Balancing Privacy and Utility with Pattern Based Activity Detection

Cassandra Carley
Duke University Computer Science
Durham, NC 27708
carley@cs.duke.edu

Abstract

The diffusion of surveillance cameras often leads to conflicts between *utility*, that is, the benefits of preserving the information the camera records, and *privacy*, that is, the ability for the people being observed to conceal information they want to protect. For example, a camera monitoring an office kitchen may be useful in identifying a food thief, but might unintentionally reveal the PIN someone enters on a mobile phone. We design a video processing system that detects private activities in surveillance video and filters them out of the recording with minimal disruption of video quality. At the core of our system is the light-weight computation of a fixed-size feature that describes the spatio-temporal aspects of human activities that extend over variable amounts of time and space. Converting events of variable length and extent to a fixed-size descriptor makes it possible to use off-the-shelf classifiers to recognize and localize activities to be protected from recording. Comparisons of our descriptor with several alternatives show improved performance with less computation. We contribute two new video datasets recorded with a kitchen security camera, and we carry out a pilot user study to show that PIN theft is a valid concern.

Framework

Nowadays, video is being collected for public and private uses at locations including our homes, offices, airports, malls, and other public spaces. Cameras are either installed at fixed locations, mounted on vehicles, or body-worn. While collected video serves some purpose, such as catching a thief, identifying a person, or tracking someone's health, it also often contains aspects of behavior that are irrelevant to the application and needlessly infringe on a person's privacy. These aspects include the identity of people in the background or any text someone may be typing on a cellphone. A fundamental tension then arises between extracting useful information from videos and protecting privacy.

We propose a video-processing framework to resolve this tension. Our pipeline feeds frame-level descriptors to a classifier and applies a suitable privacy filter that pixelates or otherwise hides the information to be protected, while keeping the video useful for its intended purpose. The main role of the descriptor is to convert patterns of activity that extend over indeterminate amounts of time and over irregular regions of image space into a vector of fixed length. As a result of this conversion, any of a number of classifiers can be used to detect whether a privacy-sensitive activity is taking place. In particular, we use convolutional neural networks, cognizant of the recent successes of these architectures.

Lightweight Trajectory Descriptor

The traditional computer vision approach to activity detection is essentially generative: If the activity involves someone's hand, as it does in our initial experiments, this approach would involve designing a model of the hands shapes and their skeletons, and estimating skeletal motions from segmented image frames.

This approach is both daunting and unnecessary. First, hands move quickly, making tracking problematic. Second, finding the hand in the image is a difficult segmentation problem. Third, the interaction of hands and the objects they manipulate (such as a cell phone) causes occlusions and requires further modeling and estimation.

All this reconstruction is overkill if the goal is the classification of activities into private ant not. Instead, we bypass the generative approach by designing a fixed-length descriptor of activity in video that is estimated through a lightweight computation. The descriptor is then fed to a classifier that answers the question in a discriminative fashion.

Our descriptor treats activity as a sort of *motion texture*, akin to the representations used in the literature to describe natural processes such as smoke and waves. The presence of an object being manipulated (cellphone) affects this texture in informative ways, rather than making estimation harder, and a liability thereby becomes an asset.

To implement this idea, we first compute trajectories $\mathbf{p}(t)$ by detecting salient points in the video and tracking them over time. We then base the temporal aspects of our descriptor on the notion of auto-covariance:

$$\mathbb{E}[\mathbf{q}(t)\mathbf{q}(t+\tau)] \quad \text{where} \quad \mathbf{q}(t) = \mathbf{p}(t) - \mathbb{E}[\mathbf{p}(t)] \; .$$

In this way, the number of auto-covariance values per point is the number of time lags τ, a fixed design parameter. To capture the spatial aspects, we introduce a variant of the cell based representations used in HOG (Histogram of Oriented Gradient) (Dalal, Triggs, and Schmid 2006) or SIFT (Shift-Invariant Feature Transform) (Lowe 2004) methods. In our descriptor, each cell contributes a vector of auto-covariance values, one vector entry per time lag.

Privacy Filter A descriptor for each frame is fed to a classifier, and image regions deemed private are concealed. In a first implementation, we black out the entire frame. However, we plan to devise concealment methods that preserve the utility of the non-offending parts of the video as much as possible. These methods include blurring or pixelating near the fingertips. We intend to explore further options to achieve a good balance between privacy and utility.

Training, Evaluation, and Pilot User Study

We created an annotated data set for both training and evaluation of our system. The video in the data set was taken in an office kitchen, with a GoPro Hero 4 Session camera we placed next to an actual surveillance camera. While the surveillance system is meant to identify food thieves or vandals, it should not allow intruding on the privacy of the kitchen users. In particular, it should not be possible to discern anything anyone types on their cell phones while they wait for their coffee to brew. With these considerations in mind, we make the following contributions:

Training Data: Effectively training the classifier requires samples for both positive and negative examples of our target activity, typing on a cell phone. For positive samples, we must address different variations in activity including hand shape, motion speed, viewpoint, and phone appearance and setting. For negative samples, we include sequences where hands are not present, hands are not the focus, phones are not present, the phone is occluded by the hand, hands that interact with the phone in ways other than typing, and hands that are present but engaged in non-phone activities. These include holding the phone without typing, getting the phone out of the pocket, and opening the fridge. To achieve better generalization we use many negative samples of hands interacting and freely moving in proximity to a cell phone and several other kitchen and desk-related objects. In total we used 3 different users and 15 different PINs to capture 18 different video sequences, totaling $87,973$ frames or 24.4 minutes of footage at 60 fps. Each frame was annotated with a detailed label corresponding to the activity which was used to create positive and negative samples given the desired target activity (we looked at both PIN entry alone and a broader category of hand-phone interaction).

Evaluation: We use our data set to plot precision-recall curves for activity classification, and to evaluate the robustness of our approach to variations in hand shape, typing speed, viewpoint, and settings. We compare our descriptor (AC) with several baseline descriptors, including raw frames, frame differences, and point trajectories. Our aim is to showcase the advantages of spatial aggregation and the effectiveness of our fixed-size descriptor in encoding temporal information over varying time intervals. Preliminary results show our descriptor achieves superior results with a smaller footprint and less training time (Figure 1).

Pilot User Study: It may be argued that it is difficult to recognize what someone is typing from a surveillance video. We conducted a pilot user study to show that this privacy intrusion is a valid concern and to inform the design of an upcoming user study and verify equipment setup. We grouped video clips with different PINs, angles, and users into four sets of three clips and had each set assessed by a different set of three lab members. Aggregating observations, we found that it took an average of 71 seconds and 2 attempts to correctly guess a user's PIN and 32 of 36, or 89%, of attempts were successful.

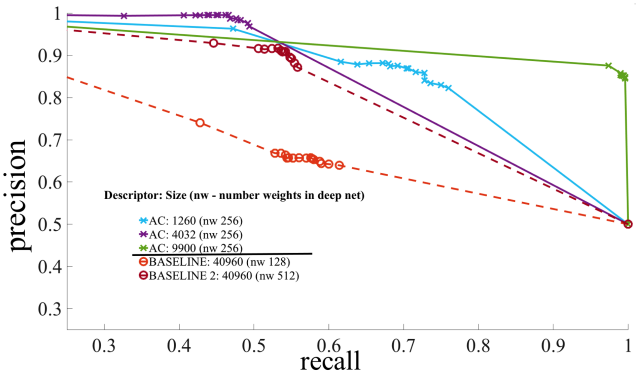

Figure 1: All descriptors consider a 100 frame history. Results show the superior performance of our AC descriptors (across several sizes) to the best BASELINEs (those built using raw image differences). For both BASELINEs the image was down-sampled to 64x64 pixels and sampled every 10 frames. Increasing the number of weights (nw) in the deep net increases the precision at minimal cost to recall. However, this requires an increase in training and testing time. Further, even with 512 net weights the BASELINE still performs worse than our AC descriptor. Experiments are done using $10,000$ training samples and $1,000$ testing samples, with an even split between positive and negative labels.

Summary and Future Work

We propose a system for the protection of private information in surveillance video, based on activity classification and an unobtrusive filter. A fixed-length descriptor of activities with variable temporal length and spatial extent allows using off-the-shelf classifiers, including convolutional neural nets. Preliminary experiments show promising precision, recall, and computational efficiency.

In future work we plan to improve activity localization in the image for minimal obtrusion, refine our descriptor for maximal discriminative ability, enlarge our training and evaluation data sets, and extend our demonstrations to other types of activities, including those that do not involve hands.

References

Dalal, N.; Triggs, B.; and Schmid, C. 2006. Human detection using oriented histograms of flow and appearance. In *European conference on computer vision*, 428–441. Springer.

Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60(2):91–110.