Characterizing and Detecting Non-Consensual Photo Sharing on Social Networks

Tengfei Zhange, Tongqing Zhou, Qiang Liu, Kui Wu, Zhiping Cai

ABSTRACT

Photo capturing and sharing have become routine daily activities for social platform users. Alongside the entertainment of social interaction, we are experiencing tremendous visual violation and photo abusing. Especially, users may be unconsciously filmed and exposed online, which is termed as the non-consensual sharing issue. Unfortunately, this problem cannot be well handled with proactive access control or dedicated bystander detection, as users are unaware of their situations and may be filmed stealthily. We propose *Videre* on behalf of the privacy of the unaware parties in a way that they would be automatically identified and warned before such photos go public. For this, we first elaborate on the predominant features encountered in non-consensual captured photos via a thorough user study. Then we establish a dataset for this context and build a classifier as a proactive detector based on multi-deep-feature fusion. To relieve the burden of person-wise unawareness detection, we further design a signature-based filter for local pre-authorization, which can also implicitly avoid classification error. We implement and test *Videre* in various field settings to demonstrate its effectiveness and performance.

CCS CONCEPTS

• **Security and privacy** → **Social network security and privacy**;
• **Human-centered computing** → **Social content sharing**;

KEYWORDS

Social network; Image sharing; Privacy preserving

1 INTRODUCTION

The pervasive use of camera-equipped smart devices has promoted wide real-world photo capturing, whereas the development of online social networks has facilitated intensive sharing of these visuals as everyday social behaviors. As reported, more than 3 billion and 4.5 billion images are uploaded to Snapchat [1] and WhatsApp [2] per day.

The unprecedented photo capturing and sharing activities put people in an awkward situation where they may be seen anytime, anywhere, thus posing severe threats to their privacy. In fact, even those cautious users with conservative social sharing practices would unavoidably and unconsciously be filmed during others' photographing, either intentionally (i.e., stealthy photographing) or unintentionally (i.e., as bystanders), which is termed as a non-consensual photo sharing issue in this paper. These photos, depicting the visuals of people who are unaware that they have been photographed[1], will accidentally disclose their life circles and daily activities if shared with the public [3–5]. Even worse, they may disseminate rapidly on social networks and reside in cloud servers, rendering privacy leakage to a broader scope and a longer period.

For example, celebrities, dignitaries, and their families are often stalked by paparazzi, live-streaming on TikTok, and pranks (e.g., recorded by some friends' mobile phones). A related event is that JK Rowling and her infant son were secretly photographed by the Big Pictures (UK) Ltd on a public street. They were unaware of any of these pictures until one was published without warning [6]. Another event is the publication of secretly taken photos of Naomi Campbell, leading to a significant court case [7]. Although both won their cases, they claimed to have suffered distress, embarrassment, and anxiety as a result of an invasion of privacy. Therefore, the problem of non-consensual photo capturing and sharing should be given adequate attention to protect the privacy of the corresponding unaware parties.

Many technical efforts to protect the privacy of photographed persons have been explored, typically by preventing photos from being taken secretly [8–11] or designing access policies to control the spread of photos on social networks [12–15]. On the one hand, some [8–11] require the proactive involvement of users by installing specific applications on the photographic equipment to detect unauthorized photo capturing. These proposals are unscalable as they are not applicable to traditional cameras, and a malicious photographer can easily bypass the detection by disabling these applications. On the other hand, access control schemes allow users to build privacy policies in advance to control whether the photo can be shown to specific viewers. However, since our non-consensual photo sharing cases are unpredictable, building an access policy beforehand [12, 13] or on a single-photo granularity [14, 15] is not feasible in practice.

The seemingly most relevant literature to non-consensual photo sharing is bystander detection. In a pioneering work in [4], a machine learning model is trained to identify bystanders in photos automatically. Nevertheless, these two terms/problems are different by definition. Namely, 'bystander' refers to one who is not intentionally captured by the photographer [4] (as shown in the right two photos in Fig. 1), while **'unaware parties' in non-consensual photo sharing denotes, more generally, one who would be more vulnerable to privacy violation if being captured** (as shown in the middle two photos in Fig. 1). As such, non-consensual photo sharing covers the severer, yet previously ignored cases of stealthy photographing. In fact, as we will show in § 6.2, the dedicated features for bystander detection used in [4] are ill-suited for depicting non-consensual photographing of unaware persons (with an average overall accuracy of ~ 60%). To this end, how to properly prevent visual privacy leakage of those unaware involvements still remains an open problem.

In this work, we discover the quantifiable characteristics of non-consensual photo sharing and propose an automatic detector sensitive to possible violation sharing behaviors. From a high level view, given the unpredictable situations (anytime, anywhere) of being

[1]We utilize the term 'unaware parties' to refer to the people who are unaware that they have been photographed.

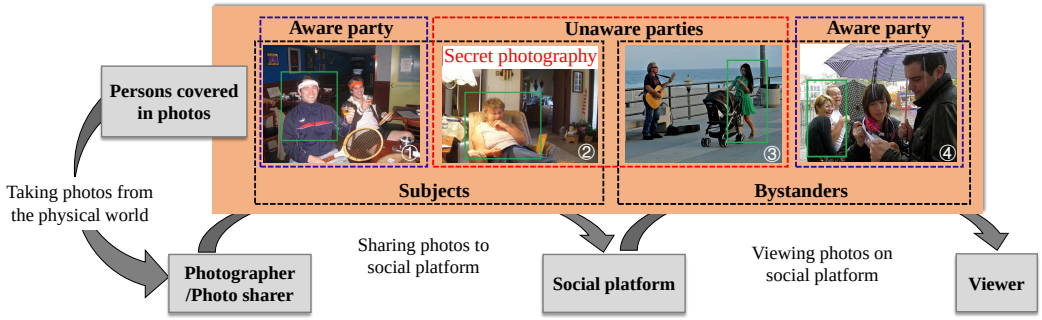


Figure 1: A general photo sharing practice process. The four example images are shared photos, and the snapped individuals can be divided into different categories according to their statuses.

unconsciously photographed, the detector shall work reactively as a privacy-preserving proxy in the platform, reminding the users whenever necessary while avoiding tedious beforehand setup.

Specifically, due to the lack of technical understanding of non-consensual photo sharing, we first conducted two user surveys. One focuses on studying the predominant features of subjective perception of unawareness in shared photos. The elaborated observations help characterize the visual cues that can be statistically found in non-consensual captured photos. The other focuses on studying the user behavior preferences about non-consensual photo sharing, which helps us establish a baseline for user habits/experience. Further, we build a photo dataset with unaware person annotation to fill the gap of lacking relevant data resources in this context. Finally, we propose the framework of *Videre* [2]. Briefly, it leverages the fused features elaborated in the user study for joint classification to identify unaware individuals in uploaded photos and uses a signature-based filter for locally collecting authorization and reducing the burden of person-wise classification.

We highlight that the reactive property of *Videre* makes it a scalable privacy guardian for all the "sensitive" while "unaware" users in social networks. To the best of our knowledge, *Videre* is the first framework for automatically identifying non-consensual photos on social networks. The main contributions of this paper are summarized as follows:

- We conduct a real-world user study for subjectively characterizing the unawareness in shared photos and find that eye gaze direction and head orientation form two predominant cues to identify non-consensual photo sharing.
- We establish a publicly available dataset, with precise manual annotations on unawareness for each person in the images, in order to promote relevant research.
- We propose the design of *Videre* by training a multi-feature fusion machine learning classifier for unaware party detection. *Videre* achieves efficiency through a pre-authorization phase performed locally with certificateless aggregate signature.
- We perform a broad spectrum of experiments and field survey to evaluate the performance of *Videre*. Our classifier achieves an accuracy of 85.1% and an F1-measure of 0.849, showing a large improvement compared with naively using the detector designed for bystander identification. It is also evaluated to yield acceptable computation and communication overhead.

[2]The Latin word for seeing and perceiving

2 RELATED WORK

Our work is closely related to the line of work for bystander privacy protection.

Various methods focus on protecting the privacy of nearby people (bystanders), which may be used to protect the unaware parties' privacy. Aditya et al. [9] presented I-Pic, where users choose a level of privacy (e.g., image capture allowed or not) based upon social context. Privacy choices of nearby users are advertised via short-range radio, and I-Pic edits the photo based on the received choices. Zhang et al. [8] designed COIN, that enables a user to broadcast his privacy requirement in much the same way as I-Pic [9] and empowers the photographer to process photos, such as erasing and tagging. Ra et al. [10] proposed Do Not Capture (DNC), where bystanders who do not want to be photographed broadcast their facial features using a short-range radio interface. When a photo is taken, the bystanders are identified by their facial features, and their faces are then blurred. Li et al. [11] proposed PrivacyCamera, which works as an App on both the photographer's and the stranger's mobile phone. When taking a photo, it can notify nearby strangers based on the GPS coordinates via peer-to-peer short-range wireless communications. Although these solutions can prevent people from being photographed unawares, they require the photographer and the bystanders to be proactive [4]. Moreover, a malicious photographer can refuse to install these software products or APPs to evade regulation.

Another set of proposed solutions attempts to design access control policies to protect bystanders' privacy, which also can be used to reduce privacy risks of the unaware parties. Henne et al. [13] proposed SnapMe, where users need to mark some locations as private in advance. Users who have marked it as private will be notified if a photo is photographed in such a location. Li et al. [12] proposed a plugin for social networking websites called HideMe, which allows users to build a scenario-based access control model by combining temporal, spatial, interpersonal, and attribute factors. Registered users can decide to blur/show their faces to photo-viewers for each scenario. A major drawback of these policy-based solutions is that these access control strategies need to be built in advance, yet the time and place of unwanted photos cannot be known in advance. Ilia et al. [14] proposed a system that takes advantage of the existing face recognition functionality of social networks and can interoperate with the current photo-level access control mechanisms. Vishwamitra et al. [15] proposed an approach to facilitate

collaborative control of individual items for photo sharing over social networks, where they shifted focus from complete photo level control to the control of individual items within shared photos. Although letting users set a privacy policy for each photo related to them is an excellent solution to preserve privacy, it is tedious and time-consuming. Several works enable a user to express their privacy deal by a physical tag, such as QR code [16] and stickers or badges [17]. Unfortunately, it is impractical for users to wear such physical tags anywhere and anytime.

Unlike the above schemes, Hasan et al. [4] trained a classifier using computer vision techniques to detect bystanders in photos automatically, achieving state-of-the-art performance. However, the concepts of the bystander and the unaware parties are essentially different. Besides, we empirically tested that the features used in [4] did not perform well and even harmed the task of identifying unaware parties.

3 MOTIVATION AND BACKGROUND

3.1 Motivating Problem

3.1.1 Non-Consensual Photo Sharing on Social Networks. A general photo sharing practice process is shown in Fig. 1, which basically involves four kinds of entities: photographers/photo sharers that take photos based on their interests and submit photos online, individuals that are covered in the shared photos, social platforms that receive the submissions and publish them on the cyberspace, viewers that view and re-share the shared photos through the social platform.

Though showing substantial societal and commercial justification, sharing real-world photos also constitutes privacy disclosure risks, as the snapped individuals may be unaware that they were being photographed, such as photo ② and ③ in Fig. 1. Sharing these photos online would obviously disclose the privacy of the unaware parties.

3.1.2 Definition of Unaware Party. In our study, we define the objective entity whose visuals are depicted by non-consensual photo as an 'unaware party'. From the perspective of photographed individuals, the notation of 'unaware party' refers to "a person who is unaware that they have been captured", such as photo ② and ③ in Fig. 1. The 'aware party' is the opposite concept to the 'unaware party', such as photo ① and ④ in Fig. 1.

3.1.3 Difference Between Unaware Party and Bystander. The current research protocols classify people in shared photos into two categories, subjects (photo ① and ②) and bystanders (photo ③ and ④), based on the importance of a person to a photo and the intention of the photographer [4]. The unaware party defined in our scheme intersects but is not identical to the bystanders. The unaware party embodies the bystander who are unaware that they were captured (photo ③), yet the bystanders who are aware that they were captured are not necessarily unaware parties, even though they were captured unintentionally. Besides, the unaware party contains the part opposite the bystanders, wherein the persons were captured under a particular circumstance, namely, secretly photographed persons. Wikipedia articulates 'secretly photographed person' as "a person who is unaware that they were being intentionally photographed or filmed" in the article 'secret photography'[3]. Undoubtedly, a 'secretly photographed person' is a subject of the photo, such as photo ②; meanwhile, the person is an unaware party. There is no logical or causal connection between the unaware party and bystander, and the concept of unaware party focuses on the unaware state of the person at the time of taking photos, which has nothing to do with whether the person is a bystander or not.

3.1.4 Design Goal. We note that the privacy of unaware parties deserves more attention, as their privacy is more easily violated and more difficult to protect. In this paper, we focus on combating the non-consensual photo sharing issue to protect unaware parties' privacy by a complementary technical approach: automatically detecting unaware parties using computer vision before the photos are posted online.

3.2 Key Challenges and Basic Ideas

This section outlines key challenges and basic ideas to combat the non-consensual photo sharing issue.

How to characterize unawareness and awareness? Based on the definition, the main common attribute of the unaware parties is centered around the concept of *unawareness* when they were being photographed. Thus, the task of automatically detecting unaware parties would be formalized as inferring whether the persons were unaware when they were photographed. However, for parties in the shared photos, terms like unawareness or awareness can have subjective and subtle interpretations that vary across individuals. Intending to enumerate the variety of these perceptions, we first conducted a user study. In the study, participants classify a person as an 'unaware party' or 'aware party' based on social norms, past experience, and visual information available in the image. Nevertheless, unlike the human observers, the detector is constrained to use only the visual information from the photo. Thus, participants were encouraged to highlight all possible visual features associated with unawareness subsequently. The results of the study showed that the participants consistently identify several visual features that distinguish unaware parties from aware parties. Section § 4.2 details the user study.

How to automatically identify the unaware parties? A prerequisite for building an automated classifier is collecting training features and labels. As a result, we first established a data set with manual annotation. As mentioned above, we selected a set of visual features strongly related to *unawareness*. To empirically test the validity of this set of visual features as predictors for automatic classification, we built several classifiers using different single features and new features derived from the multiple features' fusion, respectively. Besides, we also explored the predictive power of other image features in this task. We use the best performing classifier, which turned out to be the designed multi-feature fusion classifier. The construction process of our designed classifier is presented in Section § 5.1, and Section § 6.2 details the performance results.

How about the practical implementation issues? As decisions to share images containing unaware parties are highly subjective and consequential, we expect the automated classifier to be

[3] https://en.wikipedia.org/wiki/Secret_photography

used as part of a human-in-the-loop support system rather than fully automated. Another practical issue is that classifying each individual in each photo is time-consuming. From the sharing behavior in the real world, most people have a strong aspiration to be consulted for decisions before others upload the photos containing them on social networks. To this end, we investigated the sharing preferences of unaware parties and assessed acceptance of the behavior of consulting before sharing in Section 4.3 to establish a baseline for user habits/experience. Based on the baseline, we give the decision-making authority to unaware parties by presenting the necessary information in a user interface and provide users with a non-mandatory method for authorizing photo sharing events in advance. As an auxiliary tool, the authorization method can filter the people in the shared photos who have authorized the sharing, and these people no longer need to be detected by the classifier. We assessed user acceptance and engagement for the design of human-in-the-loop, and the user feedback is introduced in the Section § 6.4.

3.3 Primitives

We briefly introduce primitives here and provide further details in Appendix A.

Certificateless Aggregate Signature. Certificateless aggregate signature (CLAS) can eliminate the complexity of the certificate management and save communication and computation by compressing many individual signatures to a compact signature [18–20]. We adopt the state-of-the-art pairing-free scheme proposed in [20] to support users to authorize photo sharing. This scheme can be briefly introduced as follows: **Setup** is a system initialization algorithm; **PartialKeyGen** and **UserKeyGen** combine to generate public and private keys for users; users can generate a signature on some message by using the algorithm of **Sign**; multiple signatures can be aggregated into one signature by **Aggregate**; the validity of the aggregate signature can be verified using **AggVer**. Appendix A introduces more details about the construction of CLAS [20].

Fuzzy Extractor. A fuzzy extractor is defined by two procedures [21, 22]: Generation procedure *Gen* and Reproduction procedure *Rep*. *Gen* takes a biometric feature B as input and exploits a probabilistic generation function in a permissible error-tolerant manner to generate a unique random string $R \in \{0,1\}^n$ and a helper data string denoted as $HS \in \{0,1\}^*$. *Rep* takes a noisy biometric B' and HS as inputs. If the difference between B' and the original biometric B is less than a threshold value, and $(R, HS) \leftarrow Gen(B)$, then $Rep(B', HS) = R$; otherwise, there is no guarantee about the output.

4 USER-DRIVEN VIDERE DESIGN

Before designing *Videre*, we first assessed the user behavior preferences and explored the connection between visual features and unaware parties. The user data from the studies help us establish a baseline for user habits, which drives the design of *Videre*. This section first introduces the motivations, the processes, and the results of the studies. Based on the knowledge from the studies, we develop the system architecture of *Videre*.

4.1 Ethical Approval of User Studies

We performed three surveys for this research. The first user survey is about predictive features of the unaware party. The second survey is to assess user behaviour. The last survey is about the feedback on human-in-the-loop design. We introduce the first two user studies in this section and the last study in Section 6.4.

Before performing user studies, we submitted our survey design to our Institutional Review Board (IRB) and obtained their approval. In all surveys, all participants' responses were anonymous, and participation is purely voluntary. Materials related to the participants' private information, such as portraits and platform-generated photos, were deleted following the completion of the survey studies. The candidate photos are from public datasets. Overall, our study would not cause privacy and ethics concerns.

4.2 Study 1: Finding Predictive Features

The ultimate goal of *Videre* is to predict whether a person in the image is an unaware party. Nevertheless, since the difference between people who have been photographed unawares or normally is usually subtle and subjective, detecting them using visual features alone is inherently tricky. We approach this challenging problem by conducting a user study to understand the subtle concepts unveiled in a photo that participants use to distinguish between the 'unaware party' vs. 'aware party'. This understanding underpins our work, eventually enabling us to find people who may be unaware parties at scale.

The first half of the study focused on identifying the *discrepancy* of humans in classifying 'unaware party' and 'aware party' in an image. For each specially generated photo, we asked participants to speculate if they were aware or unaware that they were being photographed and, if so, why. To quantify the *consistency* of visual features influencing participants' speculations, we spurred participants to provide all possible visual features associated with the causes of the speculations in the second part of the study.

4.2.1 Survey Design. We first conducted a pilot study as the forerunner of the official user study to help us refine the study design and collect guiding answers to the questions. Early adopters of the study included experts in related fields and people without relevant work experience. In the pilot study, participants' feedback focuses on that, as a mere external viewer of an image, they cannot put themselves in the position of the person in the image to distinguish between the 'unaware party' vs. 'aware party'. To this end, we built an **immersion enhancement** platform in the official user study to encourage the participant to empathize with the person in the image. From a high-level view, the **immersion enhancement** platform generates personalized photos, which drive the remainder of the survey using portraits provided by participants. We next introduce the user study in detail.

Candidate photo set. We first selected photos in different scenes with different numbers of people from the COCO 2017 unlabeled images set [23] without using any predefined list of class names or tags. These photos form a candidate set to provide context for the generation of personalized images. To help the participants answer

Figure 2: Example photos and target person questioned in our user study.

the questions, we drew a rectangular bounding box enclosing one person as the target person for each photo, as shown in Fig. 2.

Part 1: People classification-related questions. With consent, we asked participants for their portraits. Once a participant provided a portrait, the **immersion enhancement** survey platform first selected relevant images from the candidate photo set based on the participant's gender and age. Subsequently, the faces of the participant and target persons in the bounding box were swapped to generate personalized photos using computer vision technology [24] in the platform. We generated 20 personalized photos for each participant and then used the photos to ask participants the following questions.

Q1: If you were the person in the green box, do you think you were aware or unaware that you were being photographed? with answer options ○ Definitely unaware; ○ Most probably unaware; ○ Not sure; ○ Most probably aware; ○ Definitely aware.

Q2: Depending on the response to the previous question, we asked one of the following three questions: **1) What is the main reason for thinking that you were unaware of being photographed? 2) What is the main reason for thinking that you were aware of being photographed? 3) Please describe why do you think it is hard to decide whether you were aware or unaware that you were being photographed?** Each of these questions could be answered by selecting one main option that was provided in Table 1 or input in a text box in case the provided options were not sufficient. We curated these options by extracting emergent themes from the responses of the pilot study, where participants answered this question in an open-ended discussion.

Table 1: Broad reasons used as prompts in our study

Reasons for Unaware Parties
1. I did not notice the capture device.
2. I was not in the natural state to be photographed at the time.
3. My performance is different from other persons in the photo.
Reasons for Aware Parties
1. I noticed the capture device.
2. I was in the natural state to be photographed at the time.
3. My performance was similar to other persons in the photo.

Part 2: Association between human reasoning and features. To further spur participants' thinking, we next informed them of a set of visual features as options. These specific visual features were the research team's initial hypotheses about how unawareness or awareness manifests. We explained the meaning of the terms to ensure participants understood all of them.

Q3: If you were given the opportunity to list out visual features of why you think so, what would all the visual features be? with answer options: ○ The person's gaze direction; ○ The person's head orientation; ○ The person's facial expression; ○ The person's body pose; ○ The person's motion. ○ Other [free text]. Participants answered this question by selecting one or more options that were provided.

Part 3: Demographics and attention check. We collected participants' demographics, including age, gender, and education. We also included a generic attention check question (See Appendix B).

4.2.2 Survey Implementation. We recruited volunteers to participate in the survey from May through July 2021. We balanced participants' age and gender distributions in data collection so that an approximate number of people could classify each photo. Participants were compensated $5 for this study. All participants' responses were anonymous, and all the participants' portraits and the photos generated from portraits were deleted after the survey.

4.2.3 Observations. We next present the observations of user data from study 1.

Participant profile. We collected data from 59 participants, 51 of which were valid after screening. Eight participants' data were excluded since their answers were irrelevant or not useful to the study (e.g., answering "not sure" and giving a reason "do not know"). Most participants (85%) were primarily recruited among postgraduates at age of 20-30. 62.7% of these identified themselves as male and 19 as female.

Why the person is perceived as unaware or aware by humans? For 'unaware party', the most frequently selected reason for labeling a person as an 'unaware party' is 'not notice the capture device' (67.0%). The second most frequent reason is 'not in the natural state to be photographed at the time' (28.5%). For 'aware party', the top two reasons are 'not in the natural state to be photographed at the time' (61.7%) and 'not notice the capture device'(27.5%), respectively. Intuitively, these reasons are related to the visual features from our initial hypotheses.

In addition to these, some other meaningful insights were also provided by participants. P07 said, "*I was attending an event, and there were a lot of cameras around me. Hence, I should have predicted in advance that I was photographed.*"; P26 said, "*Obviously, a white dot on this photo indicates that the photographer turned on the flash, so I was aware I was being photographed.*"; P45 said, "*It is hard for me to keep this action for a certain time. So I was waiting for the photo.*"

All these results indicate that classifying 'unaware party' and 'aware party' by humans is a complex reasoning process involving visual features that can be extracted from the images, the semantic meaning of the images, and rich inferential knowledge not available in images. Since our ultimate goal is to build classifiers that only use the images as input, we made further efforts to investigate the relationships of the human rationale with visual features that can be extracted from the image.

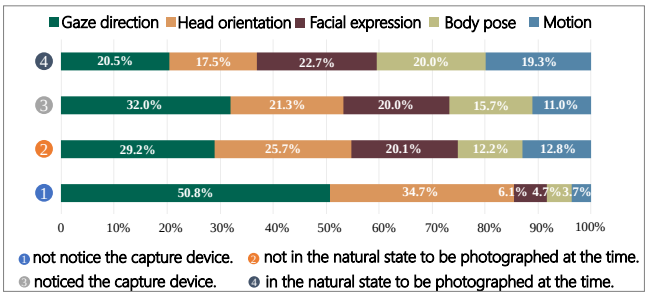


Figure 3: Frequency of different features selected for different reasons.

What factors influence human reasoning? We counted the feature selection related to the top two reasons as shown in Fig. 3. The selection number of these features indicates a correlation between features and the rationales. For 'in the natural state to be photographed', the most frequently selected factor is facial expression (22.7%), followed by gaze direction (20.5%). Gaze direction (29.2%) and head orientation (25.7%) play a decisive role in influencing participants to choose 'not in the natural state to be photographed'. For both 'not notice the capture device' and 'notice the capture device', the gaze direction was selected as the factor affecting them by most people (50.8%/32.0%), followed by head orientation (34.7%/21.3%). This implies that participants tended to agree more with the assertion that the gaze direction and head orientation of a target person causes the person to (not) notice the capture device. Associations among the other reasons and high-level visual features can be similarly interpreted.

4.3 Study 2: Assessing User Behaviors

As we expect our work to be used as a human-in-the-loop support system, we give the decision-making authority to unaware parties by presenting the necessary information in a user interface. Notwithstanding, as different roles in the photos (bystander or subject), unaware parties may require different information to decide if the photo can be shared.

On the other hand, the design of the auxiliary tool was inspired by the real-world observation of privacy-aware people's sharing behavior. They urge to be consulted for decisions before uploading their pictures. At the same time, when sharing a photo, they are also more inclined to ask about the decisions of others in the photo. Nevertheless, the level of acceptance of this sharing behavior in the general population is unclear.

Based on the above considerations, we conducted another user study that focuses on exploring the information needed to make sharing decisions for unaware parties and the acceptance of the behavior of consulting before sharing for the general population.

4.3.1 Survey Design. The user study 2 revolves around the following research questions:

Q1: If you are an unaware party and are the subject of a photo, do you allow the photo to be shared by others on social networks?

Q2: If you are an unaware party and are the bystander of a photo, do you allow the photo to be shared by others on social networks?

The following options were provided for **Q1** and **Q2**: ○ It doesn't matter, I allow it to be shared; ○ I won't allow it to be shared anyway; ○ It depends on the photo content; ○ It depends on the time of taking photos; ○ It depends on the location of the photo; ○ It depends on the identity of the photographer ○ Other [free text]. Participants can choose one or more options or input additional insights.

Capturing user's acceptance of the behavior of consulting before sharing based on participants' rating on a 5-point Likert scale (1: Strongly unwilling to 5: Strongly willing) in response to the following questions:

Q3: If you are currently a photo sharer, would you like to ask people in your photos about their willingness to share before you share them on social networks?

Q4: If someone asks if he/she can share a photo that includes you, are you willing to respond actively?

4.3.2 Survey Implementation. To avoid any possible order effect, participants were divided into two groups. Participants in group 1 were assigned the questions **Q1** and **Q3**, group 2 were assigned the questions **Q2** and **Q4**. We clearly explained the meaning of terms such as 'bystander', 'subject', and 'unaware party' to the participants in conjunction with the pictures presented. Participants were compensated $1 for answering the questions. The average completion time is 4.69 minutes.

4.3.3 Observations. We next present the observations of user data from study 2.

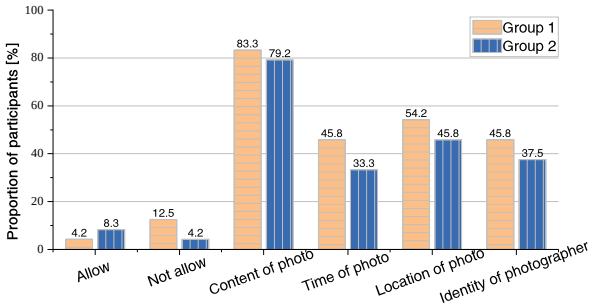


Figure 4: Proportion of participants who selected each option in Q1 and Q2.

Participant profile. Using the online public questionnaire website, we invited 70 volunteers to participate in the survey. Most participants (85%) were primarily recruited among postgraduates and fell in the age range of 20–30 years. 38 (54.3%) of these identified themselves as male and 32 (45.7%) as female. 64.7% indicated that they had been photographed unawares. If they were an unaware party in a photo, 90.2% showed reluctance to others sharing these photos directly.

Information needed for unaware parties to make sharing decision. Fig. 4 shows the proportion of participants who selected each option in **Q1** and **Q2**. From the figure, most participants chosen the option ' It depends on the photo content' with little difference across different groups (83.3%/79.2%). There are more conservatives in group 1, 12.5% of participants do not allow the photo to be shared by others when they were unaware parties and subjects in a photo. In addition to these, P05 of group 1 said, "*It depends on who can view the photo.*"

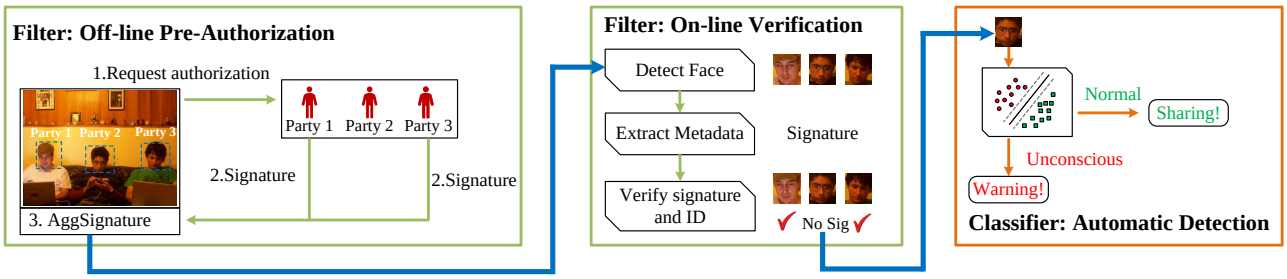


Figure 5: System architecture and workflow of *Videre*.

Acceptance of consulting before sharing. For the photo sharers, 83.3% (Median=4, Mean=4.29, SD=0.84) of participants expressed a willingness or a strong willingness to ask the share opinions of persons in the photo. For the photographed persons, 82.8% (Median=4, Mean=4.17, SD=1.01) of participants were willing or strongly willing to respond to photo shares' inquiries actively. Most of the photo sharers and the captured persons hold a positive outlook on the behavior of consulting before sharing, which motivates the design of the filter.

4.4 System Design

User habits/experiences baseline established from the observations of studies 1 and 2 directly drive the construction of two fundamental building blocks in *Videre*: **multi-feature fusion classifier** (as illustrated in orange box in Fig. 5) for automatic detection of unaware parties and **signature-based filter** (as illustrated in green box in Fig. 5) to filter authorized parties.

As shown in Fig. 5, we now present the architecture and workflow of *Videre*. Before plugging into the details, we first introduce two roles for the users in *Videre*: Photo-sharer: the user who uploads the photos on social networks; Party: the person in the uploaded photos.

► **User Registration:** New users of *Videre* need to register their identity information (such as name and face photos). *Videre* then generates a unique *ID* and key pair for each user.

► **Pre-Authorization:** Due to the high acceptance of the behavior of consulting before sharing, we provide users with a non-mandatory pre-authorization function to relieve the practical issue caused by time-consuming classifying. In this phase, the photo-sharer can request the parties' decisions in the shared photo. If one party allows the photo to be shared, they generate a signature and tag their face region to complete authorization; otherwise, the photo-sharer needs to process the photo, such as obfuscating part of this photo. After receiving some signatures, the photo-sharer aggregates them and appends the aggregated signature to the photo's metadata, which can be EXIF format.

►**Verification:** Once the photo-sharer uploads a photo, *Videre* extracts the metadata to obtain the signature. *Videre* then verifies the validity of the signature and identifies the party who has generated the signature. These parties who pass the verification will default to agree with the photo sharing and no longer need to be detected by the classifier, such as party 1 and party 3 in Fig. 5.

► **Automatic Detection:** Based on the features obtained from study 1, we built a multi-feature fusion classifier to detect unaware parties automatically. Those parties without signatures will be fed into our classifier for detection, such as party 2 in Fig. 5. If there is no attachment signature in the metadata, *Videre* will detect every party in the photo and classify them.

► **Photo Sharing:** If a party is classified as an unaware party, *Videre* will match his face with the registered face to obtain the relative user ID. In view of the results of study 2, *Videre* then sends the unaware parties a warning and the photo to let they decide whether the photo can be shared.

5 BUILDING BLOCKS OF VIDERE

In this section, we formulate the main building blocks of our scheme: **multi-feature fusion classifier** and **signature-based filter**.

5.1 Multi-Feature Fusion Classifier

We now introduce the multi-feature fusion classifier construction process, as shown in Fig. 6. First, to extract corresponding features, each image needs to be cropped to obtain the head region, which will be used as the input of the backbone networks. Second, we introduce two pre-trained learning models as backbone networks to extract different features, respectively. The model used in our design and how we extracted features are presented below.

Gaze direction extraction. We used Gaze360 [25] pre-trained on a large dataset (created in [25]) to estimate the gaze direction of a person. Gaze360 is a gaze-tracking model based on the Long Short-Term Memory (LSTM) [26] and the ResNet [27], which utilize sequences of 7 frames to predict the gaze of the central frame. A head of the target person is first processed by ResNet18 to produce a 256-dimensional feature. Then, the vector is fed to LSTMs and a fully connected layer to generate two outputs: gaze prediction and error quantile estimation. We feed the cropped images of people in our dataset to this model and extract the output of the second-to-last network layer to be used as features for our classifier.

Head orientation extraction. We used image2pose [28] pre-trained on the AFLW2000-3D dataset to estimate head orientation. The image2pose model is a two-stage network based on Faster R-CNN: the first stage is a region proposal network (RPN) with a feature pyramid network (FPN) [29], which is used to locate potential face locations; the second stage uses region of interest (ROI) pooling to extract features and pass them to two different head prediction tasks. We feed the images of people in our dataset to ROI pooling and the fully-connected layer to obtain the last face pose as the head orientation feature that we need.

Third, we concatenate the features extracted from the above backbone networks. The primary benefit of feature-level fusion is

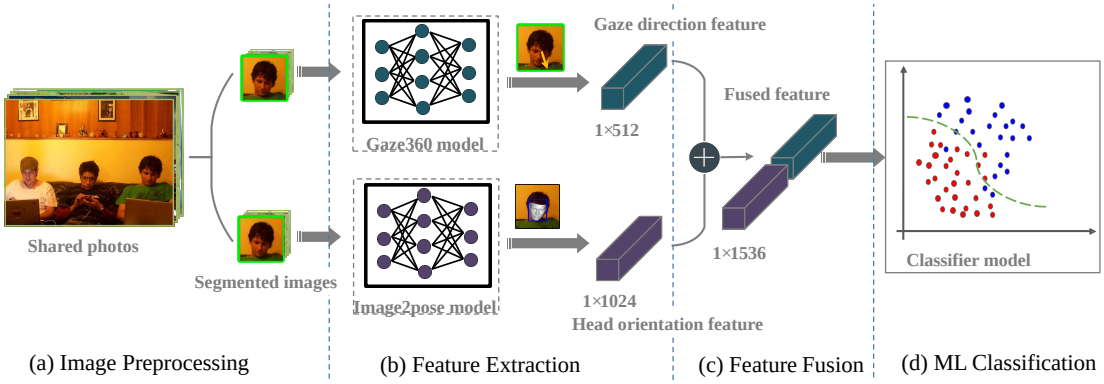


Figure 6: Overview of the classifier construction process. Gaze360 and Image2pose are pre-trained deep learning models. The gaze direction features denoted by cyan cuboid are 512 dimensional, and the head orientation features represented by purple cuboid are 1024 dimensional.

that we can derive a compact set of informative features from the related feature values generated by deep learning models, thereby improving classification accuracy. Finally, we train an MLP classifier using fused features and the established dataset.

5.2 Signature-Based Filter

In this section, we first show some notations used in the description of the filter in Table 2. Then, we sketch a strawman solution and analyze the challenging issue it faces. Finally, we present the detailed process of constructing the filter.

5.2.1 A Strawman Solution. When the photo-sharer wants to obtain the authorization of some parties, they send the photo to them and ask for signatures. If the party i allows the photo to be shared, they prepare the signature σ_i by calling **Sign** and ID_i. After receiving the signatures, the photo-sharer aggregates them by calling **Aggregate** to produce an aggregated signature. Subsequently, the photo-sharer appends the aggregated signature and IDs to the photo's metadata.

Once the photo-sharer uploads a photo, *Videre* extracts the signature from its metadata and verifies the signature by calling **AggVer** algorithm. If the signature is verified to be valid, *Videre* detects each face in the uploaded photo and matches the detected face with the registered faces to obtain related IDs'. Subsequently, *Videre* can match the signature with a specific party in the photo by comparing the obtained IDs' with the IDs in the metadata. Finally, the parties who have authorized photo sharing can be filtered, and no longer need to be detected.

However, the design is not practically viable due to the following defect: in the above solution, to match the signature with a specific party in the photo, we need to detect each face in the photos and match all the detected faces one by one with the registered faces. Nevertheless, it is computational expensive and time-consuming for large-scale subscribers in practice.

5.2.2 Our Design. Targeting the issue, we first require users who allow photo sharing to click on their faces to tag themselves. We then design a method of ID generation based on the fuzzy extractor [21], where a similar face feature can recover the ID due to the error-tolerant of the fuzzy extractor. Hence, we obtain parties' IDs directly rather than match all the detected faces with the registered faces one by one. Precisely, the method of ID generation and reconstruction consists of the following two algorithms:

Table 2: Notation and Description

Notation	Description
ID_i	the ID of user i
IDs	a set of ID
$PHF_k(\cdot)$	a k-bit perceptual hash function
RF_i	a face photo of user i
BS	a k-bit binary string
Img	an uploaded image
DF	a detected face in Img
P	a party in Img
p	the detected face of P
msk	the master secret key
S	a random string in Z_q^*
H_4	a hash function: $Z_q^* \rightarrow \{0,1\}^k$
H_5	a hash function: character string $\rightarrow \{0,1\}^k$
$d_H(\cdot,\cdot)$	the hamming distance metric method
δ_H	a fixed similarity threshold
Ω_p	the detected facial region of P

• **IDGen** (RF_i, msk, PP). This algorithm is run by *Videre* to generate ID_i for user i at the stage of user registration. After receiving the face photo RF_i submitted by the user i, *Videre* computes the perceptual hash value of RF_i as the face feature, namely, $BS_i \leftarrow PHF_k(RF_i)$, and then generates the random string and the helper string $(R_i, HS_i) \leftarrow Gen(BS_i)$. Finally, *Videre* computes $ID_i = H_4(msk + S) \oplus H_5(R_i)$ as the ID of user i.

• **IDRec** (msk, DF_i, HS_i). This algorithm is run by *Videre* to reconstruct ID_i for user i at the stage of verification. *Videre* first computes the perceptual hash value of DF_i by calling $BS_i' \leftarrow PHF_k(DF_i)$, where DF_i is a detected face of a user i in the Img. *Videre* then can reproduce the random string $R_i \leftarrow Rep(HS_i, BS_i')$ and reconstruct $ID_i' = H_4(x + S) \oplus H_5(R_i)$. If DF_i is similar with RF_i, $d_H(BS_i, BS_i') < \delta_H$. Due to the error tolerant of fuzzy extractor, we can obtain $ID_i = ID_i'$.

Now, the entire implementation process of the filter can be described as follows.

►**Setup.** *Videre* calls **IDGen** algorithm to generate ID for each registered user with their face photo. Then, the registered users generate their respective keys with **PartialKeyGen** and **UserKeyGen** algorithms using their own ID.

►**Pre-Authorization.** When the photo-sharer wants to obtain authorization of the user i, they send the photo Img to the user i and ask for a signature. The users who allow this sharing event are required to mark the position of their faces in the photo, which needs pixel coordinates of any point in their facial area. As shown in Fig. 7, if the rightmost user in the photo wants to empower the photo sharing, he just needs to click on his facial region (the area in the blue box), for example, the green dot. The pixel coordinate values of the green dot then would be returned. After that, he needs to click the "Allow" button to complete the authorization. Triggered by the click operation, *Videre* first computes the perceived hash value of Img, namely, $BS_Img = PHF_k(Img)$, as the signing messages, and then generates $\sigma_i = (T_i, \tau_i)$ as the signature of user i for Img by calling **Sign**.

Finally, $(HS_i, ID_i, pk_i, \langle x_i, y_i \rangle, \sigma_i)$ would be appended to the photo's metadata, where $\langle x_i, y_i \rangle$ is the pixel coordinates of the click location, HS_i is the helper string generated by **IDGen**. After receiving the signatures of all target users or the time to wait for responses exceeds the maximum time, *Videre* aggregates the received signatures by calling **Aggregate** to produce an aggregated signature Agg_σ. We investigated the maximum time users can tolerate in a privacy-enhanced photo sharing process through a survey, which is presented in Section § 6.4. *Videre* then appends Agg_σ to the photo's metadata. The **Aggregate** algorithm can reduce the communication cost and the computation cost of verifying signatures. For the case where the parties do not want to share the photo, they need to click the "Reject" button.

►**Verification.** Once the photo-sharer uploads a photo, *Videre* extracts the photo's metadata and computes its perceptual hash value. *Videre* then can verify the signature by calling **AggVer** algorithm. If the signature is verified, *Videre* detects faces in the uploaded photo and then obtains the facial region occupied by each party. As illustrated in Fig. 7, the bounding boxes of the facial region are generated by the online face recognition API of Tencent[4]. *Videre* then derives $(ID_i, \langle x_i, y_i \rangle)$ from the metadata. If $\langle x_i, y_i \rangle \in \Omega_p$, party P is marked by a user whose identification is ID_i at the **Pre-Authorization** stage. After that, *Videre* derives HS_i and generates ID_P by calling **IDRec** algorithm with the detected face p, the helper string HS_i, and the secret key msk. If ID_P is equal to ID_i obtained from the metadata, party P (user i) has authorized the photo-sharer to share this photo. Hence, the party P will default to be photographed normally and no longer needs to be classified by the classifier; otherwise, the party P will be sent to the classifier for further detection. Compared with the strawman solution, we avert the time-consuming face matching process by just computing the ID of parties tagged by users.

Suppose the user i rejects the sharing, the photo-sharer is obliged to take the initiative to blur the user i in the photo. However, the photo-sharer may upload the photo directly without any processing. While the user i would be detected by the classifier in our design, the photo itself may have been taken normally. The photo would then be shared on the internet without warning to the user i. As a result, the privacy of the user i is violated. This privacy leak model is a well-studied issue called privacy conflict [30, 31] in the privacy-enhanced photo sharing domain. We note that further orthogonal efforts could be integrated for a more robust system, while the privacy conflict issue is out of the scope of our focused context.

[4]https://cloud.tencent.com/product/facerecognition

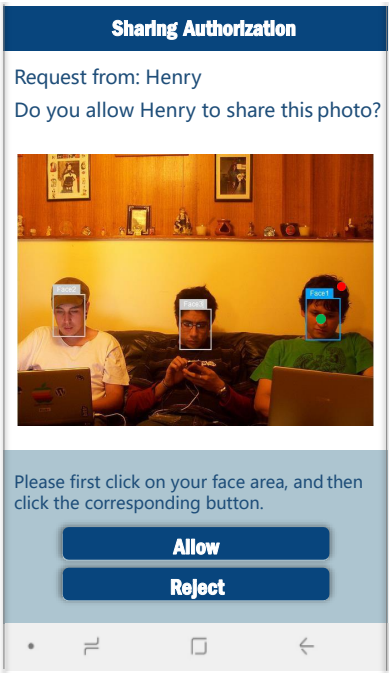


Figure 7: Illustration of interactive interface for signature generation. The green dot and red dot denote correct and wrong click point, respectively. Rectangular boxes are detected face regions by face recognition algorithm.

6 EVALUATION

6.1 Dataset and Experimental Setup

Training Data Collection and Manual Annotation: A prerequisite for developing an automated classifier is collecting training features and labels. To combat this problem, we selected 5013 images from the Flickr30k [32] and the COCO2017 [23] datasets. To get the state of each person in the images, we further segmented each person in the images, resulting in a dataset of 6437 persons. If there are N persons in an image, we made N copies of it, and each copy was pre-processed to draw a rectangular bounding box enclosing one person. We next presented these copies to at least three participants (they were recruited on our campus, and each worker was paid $50) and asked them **Q1** of user study 1. The class label of a person was determined using the mean score for question **Q1**: a positive score was labeled as 'aware party', a negative score was labeled as 'unaware party', and zero was labeled as 'neither'. In this way, we acquired 3764 (58.5%) persons with the label "unaware party" and 2673 (41.5%) persons with "aware party". In the

following experiments, 6437 persons have split into train and test sets of 4593 (71%) and 1844 (29%) persons, respectively.

Experimental Setup. We implemented our classifier in Python using some pre-trained models and open-source libraries, including expression recognition model [33], gaze prediction model [25] and popular face detection algorithm [34]. Each experimental result was obtained on a Desktop PC equipped with an Intel i9-10900K CPU with 64 GB RAM and an NVIDIA 3070 GPU running Linux. The implementation of the filter is in Java, based on the Java pairing-based cryptography [35].

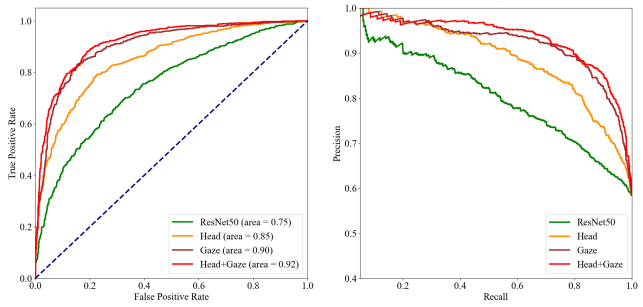
6.2 Performance of the Classifier

To perform classification for this task, we compared several established supervised learning algorithms: Support Vector Machines (SVM), Logistic Regression (LR), Multilayer Perceptron (MLP), Deep Neural Networks (DNN). The best performing classifier is MLP, followed by SVM. We compared our classifier to multiple baselines. The first was a classifier trained directly using the cropped images as features, representing a model trained with the most concrete set of features, i.e., the raw pixel values of the cropped images. The second was a classifier trained with deep features with more information, which fed the features extracted from the cropped images using a deep learning model (ResNet50 [27]) into the machine learning classification model. The next classifiers were trained with higher-level features (body pose and facial expression) and their fused feature of them. We report their results in Table 3.

Table 3 shows the performance (Accuracy, F1-score, Precision, Recall) of our trained classifiers. Accuracy denotes the number of true positives and true negatives as a percentage of all samples; Precision is the number of true positives out of the combined number of true positives and false positives; Recall is defined as the number of true positives out of the combined number of true positives and false negatives; F1-score is the harmonic mean of the precision and recall. The table shows that the classifiers trained with deep features perform better than those trained with raw images. These results suggest that the deep features extracted by the deep model contain rich deep abstraction information. However, not all deep features exhibit remarkable predictive power, and the features irrelevant to the task can even have adverse effects. Body pose and facial expression used in [4] to detect bystanders are not satisfactory in this task. For example, the accuracy of the MLP model trained with facial expression feature, a kind of higher-level semantic feature, is lower than the result of random prediction (50%). In contrast, the features selected from our user study have powerful representational abilities for the subjective privacy task. Our design classifier model achieves the best performance with an accuracy of 85.1% and an F1-measure of 0.849.

Receiver Operating Characteristic (ROC) curves are standard tools to interpret the problem of probabilistic prediction for binary classification models. The Area Under Curve (AUC) is used to present the correctness of a model. The closer the AUC value is to 1, the more accurate the model is, and the closer it is to 0.5, the more the model tends toward a random classifier.

Fig. 8 shows ROC-AUCs of the MLP classifier models trained on different features. We can observe that the MLP classifier trained on fused features has the best performance, which has an AUC of 0.92.



(a) ROC curves. (b) PR curves.

Figure 8: Receiver operating characteristic (Receiver operating characteristic) and Precision vs. recall (PR) curves of MLP classifiers trained on different features.

Interestingly, we also find that the gap between the fused feature-trained classifier and the gaze direction feature-trained classifier, which has an AUC of 0.9, is minimal. This result may be caused by the fact that the Gaze360 model [25], for achieving robust gaze direction prediction, also takes into account the head orientation.

Besides high precision, we also aim for high recall, but there is typically a trade-off between them. A precision-recall curve (PRC) allows us to see the trade-off between precision and recall when different possible cutoffs for positive classifications are used. In the PR space, the goal is to be in the upper-right-hand corner. Fig. 8 shows PRCs of the MLP classifier models trained on different features. We can observe that the MLP classifier trained on fused features also has the optimal curve. In addition, we also find that the room between the PRC of the gaze direction feature-trained model and the PRC of the fused feature-trained model gets bigger, while they are pretty close in ROC space.

In summary, the classifier trained with the gaze feature has the strongest predictive power among the classifiers trained with a single feature, which is in line with the results of our user study. Besides, by further fusion of features with high-level concepts, we trained the optimal multi-feature fusion classifier, which achieves an accuracy of 85.1%, an F1-measure of 0.849, and a recall of 0.872 for unaware class.

6.3 Performance of the Filter

In this section, we evaluate the performance of the proposed signature-based filter. Before reporting results, we first introduce the implementation tools and parameters setting. To detect a face, we implemented a popular open-source face detection algorithm [34] based on OpenCV. We utilize the fuzzy extractor in [22] to generate the random string R and the helper string HS. In addition, we need to select a robust perceptual hash algorithm $PHF_k(\cdot)$ and set the value of k and δ_H. Anunay et al. [36] evaluate commonly used PHFs for predictive performance under different transforms on images. The transforms reflect possible user actions to make images visually appealing (gamma correction), highlight image components (cropping, rescaling, and rotation), and share images (noise from compression). Anunay et al. [36] show that the PDQHash would produce the fewest false positives. Thus, we use the PDQHash [37] to deterministically map face photos to a space where proximity

Table 3: Comparison of accuracy, precision, recall, and F1-score metrics for trained classifier models. We define $\mathcal{R}$ as the feature of the raw image, $\mathcal{D}$ as the feature extracted from ResNet50, $\mathcal{B}$ as the feature of body pose, $\mathcal{F}$ as the feature of facial expression, $\mathcal{E}$ as the feature of eye gaze direction, $\mathcal{H}$ as the feature of the head orientation, "+" as the concatenating of two features. $\mathcal{B}$ and $\mathcal{F}$ are used in [4] for bystanders detection.

Features	Classifiers	Overall				Aware			Unaware		
		Acc%	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
$\mathcal{R}$	SVM	63.1	0.63	0.629	0.632	0.611	0.633	0.601	0.648	0.672	0.654
$\mathcal{D}$	SVM	73.2	0.73	0.732	0.729	0.704	0.659	0.619	0.749	0.813	0.779
$\mathcal{B}$	SVM	69.1	0.697	0.691	0.692	0.617	0.683	0.648	0.754	0.696	0.724
$\mathcal{F}$	SVM	72.1	0.801	0.782	0.750	0.630	0.876	0.766	0.872	0.589	0.733
$\mathcal{E}$	SVM	81.4	0.816	0.815	0.815	0.765	0.80	0.782	0.852	0.824	0.838
$\mathcal{H}$	SVM	72.2	0.721	0.722	0.721	0.678	0.638	0.657	0.751	0.783	0.767
$\mathcal{B}$+$\mathcal{F}$	SVM	67.4	0.677	0.674	0.675	0.604	0.638	0.621	0.729	0.700	0.714
$\mathcal{E}$+$\mathcal{H}$	SVM	**81.9**	**0.822**	**0.819**	**0.820**	**0.765**	**0.817**	**0.79**	**0.862**	**0.820**	**0.841**
$\mathcal{R}$	MLP	64.3	0.640	0.644	0.643	0.621	0.599	0.605	0.677	0.663	0.702
$\mathcal{D}$	MLP	69.2	0.690	0.692	0.691	0.638	0.603	0.620	0.726	0.756	0.741
$\mathcal{B}$	MLP	66.2	0.658	0.662	0.659	0.606	0.544	0.573	0.696	0.747	0.720
$\mathcal{F}$	MLP	41.8	0.209	0.500	0.295	0.418	1.000	0.589	0.000	0.000	0.000
$\mathcal{E}$	MLP	83.5	0.834	0.835	0.834	0.818	0.777	0.797	0.845	0.876	0.861
$\mathcal{H}$	MLP	78.0	0.781	0.78	0.78	0.729	0.753	0.741	0.819	0.813	0.811
$\mathcal{B}$+$\mathcal{F}$	MLP	45.8	0.612	0.525	0.368	0.430	0.972	0.597	0.795	0.077	0.14
$\mathcal{E}$+$\mathcal{H}$	MLP	**85.1**	**0.849**	**0.85**	**0.849**	**0.829**	**0.806**	**0.818**	**0.874**	**0.869**	**0.872**

reflects perceptual similarity. We set $k = 256$ and $\delta_H = 25$ throughout our experiments, consistent with the state of the art in [36]. Accordingly, we set the size of the fuzzy extractor's input to 256-bit and the tolerable error bits to 25-bit.

Computational Overhead. The computation cost is composed of off-line pre-authorization and online verification in the signature-based filter. Table 4 presents all the decomposed computing time. In this benchmark experiment, we specify that the number of aggregated and verified signatures is n. For a party, the time spent in generating a signature is minimal, only 9.3 ms; for a photo-sharer, the time consumed by the **Aggregate** is almost negligible, only $0.003n$ ms; as image processing is required, the computation delay is major caused by **IDGen** and **IDRec**. The time spent on them is 0.96 s and 0.98 s, respectively. Besides, *Videre* needs to spend $2.69n$ ms on **AggVer**. It immediately comes to the conclusion that the computational overhead of the filter is small.

Table 4: Run time of each stage of signature-baed filter.

Off-line			On-line	
IDGen	Sign	Aggregate	AggVer	IDRec
0.96 s	9.3 ms	$0.003n$ ms	$2.69n$ ms	0.98 s

Communication Overhead. In the signature-based filter, the communication cost consists of transfer of the image itself and the signature. Since the size of the image is not fixed, we focus on analyzing the size of the signature uploaded to *Videre* along with the photos. The size of a single signature in [20] is $|G|+\left|Z_q^*\right|$ ($|G|$ denotes the size of an element in G, $\left|Z_q^*\right|$ denotes the size of an element in Z_q^*). Suppose n is the number of people who generate signatures, the size of the aggregated signature is $n\,|G| + \left|Z_q^*\right|$. In addition, in order to match the face with the signature, we need to append the user's ID, public key pk, face coordinates $\langle x_i, y_i\rangle$, and the helper string HS to the metadata of the photo. Thus, the total size of the additional information is $n\,(|ID| + |2G| + |x| + |y| + |HS|)+\left|Z_q^*\right|$. As we use the type A curve [35] to implement the signature algorithm, $|G|$ =128-byte, $\left|Z_q^*\right|$ =20-byte, $|ID|$ =32-byte, $|HS|$ =64-byte, and $|x_i| + |y_i|$=8-byte. Hence, the total size of the appended information is (n360+20)-byte. We can conclude that the size of the additional information is acceptable, and the communication overhead between entities depends on the size of the photo.

6.4 Study 3: Feedback on Human-in-the-Loop

Despite our best efforts at automation, there is always be a need for a "human-in-the-loop" when it comes to the highly subjective and consequential privacy-enhanced photo-sharing process. Due to the unquantifiable time delay caused by "human-in-the-loop", *Videre* will face the following obstacles in practical deployment.

O1: Waiting for the users to return the signature approvals or the unaware parties to return to the sharing decision introduces a time delay.

O2: Sending warnings at irregular intervals to users by *Videre* may disturb users.

To explore the user acceptance on the human part of *Videre* and identify the aspects for future improvement, we collected feedback and suggestions through a semi-structured interview. We received 24 answers from a pool, including ordinary social platform network users and privacy-related workers. Participants were instructed about how our system works and how to use *Videre*, and informed that all the notifications are simulated and harmless. It revolves around the following research questions.

6.4.1 Survey Design. Capturing user's acceptance of **O1** and **O2** was based on participants' rating on a 5-point Likert scale (1: Strongly unwilling to 5: Strongly willing) in response to the following questions:

Q1: When sharing photos to social networks, are you willing to wait for some time to obtain the authorization of the people in the photos?

Q2: When sharing photos to social networks, are you willing to accept occasional inquiries from the system?

In addition, we asked **Q3** to investigate user's interest in using *Videre*:

Q3: Are you interested in using this service in your daily lives? with a 5-point Likert scale ranging from Strongly interested to Strongly not interested.

6.4.2 Observations. Fig. 9 shows the participants' responses to the questions **Q1**, **Q2** and **Q3**. The results indicate that most participants are receptive to the above obstacles and interested in using *Videre*. Highlights of our results are the following:

• 53.6% (Median=4, Mean=3.64, SD=1.03) of participants are strongly willing or willing to accept the time delay to obtain the authorization when sharing photos. Furthermore, we investigated the maximum time delay that users can accept in an open-ended discussion, where the minimum acceptable time delay is 30 seconds and the maximum time delay is 48 hours. The acceptable time delay for most participants (45.8%) is in the range of 5 to 30 minutes.

• 60.7%(Median=4, Mean=3.68, SD=1.19) of participants are strongly willing or willing to accept occasional inquiries.

• 79.1% of participants are strongly interested or interested in using our service (Median=4, Mean=3.96, SD=1.06). A Ph.D. from an anonymous university contacted us via WeChat. He asked us for more information and said, "*This work is interesting and meaningful.*"

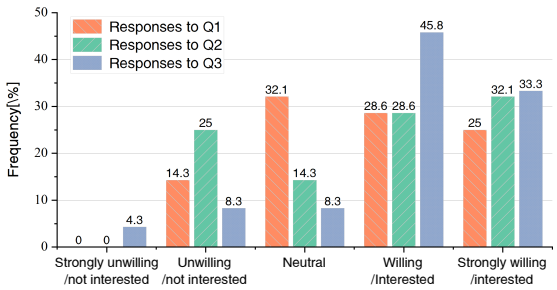


Figure 9: Participants' responses to Q1, Q2 and Q3.

6.4.3 Potential Improvements. Some interviewees suggested interesting ideas for further improving *Videre* functionalities. P19 provided a suggestion that refers to integrating some simple strategies to reduce time delay, "*Users can set a priority option in advance, such as 'not allowed to be shared', which will be implemented by default when the user does not process the photo within the specified time.*" Another suggestion consists of a more complex sharing scenario, P17 said, "*When multiple people share the same photo at the same time, only one warning should be sent to the unaware party.*"

7 LIMITATIONS

Non-consensual photo sharing is still an open challenge, for which it is unlikely to have a perfect solution in the near future. Our work naturally suffers from the following limitations that requires further investigation.

First, the constructions that we propose might result in the risk of privacy disclosure. In the **Pre-Authorization** part, the photo sharers possibly send the photo to more than one party and ask for the signatures. To make a reasonable decision, image content should be learned by all parties in the photo. However, if one of the parties refuses to share the photo, other parties have already gotten the whole photo, which causes privacy concerns.

Second, the face matching algorithm is highly related to the security of our system. If the matching is wrong, the warning would be sent to the irrelevant person, which could lead to the privacy disclosure. Given that the very intention of this proposal is to detect the persons who were unawares photographed, we state that further orthogonal efforts could be integrated for a more robust system, while the powerful face matching algorithm is out of the scope of our focused context.

Third, while we report the performance of the classifier and the filter in Section § 6, respectively, the performance of the prototype system is not evaluated. Future work should focus on collecting data from the real world as ground truth and designing a user-friendly interface. Overall, these efforts would minimize our current limitations and operationalize our work's results to evaluate the prototype system's performance.

Fourth, all of our survey participants come from the same country (although the images used had no such restriction). As a result, our formative understanding of unaware parties is likely to be situated in a particular culture and demographic.

8 CONCLUSION

Aiming to combat the non-consensual photo sharing issue, we propose an automatic detection protocol, called *Videre*, which is policy-free and does not need the victims of privacy violations to be proactive. We explore the prospect of a novel approach – identifying unaware parties solely based on the visual features of an image. In order to achieve this goal, we created a dataset and conducted a user survey in the preliminary preparations. The human-centered understanding improves the performance of automated detection. The results reported in Section § 6 suggest that the predictive ability of informative deep features and task-independent semantic features is unsatisfactory. In contrast, the features selected by human beings based on their understanding show powerful representational abilities for the subjective privacy task. According to the survey results and the dataset, we train a classifier by fusing the key features, which is the first machine learning model for the unaware party detection task with an accuracy rate of 85.1%. In addition, we design an auxiliary tool, namely, a signature-based filter, to reduce the number of automatic detections and speed up the system. Finally, performance evaluation indicates that our filter and classifier are both efficient. Since our system can automatically detect the unaware parties solely based on image data, we believe that it has the potential to protect the privacy of the unaware parties at scale.

REFERENCES

[1] Janko Rottgers. Snapchat loses 3 million daily users, but beats expectations in q2 earnings. https://variety.com/2018/digital/news/snap-posts-q2-revenue-beat-but-daily-users-down-for-first-time, 2018.

[2] 99Content. Whatsapp statistics. https://99firms.com/blog/whatsapp-statistics/#gref, 2019.

[3] Yasmeen Rashidi, Tousif Ahmed, Felicia Patel, Emily Fath, Apu Kapadia, Christena Nippert-Eng, and Norman Makoto Su. "You don't want to be the next meme": College students' workarounds to manage privacy in the era of pervasive photography. In *Proc. of Symposium on Usable Privacy and Security (SOUPS)*, pages 143–157, Baltimore,MD,USA, 2018. USENIX Association.

[4] Rakibul Hasan, David Crandall, Mario Fritz, and Apu Kapadia. Automatically detecting bystanders in photos to reduce privacy risks. In *Proc. of the IEEE Symposium on Security and Privacy (S&P)*, pages 318–335, San Francisco,CA,USA, 2020. IEEE.

[5] Majid Yar and Jacqueline Drew. Image-based abuse, non-consensual pornography, revenge porn: A study of criminalization and crime prevention in australia and england & wales. *International Journal of Cyber Criminology*, 13(2):578–594, 2019.

[6] Kate Brimsted. The jk rowling photo case–are privacy rights evolving for the online era? *Computer Law & Security Review*, 24(5):465–468, 2008.

[7] The High Court of Justice Queens Bench Division. Judgment files. http://www.courtservice.gov.uk/judgmentsfiles/j1096/Naomi_Campbell_v_Mirror.htm, 2002.

[8] Lan Zhang, Kebin Liu, Xiang-Yang Li, Cihang Liu, Xuan Ding, and Yunhao Liu. Privacy-friendly photo capturing and sharing system. In *Proc. of the International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, pages 524–534, New York,NY,USA, 2016. ACM.

[9] Paarijaat Aditya, Rijurekha Sen, Peter Druschel, Seong Joon Oh, Rodrigo Benenson, Mario Fritz, Bernt Schiele, Bobby Bhattacharjee, and Tong Tong Wu. I-pic: A platform for privacy-compliant image capture. In *Proc. of the Annual International Conference on Mobile Systems, Applications, and Services (MobiSys)*, pages 235–248, Singapore,Singapore, 2016. ACM.

[10] Moo-Ryong Ra, Seungjoon Lee, Emiliano Miluzzo, and Eric Zavesky. Do not capture: Automated obscurity for pervasive imaging. *IEEE Internet Computing*, 21(3):82–87, 2017.

[11] Ang Li, Qinghua Li, and Wei Gao. Privacycamera: Cooperative privacy-aware photographing with mobile phones. In *Proc. of the IEEE International Conference on Sensing, Communication, and Networking (SECON)*, pages 1–9, London,UK, 2016. IEEE.

[12] Fenghua Li, Zhe Sun, Ang Li, Ben Niu, Hui Li, and Guohong Cao. Hideme: Privacy-preserving photo sharing on social networks. In *Proc. of the International Conference on Computer Communications (INFOCOM)*, pages 154–162, Paris,France, 2019. IEEE.

[13] Benjamin Henne, Christian Szongott, and Matthew Smith. Snapme if you can: Privacy threats of other peoples' geo-tagged media and what we can do about it. In *Proc. of the ACM Conference on Security and Privacy in Wireless and Mobile Networks (WISEC)*, pages 95–106, Budapest,Hungary, 2013. ACM.

[14] Panagiotis Ilia, Iasonas Polakis, Elias Athanasopoulos, Federico Maggi, and Sotiris Ioannidis. Face/off: Preventing privacy leakage from photos in social networks. In *Proc. of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 781–792, Denver Colorado,USA, 2015. ACM.

[15] Nishant Vishwamitra, Yifang Li, Kevin Wang, Hongxin Hu, Kelly Caine, and Gail-Joon Ahn. Towards pii-based multiparty access control for photo sharing in online social networks. In *Proc. of the ACM on Symposium on Access Control Models and Technologies*, pages 155–166, New York,NY,USA, 2017. ACM.

[16] Cheng Bo, Guobin Shen, Jie Liu, Xiang-Yang Li, YongGuang Zhang, and Feng Zhao. Privacy. tag: Privacy concern expressed and respected. In *Proc. of the ACM Conference on Embedded Network Sensor Systems*, pages 163–176, Memphis,TN,USA, 2014. ACM.

[17] Frank Pallas, Max-Robert Ulbricht, Lorena Jaume-Palasí, and Ulrike Höppner. Offlinetags: A novel privacy approach to online photo sharing. In *Proc. of the Extended Abstracts on Human Factors in Computing Systems (CHI)*, pages 2179–2184, Toronto Ontario,Canada, 2014. ACM.

[18] Yong Ding, Bingyao Wang, Yujue Wang, Kun Zhang, and Huiyong Wang. Secure metering data aggregation with batch verification in industrial smart grid. *IEEE Transactions on Industrial Informatics*, 16(10):6607–6616, 2020.

[19] Attila Altay Yavuz, Anand Mudgerikar, Ankush Singla, Ioannis Papapanagiotou, and Elisa Bertino. Real-time digital signatures for time-critical networks. *IEEE Transactions on Information Forensics and Security*, 12(11):2627–2639, 2017.

[20] Wenjie Yang, Shangpeng Wang, and Yi Mu. An enhanced certificateless aggregate signature without pairings for e-healthcare system. *IEEE Internet of Things Journal*, 8(6):5000–5008, 2020.

[21] Yevgeniy Dodis, Leonid Reyzin, and Adam Smith. Fuzzy extractors: How to generate strong keys from biometrics and other noisy data. In *Proc. of International Conference on The Theory and Applications of Cryptographic Techniques*, pages 523–540, Interlaken,Switzerland, 2004. Springer.

[22] Ran Canetti, Benjamin Fuller, Omer Paneth, Leonid Reyzin, and Adam Smith. Reusable fuzzy extractors for low-entropy distributions. In *Proc. of the Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 117–146, Vienna,Austria, 2016. Springer.

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proc. of European Conference on Computer Vision (ECCV)*, pages 740–755, Zurich,Switzerland, 2014. Springer.

[24] Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. Fast face-swap using convolutional neural networks. In *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, pages 3677–3685, Honolulu,HI,USA, 2017. IEEE.

[25] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6912–6921, Long Beach,CA,USA, 2019. IEEE.

[26] Alex Graves, Santiago Fernández, and Jürgen Schmidhuber. Bidirectional lstm networks for improved phoneme classification and recognition. In *Proc. of the International Conference on Artificial Neural Networks*, pages 799–804, Warsaw,Poland, 2005. Springer.

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas,NV,USA, 2016. IEEE.

[28] Vitor Albiero, Xingyu Chen, Xi Yin, Guan Pang, and Tal Hassner. img2pose: Face alignment and detection via 6dof, face pose estimation. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7617–7627, Nashville,TN,USA, 2021. IEEE.

[29] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2117–2125, Honolulu,HI,USA, 2017. IEEE.

[30] Kemi Ding and Junshan Zhang. Multi-party privacy conflict management in online social networks: a network game perspective. *IEEE/ACM Transactions on Networking*, 28(6):2685–2698, 2020.

[31] Jose M Such, Joel Porter, Sören Preibusch, and Adam Joinson. Photo privacy conflicts in social media: A large-scale empirical study. In *Proc. of the ACM Conference on Human Factors in Computing Systems (CHI)*, pages 3821–3832, New York,NY,USA, 2017. ACM.

[32] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[33] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2852–2861, Honolulu,HI,USA, 2017. IEEE.

[34] Paul Viola and Michael J Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[35] Angelo De Caro and Vincenzo Iovino. jpbc: Java pairing based cryptography. In *Proc. of the IEEE Symposium on Computers and Communications (ISCC)*, pages 850–855, Kerkyra,Greece, 2011. IEEE.

[36] Anunay Kulshrestha and Jonathan Mayer. Identifying harmful media in end-to-end encrypted communication: Efficient private membership computation. In *Proc. of the USENIX Security Symposium (USENIX Security)*, pages 893–910, Virtual, 2021. USENIX Association.

[37] Facebook. The tmk+pdqf video-hashing algorithm and the pdq image-hashing algorithm. https://github.com/facebook/ThreatExchange/blob/master/hashing/hashing.pdf, 2019.

A INSTANTIATION OF THE CLAS SCHEME

Since we adopt the CLAS scheme [20], we summarize its construction to make this paper self-contained. The process is as follows.

(1) Setup $\left(1^{\lambda}\right)$. Let G be a q-order group with a generator Q. *Videre* randomly selects two elements $a, S \in Z_q^*$ and computes $P_{pub} = aQ$. Then, *Videre* sets three hash functions $H_1, H_2, H_3 : \{0,1\}^* \rightarrow Z_q^*$. Finally, it keeps the master secret key $msk = a$ secretly and makes the system parameters $PP = \left\{P_{pub}, G, q, Q, H_1, H_2, H_3\right\}$ public.

(2) PartialKeyGen (PP, msk, ID, B). *Videre* selects a random $r_i \in Z_q^*$ and computes $Y_i = r_i Q$. Suppose the identity of user i is ID_i,

Videre then computes $h_i = H_1\left(ID_i, Y_i, P_{pub}\right)$ and $b_i = r_i + h_i a$. Finally, *Videre* returns $(b_i, Y_i,)$ to user i.

(3) UserKeyGen (PP, msk, ID). User i randomly selects $a_i \in Z_q^*$ and computes $X_i = a_i Q$. Subsequently, the user i generates his private key $sk_i = (a_i, b_i)$ and public key $pk_i = (a_i, Y_i)$.

(4) Sign (M_i, sk_i, PP). User i randomly selects $t_i \in Z_q^*$ and computes $T_i = t_i Q$. User i then computes the following hash values: $s_i = H_2\left(T_i, ID_i, pk_i, P_{pub}\right)$, $k_i = H_3\left(M_i, T_i, ID_i, pk_i, P_{pub}\right)$. Subsequently, the user i computes $\tau_i = t_i + k_i\left(s_i a_i + b_i\right)$, and generates $\sigma_i = (T_i, \tau_i)$ as the signature of M_i.

(5) Aggregate $\left(PP, \{M_i, ID_i, pk_i, \sigma_i\}_{i=1}^n\right)$. Any user can run the aggregation algorithm with the public parameters PP, a set of four-tuples $(M_i, ID_i, pk_i, \sigma_i)$ as inputs. It computes $\tau = \sum_{n=1}^n \tau_i$ and generates $\sigma = (\tau, T_1, T_2, ..., T_n)$ as the aggregate signature on these four-tuples $\{M_i, ID_i, pk_i, \sigma_i\}_{i=1}^n$.

(6) AggVer $\left(PP, \{M_i, ID_i, pk_i\}_{i=1}^n, \sigma\right)$. It takes as input these three-tuples $\{M_i, ID_i, pk_i\}_{i=1}^n$ and σ. The verifier first computes the following hash values: $h_i = H_1\left(ID_i, Y_i, P_{pub}\right)$, $s_i = H_2\left(T_i, ID_i, pk_i, P_{pub}\right)$ and $k_i = H_3\left(M_i, T_i, ID_i, pk_i, P_{pub}\right)$. It then checks whether the equation Equation (1) holds.

$$\tau Q = \sum_{n=1}^{n} T_i + k_i\left(s_i X_i + Y_i + h_i P_{pub}\right) \tag{1}$$

If yes, it outputs "Accept"; otherwise, it outputs "Reject".

B ATTENTION-CHECK QUESTION

Figure 10: Image used for attention check question.

We used Fig. 10 to check participants' attention: **Which of the following statement is true for the photo?** ○ There is a child in the photo. ○ The man in the photo is eating an apple. ○ This photo was taken indoors. ○ The man in the photo is eating an orange. ○ Prefer not to answer.