Sanders. Samizdat 2.0: The dymovsky case and the use of streaming video as a political tool in contemporary russia. *International Journal of E-Politics*, 3(1):26–41, 2012.

[24] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[25] M. Korayem, R. Templeman, D. Chen, D. Crandall, and A. Kapadia. Enhancing lifelogging privacy by detecting screens. In *CHI*. ACM, 2016.

[26] P. Korshunov and T. Ebrahimi. Using face morphing to protect privacy. In *AVSS*, pages 208–213, Aug 2013.

[27] J. E. Kyprianidis and J. Döllner. Image abstraction by structure adaptive filtering. In *Proc. EG UK Theory and Practice of Computer Graphics*, pages 51–58, 2008.

[28] H. Li, G. Liu, and K. N. Ngan. Guided face cartoon synthesis. *IEEE Transactions on Multimedia*, 13(6):1230–1239, 2011.

[29] Y. Nakashima, T. Koyama, N. Yokoya, and N. Babaguchi. Facial expression preserving privacy protection using image melding. In *ICME*, pages 1–6, June 2015.

[30] C. Neustaedter, S. Greenberg, and M. Boyle. Blur filtration fails to preserve privacy for home-based video conferencing. *ACM Trans. Computer Human Interactions*, 13(1):1–36, Mar. 2006.

[31] J. R. Padilla-López, A. A. Chaaraoui, and F. Flórez-Revuelta. Visual privacy protection methods. *Expert Syst. Appl.*, 42(9):4177–4195, June 2015.

[32] P. Pérez, M. Gangnet, and A. Blake. Poisson image editing. In *ACM Transactions on Graphics*, volume 22, pages 313–318, 2003.

[33] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *CVPR*, pages 2847–2854. IEEE Computer Society, 2012.

[34] M.-R. Ra, R. Govindan, and A. Ortega. P3: Toward privacy-preserving photo sharing. In *NSDI*, pages 515–528, 2013.

[35] N. Raval, A. Srivastava, K. Lebeck, L. Cox, and A. Machanavajjhala. MarkIt: Privacy Markers for Protecting Visual Secrets. In *UbiComp '14 Adjunct*, pages 1289–1295, 2014.

[36] C.-H. Rhee and C. H. Lee. Cartoon-like avatar generation using facial component matching. *IJMUE*, 8(4):69–78, 2013.

[37] R. Templeman, M. Korayem, D. Crandall, and A. Kapadia. PlaceAvoider: Steering first-person cameras away from sensitive spaces. In *NDSS*, pages 23–26, 2014.

[38] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *ICCV*, pages 839–846, 1998.

[39] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991.

[40] J. Wang, Y. Xu, H.-Y. Shum, and M. F. Cohen. Video tooning. In *ACM SIGGRAPH*, pages 574–583, 2004.

[41] M. Wang, R. Hong, X.-T. Yuan, S. Yan, and T.-S. Chua. Movie2comics: Towards a lively video content presentation. *IEEE Transactions on Multimedia*, 14(3):858–870, 2012.

[42] H. Winnemöller, S. C. Olsen, and B. Gooch. Real-time video abstraction. In *ACM Transactions On Graphics*, volume 25, pages 1221–1226, 2006.

[43] Z. Xu, H. Chen, S.-C. Zhu, and J. Luo. A hierarchical compositional model for face representation and sketching. *PAMI*, 30(6):955–969, 2008.

[44] YouTube Statistics, 2016. `https://www.youtube.com/yt/press/statistics.html` (Accessed March 20, 2016).

[45] L. Zhang, T. Jung, C. Liu, X. Ding, X. Y. Li, and Y. Liu. Pop: Privacy-preserving outsourced photo sharing and searching for mobile devices. In *ICDCS*, pages 308–317, June 2015.

[46] W. Zhang, X. Wang, and X. Tang. Lighting and pose robust face sketch synthesis. In *ECCV*, pages 420–433, 2010.

[47] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, pages 2879–2886, 2012.