Crowdsourcing Approach for Evaluation of Privacy Filters in Video Surveillance

Pavel Korshunov
Multimedia Signal Processing Group, EPFL
CH-1015 Lausanne, Switzerland
pavel.korshunov@epfl.ch

Shuting Cai
Multimedia Signal Processing Group, EPFL
CH-1015 Lausanne, Switzerland
shuting.cai@epfl.ch

Touradj Ebrahimi
Multimedia Signal Processing Group, EPFL
CH-1015 Lausanne, Switzerland
touradj.ebrahimi@epfl.ch

ABSTRACT

Extensive adoption of video surveillance, affecting many aspects of the daily life, alarms the concerned public about the increasing invasion into personal privacy. To address these concerns, many tools have been proposed for protection of personal privacy in image and video. However, little is understood regarding the effectiveness of such tools and especially their impact on the underlying surveillance tasks. In this paper, we propose conducting a subjective evaluation using crowdsourcing to analyze the tradeoff between the preservation of privacy offered by these tools and the intelligibility of activities under video surveillance. As an example, the proposed method is used to compare several commonly employed privacy protection techniques, such as blurring, pixelization, and masking applied to indoor surveillance video. Facebook based crowdsourcing application was specifically developed to gather the subjective evaluation data. Based on more than one hundred participants, the evaluation results demonstrate that the pixelization filter provides the best performance in terms of balance between privacy protection and intelligibility. The results obtained with crowdsourcing application were compared with results of previous work using more conventional subjective tests showing that they are highly correlated.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*perceptual reasoning, representations, data structures, and transforms*; H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems—*evaluation/methodology, video*

Keywords

Privacy protection tools, crowdsourcing, video surveillance, evaluation methodology, intelligibility.

1. INTRODUCTION

The alarming rate at which video surveillance is being adopted has raised concerns among public and motivated development of privacy protection tools. Typical techniques (i.e., filters) used for obscuring personal information in a video in order to preserve privacy include blurring and pixelization of sensitive regions or their masking. More advanced privacy protection techniques have also been developed recently, such as scrambling [3], encryption of faces in video [1], obscuring [2] and complete removal of the body silhouettes [7], anonymization [6], etc.

However, there is a noticeable lack of methods to assess the performance of privacy protection tools and their impact on the surveillance task. While many evaluation protocols and tools (most notably those developed as part of PETS[1] workshops and grand challenges) are available to test the robustness, accuracy, and efficiency of video analytics for surveillance, little attention has been given to the privacy aspects. Therefore, a formal methodology for evaluation of the privacy protection is needed.

Since typical end user of privacy filters is a human subject, the ground truth required for evaluating their performance is also subjective. In this paper, we propose a subjective evaluation methodology to analyze the tradeoff between the preservation of privacy offered by privacy protection filters and the intelligibility of activities under video surveillance. We focus on several use cases of benign and suspicious behavior in indoor video surveillance, and apply commonly used privacy protection filters, such as blurring, pixelization, and masking to obscure the privacy-sensitive regions. Then, we ask human subjects to evaluate the resulting videos in terms of degree of privacy preservation and intelligibility of the surveillance events. The proposed evaluation method allows to identify the weaknesses of existing privacy protection tools and provide a reference for evaluation of other techniques.

One important objective in organization of subjective evaluations is getting enough reliable subjects to participate in the study. A possible solution is crowdsourcing, which is an increasingly popular approach for solving problems benefiting from large number of participants. Since evaluation scenario considered in this paper assumes a human guard sitting in a surveillance room and observing monitors with privacy protected video from surveillance cameras, the crowdsourcing seem to be a well-fitting and useful approach.

For the crowdsourcing based evaluation, we have built an online application VideoRate[2] that utilizes Facebook ID for login and reliable user authentication. Compared to laboratory based evaluations, this approach allows some flexibility to users, as one could stop participation in the experiments at any time, as well as simulates the real-time scenario better, since the users evaluate videos in different lighting conditions and using monitors with different

[1]IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)
[2]https://ltslinux18.epfl.ch/

resolutions and color settings. We disseminated the call to participate in the subjective test using VideoRate application via such social networks like Facebook, Twitter, and LinkedIn, as well as various research mailing lists. With an estimated outreach to more than 1500, some 120 among them used the application and submitted subjective scores. We then compared these subjective results with the results from a similar evaluation conducted by a conventional approach in a designated research test laboratory at EPFL. The results demonstrate high correlation with only some minor differences favoring the crowdsourcing method, which means that it can be considered as a reliable and effective approach for subjective evaluation of visual privacy filters.

VideoRate application was built using Facebook platform, because Facebook's users are generally verified individuals with very little number of them with fake IDs. Since we disseminated the application only to either friends, family, or friends of friends, we have, in a way, propagated trust (since we can trust our friends, and they can trust their friends). Such measures insured significantly more reliable results of the subjective tests, compared to a classical crowdsourcing scenario. Otherwise, we could use statistical analysis for determining the behavioral outliers as presented in [4]. Also, we used Facebook, instead of such tools like Amazon Mechanical Turk[3], to avoid paying people, which could lead to situation when people maximize their profits at the expense of honest evaluation results. Therefore, our evaluation environment is trustworthy and, by extension, we utilize trustworthy mechanisms to insure that the results are reliable.

2. EVALUATION METHODOLOGY

This section describes the evaluation methodology based on crowdsourcing that is designed for effectiveness assessment of the various visual filters to protect privacy of individuals on one hand, and their impact on the intelligibility of the surveillance task on the other. We start with describing the underlying use cases and dataset used in the evaluations, followed up with details on evaluation protocol and the crowdsourcing tool.

2.1 Use cases and underlying database

Privacy and surveillance are both heavily context dependent concepts. Therefore, any evaluation methodology should take into account the context in which the surveillance task is performed. In this paper, we focus on a simple use case, namely, a monitoring situation, without recording, where an observer (test subject) watches a video of an indoor scene under surveillance with a single standard definition camera. In the monitored scene, individuals move in front of the camera, either behaving normally, or acting abnormally.

To evaluate this use case, we used a dataset consisting of 9 different video sequences with a duration of 10 seconds each. Different indoor video surveillance scenarios were considered, such as a person walking towards and away from the camera (normal scenario), blinking into the camera (suspicious), and wearing hat, sunglasses, or scarf around the mouth (suspicious) to hide personal identity. Table 1 provides a short description of each video sequence in the database.

To each video sequence in the dataset, a semi-automatic segmentation and tracking algorithm was applied in order to obtain a binary mask[4], identifying a foreground object of interest, which not only plays a certain role in the understanding of the situation under surveillance, but also may contain potentially privacy sensitive information. Different privacy protection filters were then applied to the extracted foreground objects. Blurring, pixelization, and masking (black foreground shape covering the region of interest) privacy filters were selected (see examples in Figure 1) to generate different versions for each video sequence. Thus, a total of 27 processed video sequences were produced and used in the subjective evaluation, as described in the next section.

Table 1: Description of Video Sequences Used in the Evaluation

Seq. 1	White male, sunglasses, walks away and towards the camera
Seq. 2	White female, walks towards the camera, blinks three times
Seq. 3	Asian male, glasses, walks in from the right side, blinks three times to the camera
Seq. 4	White male, walks toward the camera, blinks three times
Seq. 5	Asian female, walks towards the camera, blinks three times
Seq. 6	White female, walks toward the camera, blinks three times
Seq. 7	Asian male, glasses, walks toward the camera, blinks three times
Seq. 8	White female, walks toward the camera
Seq. 9	White female, wears scarf around her face, walks toward the camera

2.2 Evaluation Protocol and Crowdsourcing Application

The overall goal of the subjective evaluation was to assess whether the detection of the normal or abnormal behaviors in the scene was possible, while various privacy protection filters were applied. At the same time, the effectiveness of privacy protection was assessed, as the identities of the individuals in the sequences might have been hidden. Particularly, each subject, i.e., the user of VideoRate application, was asked to watch a video sequence and then answer to questions presented in Table 2.

Table 2: Questions Asked During the Assessment

1. What is the gender of the person?	☐ Female	☐ Male	☐ I don't know
2. What is the race of the person?	☐ White	☐ Asian	☐ I don't know
3. Does the person wear glasses?	☐ Yes	☐ No	☐ I don't know
4. Does the person wear sunglasses?	☐ Yes	☐ No	☐ I don't know
5. Does the person wear a scarf?	☐ Yes	☐ No	☐ I don't know
6. Does the person blink into the camera?	☐ Yes	☐ No	☐ I don't know

An important issue to resolve was the memory effect during viewing, when observation of a video could potentially affect the evaluation of a following video. In our case, the main concern came from interactions between different versions of the same video, since different details could be visible in different video of the same scene obfuscated by different privacy filters. For instance, observa-

[3]https://www.mturk.com/

[4]MIT annotation tool:
http://people.csail.mit.edu/celiu/motionAnnotation/

(a) Original, no filters

(b) Blurring

(c) Pixelization

(d) Masking

Figure 1: An Example of Video Sequence (Seq. 9 from Table 1) with Privacy Filters Applied

tion of a blurred video could provide information otherwise invisible in the pixelated version of the same video. Consequently, if the former precedes the latter, the memory effect could affect the evaluation of the latter.

To avoid such memory effect in the assessment of the privacy protection and the task performed in video surveillance, each subject was shown each of the 9 contents only once. To insure that, the 27 processed video sequences were divided into three disjoint sets designated as A, B, and C, with each set containing 9 sequences including all the different contents. Furthermore, every set contained an equal number of blurred, pixelated, and masked video. Table 3 illustrates how video sequences were divided into these three sets.

A set of video sequences (A, B, or C) is chosen randomly for each user at his/her first login to VideoRate application. A user is then presented with a formal disclaimer stating the research purpose of the web application and that the personal data is not going to be kept longer than the related research activities. Once a user agrees with the disclaimer, application displays its home screen with a notice describing the nature of the evaluation and the suggestion to try a demo video sequence, which helps users understand how to proceed with the actual evaluation. Then, the user can choose to press a start button, after which, one of the video sequences (randomly chosen) from the preselected set (A, B, or C) is played automatically. After the video is finished playing, the questions from Table 2 are presented. The user is asked to answer all questions by clicking the corresponding radio buttons. Once the answers are submitted, the user is brought back to the home screen and can continue to view the rest of video sequences (which will be displayed in order according to Table 3) from the set until all 9 video sequences are finished.

Video sequences in VideoRate are played at the native camera resolution 640×480 with 30 fps. Videos were encoded with MPEG-4 and placed in Ogg Vorbis and MP4 file formats to support the playback by different browsers. The application is able to correctly run on most of the commonly used browsers including mobile versions, except Internet Explorer 6 and 7, since they do not support HTML5, which was used in VideoRate.

A few important aspects of the evaluation process should be noted. No video controls are given to the user while a video sequence is played automatically, thus, preventing user from stopping a video to view it in details. Once a video sequence is played, it cannot be played again. Such restrictions allow us to simulate the studied real-time video surveillance scenario. By randomly choosing a set of 9 video sequences and which video from this set to play first, we insuring an equal chance for each of the total 27 video sequences to be evaluated by subjects. Also, such arrangement insures that every user has a balanced overview of the used privacy filters, which helps avoiding bias in the results. The application allows the user to stop the evaluation session at any time. When the user later returns back to the application, it is possible to continue

evaluating the remaining (those that are not played yet) video sequences form the set. In case a video was already played, but no answers were submitted, it will not be counted towards evaluation results.

Table 3: The Arrangements of the Filtered Video Sequences into Evaluation Sets A, B, and C

Seq.	Blurring	Pixelization	Masking
Seq. 1	A	C	B
Seq. 2	B	A	C
Seq. 3	C	B	A
Seq. 4	A	C	B
Seq. 5	B	A	C
Seq. 6	C	B	A
Seq. 7	A	C	B
Seq. 8	B	A	C
Seq. 9	C	B	A

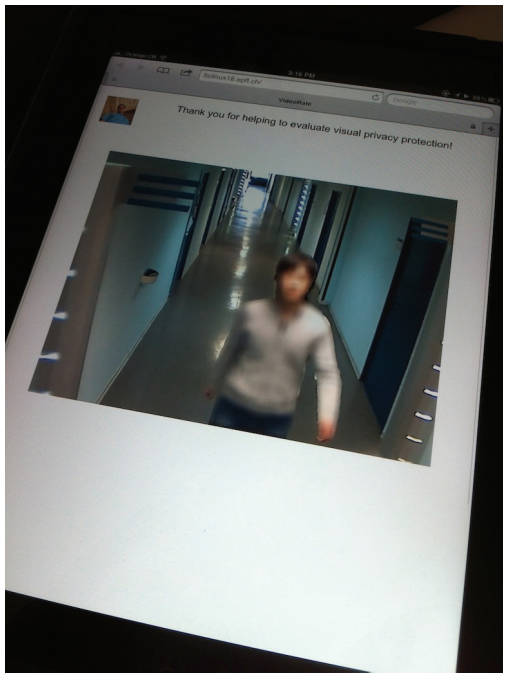

Figure 2: A Photo of VideoRate Application Running on iPad.

3. EVALUATION CRITERIA

Given the context dependent nature of privacy and intelligibility, in the surveillance scenario under consideration, the first three questions from the Table 2 are assumed to be relevant to privacy and the last three questions to intelligibility. Information about gender, race, and glasses (first three questions) are privacy related. These characteristics do not carry anything unusual, given the surveillance scenario, while they can be used to identify people in the indoor environment, and they can be discriminated against based on these features. Therefore, such privacy features should not be recognizable to the observers. On the other hand, blinking three times into the camera, which looks like a sort of secret code (at least, it's an unusual behavior), sunglasses worn indoor (possibly for hiding eyes), and scarf around the face (to hide the identity) are considered unusual and alarming, since neither of these characteristics are typical for an indoor environment. These unusual features therefore are set as related to intelligibility and should be visible to the observers.

Therefore, the following criteria is used for understanding how well a given filter protects privacy. If an observer correctly answers the privacy related question for a given video sequence and privacy protection filter, the privacy is not protected in this case. Incorrect answer or no answer (option "I don't know") means that the privacy is preserved. For intelligibility, on the other hand, a correct answer to the corresponding question means that surveillance task can be performed successfully, while incorrect or uncertain answers leads to the failure of recognizing an important unusual event.

Such tradeoff between privacy and intelligibility can be used to compare different privacy protection techniques and understand how these techniques perform, given various video contents.

If an observer correctly answers to the privacy related question, the privacy value is 0, since the privacy was not protected in this case. Incorrect answer or no answer (option "I don't know") yield 1. Then, the average privacy score of all three privacy related questions across all test subjects is computed for each type of filter and each video sequence. For intelligibility, it is reversed: correct answers to intelligibility related questions result in value 1 and incorrect or no answer in value 0.

4. EVALUATION RESULTS

More than 1500 people were reached out with a call for participation in the online crowdsourcing evaluation via several social networks, such as Twitter, Facebook, and LinkedIn, as well as several research mailing lists. The total number of people that ended-up using VideoRate application was 120, which was achieved over the course of 7 days from when the application was released to the public.

As mentioned earlier, each user could only view one out of three sets of video sequences, therefore, the number obtained evaluations scores per filtered video oscillated between 33 to 43, with average 38 subjects providing a score for each video. It can be noticed that not all users completed their sets of video sequences, otherwise all video sequences would be evaluated by about 40 subjects (120 total divided by 3). Among all users, 22 were female. Participants had highly diversified ethnical and cultural backgrounds with well distributed geographical locations. The age of the users varied from twenties up to late forties.

For each privacy filter, the aggregated results are illustrated on a two dimensional space in Figure 3, with the amount of privacy preservation and the degree of recognition of activities under surveillance (i.e., intelligibility), as vertical and horizontal axes respectively. The privacy score is ranging from 0 (no privacy protection) to 1 (fully protected), which is the average of the scores of the privacy related questions from the test subjects, as described in the previous section. Each point in the figure corresponds to a different video sequence and a different privacy protection filter. Points corresponding to a privacy filter are marked with distinguishing point-style.

The best privacy preserving filter would be a blacked out camera with no video feed, but, in such a case, there would be no surveillance possible and intelligibility would be zero. Therefore, a usable privacy protection filter should have a balance between privacy and intelligibility. In an ideal situation, the evaluation scores for such filter would lie in the top right corner of the tradeoff graph, having the highest values of privacy and intelligibility.

Figure 3, with evaluation scores of the typical privacy filters, demonstrates that blurring filter yields the highest intelligibility

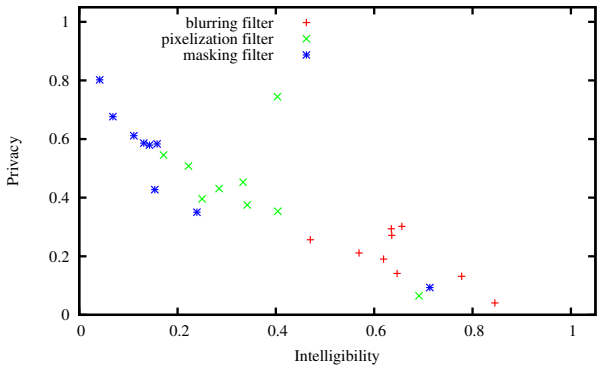

Figure 3: Intelligibility vs. Privacy for Different Filters from Crowdsourced Data

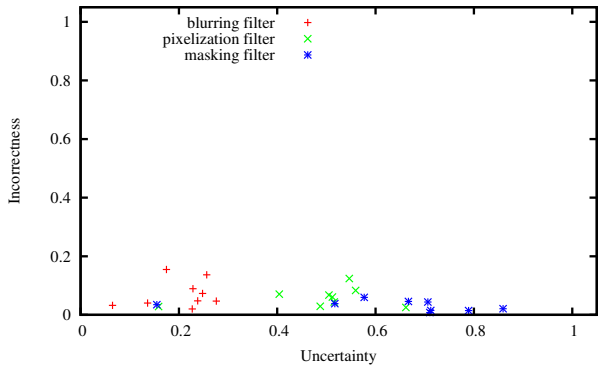

Figure 4: Uncertainty vs. Incorrectness for Different Filters from Crowdsourced Data

while providing the lowest privacy protection. Not surprisingly, the masking filter shows the highest privacy protection, while having the lowest intelligibility, since a person from the video sequence is replaced with a black silhouette. However, the highest privacy score for the masking filter is still below 0.8, which means that at least 20% of the answers to the privacy questions were still correct. By looking into details, we noticed that the largest number of correct answers for masking filter is to the gender question, which is because people can recognize gender by the shape of a person in the video. Therefore, the shape of persons' masks should be distorted to hide the actual shape of the person. A surprising result shows pixelization filter demonstrating high privacy protection while still yielding high degree of the activities recognition, which makes it the filter with the best balance of privacy and intelligibility.

It can be noted in Figure 3 that one video sequence demonstrates an odd results for every filter, having the smallest value of privacy and a significantly high intelligibility (the points of each different color with the lowest privacy scores). In this video, the face of the person walking from a distance was left visible (unprotected by a filter) just for a couple of frames, which immediately rendered privacy protection filters useless. This video sequence indicates that even a slight inaccuracy or inconsistency in the way the filters are applied can lead to the complete loss of the privacy protection effort.

Figure 4 demonstrates the effect of different privacy protection filters on the uncertainty and incorrectness in the answers of the test subjects. Uncertainty axis reflects the average normalized amount of "I don't know" answers, which were given to both privacy and intelligibility questions. Incorrectness is computed as normalized average of wrong answers, when subjects were certain but wrong. Each point on the graph corresponds to one video sequence distorted by one privacy protection filter. This figure shows masking filter yielding the largest uncertainty, while blurring filter results in the largest false positive (incorrectness). Such unbalance indicates that blurring filter is less applicable in the surveillance scenarios with little tolerance for false positives. The masking filter on the other hand would be better in a typical surveillance system when some uncertainty can be tolerated, i.e., an uncertain observation can be checked via other means, such as an additional security check, but false positive is required to be low.

4.1 Comparing with Conventional Evaluations

Similar to the described crowdsourcing evaluations with VideoRate application, the set of subjective tests of visual privacy filter was performed previously in a designated evaluation laboratory [5]. Only 36 subjects participated in these offline tests with 12 subjects evaluating each video sequence. To understand the differences and consistencies between conventional laboratory testing and crowdsourcing approaches, in this section, we compare both sets of evaluation scores.

There are few notable differences in how offline and online experiments were conducted, which may affect the final evaluation results:

- In offline case, several subjects (up to 6) were seated in designated laboratory controlled environment for subjective tests with the only task to evaluate video sequences, while in online experiments, users watch video sequences at their convenient time and place.
- In the offline evaluations, video sequences were played at fullscreen 1280×1024 resolution compared to 640×480 camera resolution played online.
- Subjects were given strictly 25 seconds to answer the questions after each video sequence in the offline tests, while no such constraints were imposed in online experiments.

We plot evaluation scores of online tests vs. offline tests on Figures 5 and 6 for privacy and intelligibility related questions respectively. Points corresponding to blurring, pixelization, and masking privacy filters are marked with different point-styles. These figures demonstrate that scores of online and offline evaluations are highly consistent with each other, especially in the case of privacy, since the points fit to the diagonal lines of the figures quite well. To have a better understanding of this correlation, we also computed Spearman and Pearson correlation coefficients presented in Table 4. The high values of Spearman and Pearson indicate high correlation of the corresponding scores.

It can be noted however that there is a higher correlation of the privacy scores compared to the intelligibility scores. Spearman coefficient for the intelligibility of pixelization privacy filter is especially low. If we analyze Figures 5 and 6 in more details, we can notice that privacy points mostly lie above the diagonal line, while intelligibility lie below the diagonal. It means that the participants in the offline tests were more accurate in general. Such tendency can be explained by the larger resolution of the played video sequences in offline tests (1280×1024) compared to online (640×480), which allowed offline participants to notice finer details in the video sequences. And since intelligibility is based on the questions requiring notice of the finer details than privacy (i.e.,

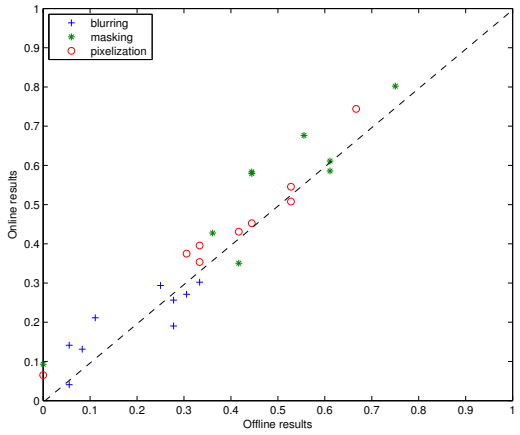

Figure 5: Privacy of Online vs. Offline Evaluations

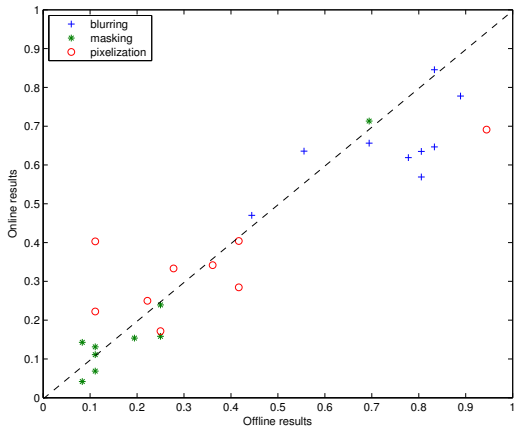

Figure 6: Intelligibility of Online vs. Offline Evaluations

blinking, see Table 2), the intelligibility scores of online evaluation deviate from the offline counterparts more significantly.

5. CONCLUSION AND FUTURE WORK

This paper defines a crowdsourcing online based methodology for evaluation of privacy protection tools for video surveillance. In the proposed evaluation protocol, we focus on two important aspects: (i) how much of the privacy is protected by such tool and (ii) how much it impacts the efficiency of the underlying surveillance task (intelligibility). The pixelization filter shows the best performance in terms of balancing between privacy protection and allowing high intelligibility. Masking filter, on the other hand, demonstrates the highest privacy protection with low incorrectness and high uncertainty, which can be suitable for the higher security surveillance applications.

The results of the online crowdsourcing subjective evaluations were also compared with results obtained from offline tests conducted in a laboratory under controlled environment. In general, there is high correlation between these two sets of results, with online results demonstrating a more modest accuracy specifically for intelligibility scores, which rely on observations of finer visual details. Overall, the crowdsourcing based results are more preferable, because the evaluation setup simulates the real-life surveillance scenario better, and it is easier and faster to get large number of diverse participants.

Table 4: Pearson and Spearman Correlation Coefficients of the Results from Online and Offline Evaluations.

Aspects	Filters	Pearson	Spearman
privacy	blurring	0.85	0.82
	masking	0.94	0.92
	pixelization	0.98	0.97
intelligibility	blurring	0.66	0.61
	masking	0.98	0.85
	pixelization	0.82	0.56

Future work includes extending the set of evaluation questions to identify other tradeoffs in privacy protection. The effect of applying protection tools with different levels of strength also need to be evaluated. We also plan to extend the dataset to include both more content and several additional privacy filtering tools.

6. ACKNOWLEDGMENTS

The authors would like to thank Zdenek Svachula for help in generating some of the masks used in this work and with the score sheets. The authors also thank EURECOM for generously providing the video dataset used in the evaluations and Claudia Araimo for helping preparing it. A special thanks to Francesca De Simone for fruitful discussions and for help in editing few drafts of the paper. This work was conducted in the framework of the EC funded Network of Excellence VideoSense.

7. REFERENCES

[1] T. E. Boult. PICO: Privacy through invertible cryptographic obscuration. In *IEEE Workshop on Computer Vision for Interactive and Intelligent Environments*, pages 27–38, Lexington, KY, Nov 2005.

[2] S.-C. S. Cheung, M. V. Venkatesh, J. K. Paruchuri, J. Zhao, and T. Nguyen. *Protecting privacy in video surveillance*, chapter Protecting and Managing Privacy Information in Video Surveillance Systems, pages 115–128. Springer-Verlag, 2009.

[3] F. Dufaux and T. Ebrahimi. Video surveillance using JPEG 2000. In *proc. SPIE Applications of Digital Image Processing XXVII*, volume 5588, pages 268–275, Denver, CO, Aug 2004.

[4] I. Ivanov, P. Vajda, J.-S. Lee, and T. Ebrahimi. In tags we trust: Trust modeling in social tagging of multimedia content. *IEEE Signal Process. Mag.*, 29(2):98–107, 2012.

[5] P. Korshunov, C. Araimo, F. D. Simone, C. Velardo, J.-L. Dugelay, and T. Ebrahimi. Subjective study of privacy filters in video surveillance. In *accepted in 14th international workshop on multimedia signal processing MMSP2012*, Banf, Canada,, September 2012.

[6] C. Velardo, C. Araimo, and J.-L. Dugelay. Synthetic and privacy-preserving visualization of video sensor network outputs. In *5th ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC'11)*, pages 1–5, Ghent, Belgium, Aug 2011.

[7] J. Wickramasuriya, M. Datt, S. Mehrotra, and N. Venkatasubramanian. Privacy protecting data collection in media spaces. In *Proceedings of the 12th annual ACM international conference on Multimedia (ACMM'04)*, pages 48–55, New York, NY, USA, Oct 2004.