Kaleido: You Can Watch It But Cannot Record It

Lan Zhang[1], Cheng Bo[2], Jiahui Hou[3,4], Xiang-Yang Li[1,3]
Yu Wang[2], Kebin Liu[1], Yunhao Liu[1]
[1]School of Software and TNLIST, Tsinghua University, China
[2]Department of Computer Science, University of North Carolina at Charlotte, USA
[3]Department of Computer Science, Illinois Institute of Technology, USA
[4]Department of Computer Science, University of Science and Technology of China, China
{lan, kebin, yunhao}@greenorbs.com, {cbo1,Yu.Wang}@uncc.edu, {jhou11,xli}@cs.iit.edu

ABSTRACT

Recently a number of systems have been developed to implement and improve the visual communication over screen-camera links. In this paper we study an opposite problem: how to prevent unauthorized users from videotaping a video played on a screen, such as in a theater, while do not affect the viewing experience of legitimate audiences. We propose and develop a light-weight hardware-free system, called KALEIDO, that ensures these properties by taking advantage of the limited disparities between the screen-eye channel and the screen-camera channel. KALEIDO does not require any extra hardware and is purely based on re-encoding the original video frame into multiple frames used for displaying. We extensively test our system KALEIDO using a variety of smartphone cameras. Our experiments confirm that KALEIDO preserves the high-quality screen-eye channel while reducing the secondary screen-camera channel quality significantly.

Categories and Subject Descriptors

C.21 [**Network Architecture and Design**]: Network Communications

Keywords

Screen-Camera Communication; Image Privacy; Image Copyright Protection

1. INTRODUCTION

Recently we have witnessed a blooming of electronic visual displays deployed for a variety of purposes (*e.g.*, for news and entertainment, for advertising, for tour guide, or for human-computer interaction) and in a wide range of devices (*e.g.*, phone-screen, tablet, TV, electronic board). The volume of information exchanged between these visual displays and their audiences is tremendous. For example, video playback has contributed to about 80% of the Internet traffic [15]. Researchers recently propose to encode information into the screen-camera side-channel by taking advantage of the extra signal that can be captured by camera but not the human eye. A number of innovative systems have been developed to implement and improve the visual communication over screen-camera links [14, 20, 22–24, 26–28, 35, 44, 45, 50]. In this paper we study a relevant but different problem: how to prevent unauthorized users from videotaping a video played on a (projected) screen (such as in a cinema or a lecture hall) for high-quality redisplay while do not affect the viewing experience of live audiences. While existing techniques try to maximize decodability of screen-camera channel, we seek to maximize the quality degradation of the display-camera channel while retain the quality of the screen-eye channel. A technology developed in this regard will have a lasting effect in protecting the copyright of the video, preventing audiences from taking a high-quality pirated copy of the video (called *pirate video* hereafter in this work). Film piracy causes lost of revenue about $\$20.5$ billion annually according to a recent survey [6], and over 90% of this illegal online content is delivered from these pirate movies [4]. Unauthorized videotaping during the exhibition could cause unwanted information leakage. So, for the purpose of copyright issues, many exhibitions have strict no-camera policies. Moreover, videotaping a presentation or project demonstration could also cause infringement of copyright and even plagiarism.

Copyright protection is becoming increasingly important. With the rapid spread of camera-enabled mobile devices or wearable devices, recording video in a cinema or a lecture hall is extremely easy and hard to detect. Traditionally, to protect the copyright, copyright is first filed for a digital property indicating that it is protected by law and unauthorized usage is illegal. Various technologies have been developed in the industry and research community for conveying this copyright protection information and/or protecting the digital property copyright from being violated. For example, watermarking [16, 48, 49] schemes are often used to claim the ownership of the released digital property, where a watermark is added to the digital property (*e.g.*, images or videos) while broadcasting. Film industry often implements expensive security strategies or equips guards with night vision goggles [4] to prevent film piracy. Unfortunately, these technologies are ineffective in preventing attendees from taking pirate video for later redisplay. Such intellectual property disrespect behavior has sparked new technologies to prevent illegal video recording, such as inserting extra frames to obscure a recording [56], or projecting ultraviolet or infrared light onto the screen so as to wash out recorded picture [5]. Although such recorded videos contain both valuable video and obscure image or shade, the content could still be more likely perceived by human in large extent. In addition, some technologies cannot be adopted in other display devices, *e.g.*, large LCD screens for personnel use. The goal of this work is to develop a *universal* technology that can be used to protect the video displayed in a variety of devices without introducing extra hardware from pirate videotaping using typical mobile devices, such as smartphone or smartglasses.

MobiCom'15, September 7–11, 2015, Paris, France.

DOI: http://dx.doi.org/10.1145/2789168.2790106.

In contrast to effective screen-camera communications, achieving such a lofty goal of preventing mobile pirate recording is extremely challenging. The main challenging issue is that we have to ensure a smooth view experience/quality of legitimate live audiences while still preventing a good quality pirate-video recording by the adversary audiences. Notice that the recording schemes of smart devices are typically designed to mimic the basics of how human views the surrounding world. Thus, there is a very small design opportunity for such a transformative technology. In this work, we will first extensively explore the basics of human vision, video encoding, screen display, and video-recording mechanisms. By taking advantage of the limited disparities (*e.g.*, the *spectral and temporal color mixture*, the *flicker* effect and *critical flicker frequency*, and the *rolling shutter* of commonly used digital camera) between human vision and video-recording, we develop KALEIDO, a secure video encoding and displaying system. KALEIDO provides smooth viewing experience for human audience but prevents adversaries from recording a good quality pirate-video using mainstream camera-enabled smart devices. As we do not want to use any extra hardware device in our system, KALEIDO is designed purely based on re-encoding the video for screen-display. Notice that the original video is filmed in $24fps$ or $30fps$ (called *video rate*, whose unit is frame per second) typically, while for ensuring better viewing experience, the *refresh rate* of a display screen could be as large as 240fps. To address such disparity, when being played, each frame of the original video will be duplicated so as to fit the refresh rate of each display system. KALEIDO designs a new scheme to "duplicate" the original frame, which we call it *watch-only video re-encoding*. Instead of directly duplicating the original frame, a sequence of d frames ($d \simeq \frac{\text{refresh rate}}{\text{video rate}}$) for each original frame is carefully produced, such that the visual perception of the newly designed frames is same as the original video frame while there is a significant information loss in the pirate video. We have implemented KALEIDO and conducted a comprehensive evaluation over a variety of common mobile cameras.

Summary of Results: We test KALEIDO over 30 videos of different styles, use both LCD monitor and projectors for displaying watch-only videos, and use a variety of smart devices to videotape the video. We evaluate the video quality of watch-only video and pirate video respectively using both subjective video quality measurement (via extensive survey of 50 audiences) and objective video quality measurement with a number of different metrics (such as PSNR and SSIM). Our experiments confirm that KALEIDO preserves the high-quality screen-eye channel while reducing the secondary screen-camera channel quality significantly. First, the viewing experience of legitimate audience is not affected: over 90% of the surveyed audiences do not see any quality differences between the original video and watch-only video. The average score of the survey is over 4 out of 5, indicating the video quality degradation is almost unnoticeable. Second, the scheme is very effective for preventing pirate videotaping: among surveyed audiences, 96% experience a significant quality drop in the pirate video; and the objective video quality measurement of pirate video also confirms this observation (PSNR dropped over 60%, SSIM dropped over 40%). Notice that, due to various techniques used in reducing the quality of the video, the pirate video actually experiences a larger quality degradation when played in real-time than in each of the frames in the pirate video.

The rest of paper is organized as follows. In Section 2, we briefly review the preliminary knowledge about human vision and coloring, the video encoding, the video display, and the video-taping. In Section 3, we highlight the design space and principles, and the design challenges and opportunities. We then introduce our KALEIDO for generating watch-only video. We report results from our extensive evaluation of KALEIDO in Section 4. We review the related work in Section 6 and conclude the paper in Section 7.

2. BACKGROUND AND PRELIMINARY

In this work, we design a special video type, a *watch-only video*, *i.e.* the video can be displayed on common devices and be watched by human with the same visual quality as the original video, but the pirate version, captured by pirates' mobile cameras, will suffer a severe quality degradation. This is challenging as nowadays mobile devices are equipped with sophisticated cameras which are imitation of the human eye. Before presenting our design, we briefly review the properties of the human eye as the information receiver and the constraints that the display and camera technologies place on the transmission of the light signal.

2.1 Characterizing Human Vision

Human possess a photopic vision system, which is driven by the cone-cells in the retina. When we see the rich light spectra of objects, different light wavelengths stimulate the three kinds of cone-cells of a viewer in different degrees, providing her perception of distinct colors. Color is usually recognized by the viewer with two aspects: (1) *luminance*, which is the indication of the "brightness" of the light; (2) *chromaticity*, which is the property that distinguishes the composition of the light spectra.

Color Description: Various models are designed to quantify human color vision. The commonly used 1931 CIE color spaces are the first defined quantitative links between the physical pure colors (*i.e.*, wavelengths) in the electromagnetic visible spectrum and the physiological perceived colors in human color vision. It converts the spectral power distribution of light into the three tristimulus values X, Y, Z. Here Y determines the illuminance (brightness), and X and Z give chromaticity (hue) at that luminance. The chromaticity values can be presented in a CIE chromatic diagram as illustrated in Fig. 1, where coordinates are defined by $x = \frac{X}{X+Y+Z}$ and $y = \frac{Y}{X+Y+Z}$. The diagram represents all of the colors visible to the average person. In the rest of this paper, we use (x, y, Y) values to describe the chromaticity and illuminance of a specific color.

Spectral Color Additive Rule: The colors along any line-segment between two points can be made by mixing the colors at the end points, which is called the *chromatic additive rule*. Specifically, if we have two colored light C_1 and C_2 with values (x_1, y_1, Y_1) and (x_2, y_2, Y_2), and mix the two colors by shining them *simultaneously*, we obtain the mixed color (x, y, Y) denoted by

$$\begin{cases} (x, y) = \frac{Y_1}{Y_1+Y_2}(x_1, y_1) + \frac{Y_2}{Y_1+Y_2}(x_2, y_2) \\ Y = (Y_1 + Y_2)/2 \end{cases} \quad (1)$$

The rule shows that, the chromaticity for the mixed color lies on the line segment joining the individual chromaticities, with the node position on the line segment depending on the *relative brightness* of the two colors being mixed. Clearly, the combination of colors to produce a given perceived color is not unique. For example, the pair C_1C_2, C_3C_4, C_5C_6 in Fig. 1 can each produce the same color C if combined in the right proportions.

Temporal Color Additive Perception: When people watch temporal varying colors, they receive both illuminance change and chromaticity change. When two isoluminant colors alternate at frequencies of $25Hz$ or higher, an observer typically perceives only one fused color, whose chromaticity is determined based on the chromatic additive rule previously discussed. This may also relate to *persistence of vision*, the theory where an afterimage is thought

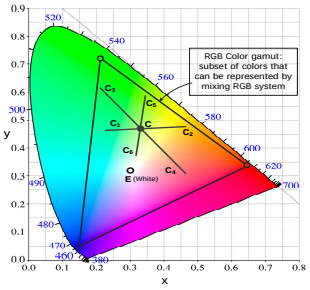


Figure 1: CIE 1931 chromatic diagram and color mixture.

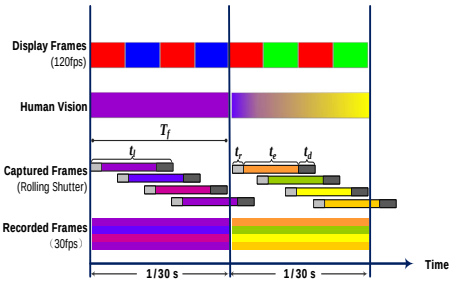


Figure 2: Color perception by human eyes and image capturing by CMOS cameras.

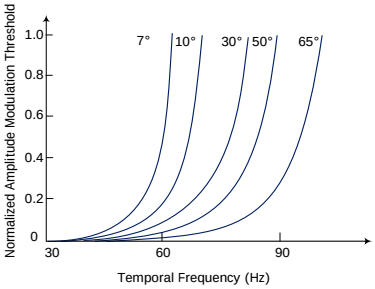


Figure 3: Flicker regression equation for different display field sizes.

to persist for approximately $\frac{1}{16}$ of a second on the retina, which is also believed to be the explanation for motion perception. Fig. 2 illustrates the color fusion result by the human eye. For example, people perceive alternate red and blue as magenta, and alternate red and green as yellow.

Flicker and Critical Flicker Frequency (CFF): Although, human's visual system is very efficient and powerful, its ability to interpret the temporal information presented on video displays is limited. When the change frequency is smaller than eye's *temporal resolution*, called *critical flicker frequency* (CFF), *flicker* happens [8, 25, 32]. Illuminance flicker is a visible illuminance fading between frames displayed on screen when the brightness drop for time intervals sufficiently long to be noticed by the human eye. Chromatic flicker is defined similarly. On the other hand, when the flicker frequency is larger than the CFF threshold, the illuminance flicker stimulus and chromatic flicker stimulus from a sequence of continuous frames are only perceived by human as time-averaged luminance and time-averaged wavelength respectively.

Typically, human eyes can only resolve up to $50Hz$ to *luminance flicker* and $25Hz$ to *chromatic flicker* [25]. Thus our eyes cannot capture fast moving objects and high frequency flickery images [45]. In practice CFF depends on both the spatial and temporal modulation of luminance across the display [18,43]. If the absolute amplitude of the main frequency of the display luminance modulation is greater than a pre-determined frequency-dependent threshold (denoted as $A(f)$) the observers will perceive flicker. Typically

$$A(f) = a \cdot e^{bf} \tag{2}$$

where f is the refresh frequency and a and b are constants that depend on the size of the luminous area. Fig.3 illustrates the equation for different display field sizes where the degree value measures the angle of the smallest cone apexed at eye to cover the display field. Then we have $CFF = \frac{\ln[A(f)/a]}{b}$.

2.2 Video Encoding and Display

Most screens produce a wide range of colors by stimulating the cones of eyes with varying amounts of three primary colors - red, green and blue. The most widely used display devices are LCD monitors and projectors. We cannot display the full range of human color perception with these devices, because the gamut of normal human vision covers the entire CIE diagram while the gamut of an RGB display can be represented as a triangular region within the CIE diagram, with three vertexes are red, green and blue (see Fig. 1).

Video is typically stored in compressed form to reduce the file size, and a number of video file formats were developed. In this work, we will consider a generic video stream consisting of sequential still images, referred to as *frames*. Each frame is a matrix of color pixels. During playing the video, the video stream is decoded and presented by the display system frame by frame. Displaying frames in high frequency (or called *refresh rate*) creates the illusion of moving images because of the phi phenomenon [19]. Modern off-the-shelf LCD monitors and projectors support $120Hz$ refresh rate, and the refresh rate for some game LCD monitors could reach $240Hz$. Since most films are shot in 24 or 30 frame per second, it is common that each original video frame will be repeated several times while being displayed on screen. In KALEIDO, instead of repeating each original video frame directly, we will carefully design these *displayed frames* by changing the color pixel such that the viewing experience of live audiences is not affected and it also prevents high-quality videotaping from third-party cameras.

2.3 Video Recording

When a video is displayed in screen, two communication channels will be investigated: *screen-eye* channel and *screen-camera* channel. The screen-eye channel represents how a human will perceive the displayed video, which has been discussed in Subsection 2.1. Here we will review some important specifics of screen-camera channel, which later we use to design our watch-only video.

Varying Recording Rate: Cameras sense color similarly as the human eye. Each pixel receives light of different wavelength during the exposure time, and fuses them to compute the illuminance and chromaticity values of this pixel. Onboard cameras now could capture high-resolution mega-pixel images at fast frame rate (called *record rate*), which even exceed the perception capability of retina. For example, the record rate of traditional onboard cameras is 24, 30 or $60fps$, while some of latest mobile smartphones, *e.g.*, iPhone 5, iPhone 6 and Samsung Note 4, support up to 120 and $240fps$ in high quality.

Rolling Shutter: CMOS image sensors have become mainstream in onboard cameras for mobile devices, which expose and readout each rows of pixels consecutively [38]. Most of consumer-grade cameras implement such image sensor due to its low energy cost, but this leads to geometric distortion of captured image, called *rolling-shutter effect*. We explain the rolling-shutter mechanism by a simple example as shown in Fig. 2. Assume that, before exposure, each line of a video frame requires duration of t_r second by the camera sensor for resetting the line to query the data. The sensor scans the scene line by line to synthesize the complete image, and for each line, the duration for the sensor exposed to the light is t_e before it takes t_d line scan acquisition time for the driver to dump the data. Then the total acquisition duration (denoted as t_l) for retrieving a line is

$$t_l = t_r + t_e + t_d. \tag{3}$$

Assume the recording rate of a camera is $30fps$, so that the duration for constructing a single frame is $1/30s$, denoted as T_c. Since each frame contains multiple batches of line scans with each line of duration t_l, which are exposed and dumped sequentially and overlapped in parallel, we define *effective light sampling frequency* as the number of lines being captured in one second. Although typical rolling-shutter camera captures an image at its reported frame rate $f_c = \frac{1}{T_c}$, its effective sampling frequency is $f_s = f_c \times n$, where n denotes the actual number of lines in individual images.

Unstable Inter-frame Interval: Generally, the shutter is required to open for a certain duration for sufficient light to complete a single frame, and cameras generate single frame continuously in pre-defined high frequency to record a video. The exposure duration depends on the sensitivity of sensor itself and the actual lighting condition, including the contrast and intensity. Most consumer-graded cameras adjust their frame rate automatically to ensure the frame visual quality for the whole captured video. According to the experiments conducted by [23], some off-the-shelf mobile devices cannot reach nominal frame rate when recording, and the inter-frame time intervals often fluctuate.

3. SYSTEM DESIGN

In this section, we will discuss the design space and principles of KALEIDO, the design challenges and opportunities for implementing KALEIDO, and finally the architecture and our detailed design of watch-only video for KALEIDO.

3.1 Design Space and Principles

We assume that an original video is produced with 24fps or 30fps video rate, and the video will be displayed in some screens with larger refresh rate, say $120fps$. Our goal is to protect the copyrighted video from undesired recording by commercial mobile devices with diversified recording rates, instead of prohibiting the display of the video using mobile devices. The pirate shooting is limited to those using commercial onboard cameras of mobile devices. The professional high-end film cameras are thus excluded. We aim at designing a more radical and effective method to generate a legitimate watch-only version of the videos from the original video. For each original video frame, we will generate a sequence of watch-only frames (precisely $\frac{\text{refresh rate}}{\text{video rate}}$ frames). The watch-only video can be displayed by any off-the-shelf display device. When the watch-only video is displayed normally, viewers will not notice any quality difference from the original one, *e.g.*, without color distortion, artifacts or flicker. But when watching the pirated version recorded by a camera, viewers will suffer a severe intolerable quality degradation.

Leveraging opportunities offered by the limited resolution of human vision system, rolling shutter of the camera and asynchronization between display and camera systems, we propose a system KALEIDO which takes the original video as input and produces a *watch-only* version. As shown in Fig. 4, KALEIDO is an add-on for the current video play system without extra hardware. The piracy procedure typically consists of four steps:

Step 1: the legitimate video is re-encoded and displayed;
Step 2: the pirate shoots the displayed video with a camera;
Step 3: the captured frames are recorded into a video file, which is the pirate version of the original one;
Step 4: the pirate video is displayed for the viewer.

The core of our solution is to re-encode the original video into a *watch-only* one, under the constraint that the viewer's watching experience should be reserved in Step 1; but after the Step 2 and Step 3, the watching experience degradation should be maximized in Step 4.

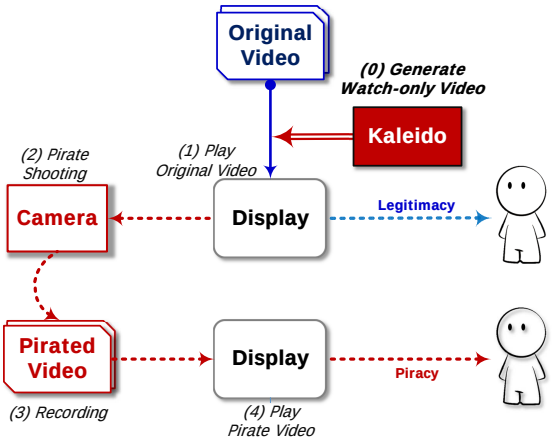


Figure 4: Original display v.s. pirated video display.

3.2 Design Challenges and Opportunities

Our basic approaches for causing a quality loss into the viewing experience of the pirated video are to introduce illuminance flicker and chromatic distortion into the re-encoded frames. The most challenging part of KALEIDO is to ensure the encoded flicker and distortion are imperceivable to the legitimate viewers at first, and then become perceivable after a piracy procedure.

To address these challenges, we will reinvestigate the disparity between the human vision system and the camera system. As shown in Fig. 2, the human eye receives light illuminance and chromatic perturbations in a continuous but low-pass manner, while camera captures light as a discontinuous sampling system with a higher temporal resolution. Taking the continuous frame stream as a varying light signal with specific spatial and temporal color distribution, we exploit the information loss and distortion by camera shooting to look for opportunities. Let the refresh rate of display be f_d and frame record rate of camera be f_c. Then the display duration for each frame is $T_d = \frac{1}{f_d}$ and recording window of a recorded frame is $T_c = 1/f_c$. We analyze the following two complementary cases.

Case 1: $f_d > f_c$. (Display rate is larger than record rate)

In this case, there are multiple frames displayed during a single capture time window T_c by the camera. Remember that the rolling shutter effect of camera causes a line's exposure time t_e be less than T_c. In practice t_e could be less than the half of T_c. Hence, for a specific line in the recorded frame, its exposure time is not enough to record the complete light signal during T_c. If the signal is time-invariant, the line doesn't lose any information. That's why we can record a traditional video ($30fps$) displayed on a $120Hz$ screen using a $30fps$ camera, since it repeats each frame four times. If the signal is time-varying (*i.e.*, re-encoded frames from a single original frame are different), then part of the variation cannot be recorded, *i.e.*, the temporal distribution of the recorded signal in the pirate video deviates from the original video frame. Because eye perceives time-averaging chromaticity and illuminance, the temporal variation loss could cause a perceivable distortion. Besides, different lines in the recorded pirate video lose different portions of the temporal variation, which could cause a spatial deformation of each recorded frame. Fig. 2 presents an example, where $f_d = 120Hz$

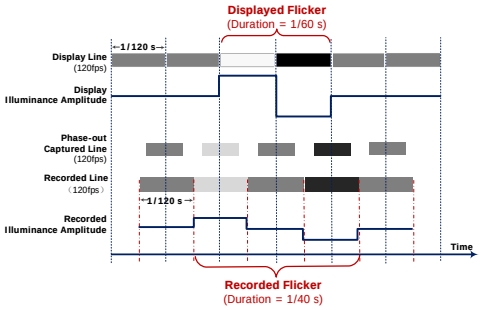


Figure 5: Flicker pollution due to out-phase camera sampling.

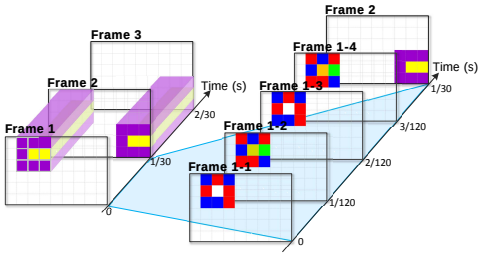


Figure 6: Color decomposition for display frames.

and $f_c = 30fps$. It is a common setting for commercial display devices and cameras. In this figure, two intervals defined by three vertical lines represent two original video frames. The first row represents four encoded display frames for each of the original video frame. During one capture time window, four frames display alternate colors in this case. The second row denotes the colors to be perceived by the human eye. The human vision perceives one color by fusing them equally. The third row denotes the video capture procedure by camera with rolling shutter effect, while the fourth row denotes the recorded two frames by the camera. Each line of the camera captures only part of the display frames, which results in distorted color fusion results for each line. And the recorded image presents a stripped pattern.

Case 2: $f_d \leq f_c$. (Display rate is less than record rate)

In this case, every displayed frame can be captured by at least one recorded frame. If the display system and camera system are ideally synchronized, then during the exposure time of any line of the recorded frame the light signal is constant (from a single display frame) and the camera can record the displayed light signal with high fidelity. In practical applications, with high probability, the camera is asynchronized with the display, which causes out-phase lines in each recorded frame. As illustrated in Fig. 5, one out-phase line captures light signal from two successive displayed frames. If there is a flicker (two successive darker and lighter frames, or vice versa) at a frequency $f_f = f_d/2$, the perturbation will be captured by $2\frac{f_c}{f_d} + 1$ temporal successive out-phase lines. As a result, the flicker is recorded but its frequency is down-converted to $\frac{2}{2+f_d/f_c} f_f$. In the example of Fig. 5, the flicker frequency is down-converted from 60Hz to 40Hz. With this observation, we have an opportunity to encode invisible noises, whose frequency is larger than CFF, to the original video. After the down-conversion by camera recording, the noise could become visible because its frequency now falls below the CFF.

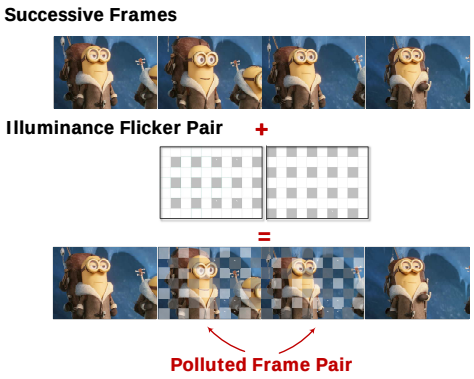


Figure 7: Illuminance frame pollution in display frames.

Additionally, unstable inter-frame intervals of most commercial onboard cameras aggravate the information loss and distortion for both Case 1 and Case 2, hence making the color distortion of recorded frames even worse.

Consequently, during a single capture time window (T_c), rolling shutter effect and unstable intervals could cause a temporal information loss or deviation in the recorded frame if the displayed frames for each original video frame are time-varying. We then propose two techniques to aggravate the temporal information distortion.

Technique 1: Chromatic Frame Decomposition. Since most current videos are $24fps$ or $30fps$, with high refresh rate display devices (*e.g.*, $120Hz$) one video frame can be decomposed into n (*e.g.*, 4 or 5) successive display frames following the temporal chromatic additive rule of human eyes. We propose to decompose one invariant chromatic signal into chromatic flickers, which can be fused by human eye. Fig. 6 shows an example of the color frame decomposition. The chromatic flicker frequency is 60, which is larger than the chromatic CFF. Note that our design is different from the visual cryptography [33], based on the visual effect produced by overlapping multiple transparent slides.

Technique 2: Illuminance Frame Pollution. When there is illuminance fluctuation, human eye works as a low pass filter to eliminate the high-frequency flicker and perceives the averaging illuminance. We propose to add imperceivable illuminance flickers to pollute the frames. As illustrated in Fig. 7, each flicker is a pair of pollution frames. The time averaging illuminance of each pixel from two pollution frames equals 0, which cancels out illuminance change for human eye. But if there is a temporal information distortion in the recorded frames, the flicker cannot be balanced out. Besides, the flicker's frequency is just above the illuminance CFF and its amplitude will be maximized. So if any down-conversion happens, the flicker will become perceivable.

Technique 3: Embrace Spatial Deformation. In company with temporal distortion achieved by chromatic frame decomposition and illuminance pollution, we also design KALEIDO to deform each decomposed frame's shape to prevent image capturing during the video play. Our goal is to make display frames' colors appear as random as possible. Randomizing each display frame's color, while preserving the view experience of legitimate audiences, is possible due to the metamerism. Note that a color can be decomposed to an infinite number of different color pairs. Randomizing different decomposition color pairs will make each display frame like a random noise.

3.3 Kaleido: Watch-Only Video Generation

We are now ready to present KALEIDO by exploiting the main techniques (chromatic frame decomposition, illuminance frame pollution, and spatial deformation) and integrating them to generate watch-only videos.

For simplicity, we consider a $30fps$ video, which consists of N sequential frames $\{V^1, V^2, \cdots, V^N\}$. Each frame V^k is a $R \times C$ matrix, with each pixel's P_{ij}^k color is $C_{ij}^k = (x_{ij}^k, y_{ij}^k, Y_{ij}^k)$. Recall that, (x, y) determines the pixel's chromaticity, *i.e.* coordinates in the CIE diagram, and Y is the illuminance level. In KALEIDO, we focus on an off-the-shelf display device. As a running example to demonstrate our design, we assume that the refresh rate of the display device is $120Hz$. Our scheme can easily be modified to adapt to different refresh rates. We decompose each original frame V^k into 4 display frames (called *sub-frames*) $\{V^{k,1}, V^{k,2}, V^{k,3}, V^{k,4}\}$. Note that all sub-frames have the same duration $1/120s$. To guarantee the flicker frequency greater than CFF and the sub-frames can be fused by human eye, we decompose each frame into two different sub-frames, referred to as fusion pair, and repeat the fusion pair. We then need to determine the $(x_{ij}^{k,l}, y_{ij}^{k,l}, Y_{ij}^{k,l})$ values of each pixel $P_{ij}^{k,l}$ in sub-frames. For $24fps$ video, it can be easily converted to $30fps$ using standard pulldown tools, or each frame can be decomposed to 5 frames, which makes the decomposition more complex, but the principle and techniques are the same.

According to the chromatic additive rule and flicker fusion rule, given the color of a pixel $C = (x, y, Y)$, we need to decompose it to two colors $C_1 = (x_1, y_1, Y_1)$ and $C_2 = (x_2, y_2, Y_2)$, which satisfies

$$\begin{cases} (x, y) & = \alpha(x_1, y_1) + (1 - \alpha)(x_2, y_2) \\ Y & = (Y_1 + Y_2)/2, \end{cases} \tag{4}$$

where $\alpha = \frac{Y_1}{Y_1+Y_2}$. Since the mixed chromaticity (x, y) is a weighted average depending on the relative illuminance of the decomposed two colors, we should determine the illuminance of each pixel first. So, based on the pixel illuminance of the original video, KALEIDO firstly determines the illuminance pollution of the subframe sequence, which gives the final illuminance level (*i.e.*, Y values) of every pixel. Then the illuminance ratio α is fixed. Secondly, (x_1, y_2) and (x_2, y_2) are selected to (approximately) maximize the temporal distortion and spatial deformation.

3.3.1 Illuminance Frame Pollution

Let the initial illuminance levels of every pixel in sub-frames equal to the illuminance in its original frame, say Y_{ij}^k. Given two successive sub-frames $\{V^{k,1}, V^{k,2}\}$, a pixel pair $P_{ij}^{k,1}$ and $P_{ij}^{k,2}$ have illuminance levels $Y_{ij}^{k,1} = Y_{ij}^{k,2} = Y_{ij}^k$. If we can add an illuminance complementary perturbation $(+\delta, -\delta)$ to the pixel pair, it changes their illuminance levels to $Y_{ij}^{k,1} + \delta$ and $Y_{ij}^{k,2} - \delta$. Then the human eye perceives an average illuminance Y_{ij}^k, which equals the original illuminance level if the refresh frequency is above the CFF. In this way, the added complementary perturbation is imperceivable. However, when there is a temporal information loss (as in Case 1), the perturbation cannot be canceled out; when outphase captures happen (as in Case 2), the perturbation's frequency is down-converted. In those situations, the perturbation becomes perceivable flicker to human.

Based on this rule, we can add imperceivable flicker (either $(+\delta, -\delta)$ or $(-\delta, +\delta)$) to pixel blocks of two successive sub-frames to pollute the displayed video. The values of the amplitude δ and block size should be maximized to aggravate the pollution. Remember that the CFF increases as greater amplitude and larger block size, which could cause the flicker perceivable once CFF $> 60Hz$. For example, the CFF is about $90Hz$ when the block size, together with a viewing distance, resulting a 65° angle of view and the normalized amplitude is greater than 0.4 (Fig. 3). So our pollution mechanism chooses the desired CFF as between 50-$55Hz$. Then when the recorded video converts the $60Hz$ flicker down to a lower frequency, *e.g.*, $40Hz$, it will be below the CFF. To obtain a larger space for amplitude modulation, we design the block size as about 10° with respect to that human's vertical field of view at about 120°. Hence, based on Fig. 3, the normalized amplitude of the flicker could be as large as 0.2. For each original frame, we add the block-pattern flicker pair to its second and third sub-frames. Fig. 7 shows an example of illuminance pollution.

3.3.2 Chromatic Frame Decomposition

After the illuminance frame pollution, the illuminance of each pixel in the sub-frames is determined. For a fusion pair, any pair of corresponding pixels have an infinite number of chromaticity combinations to achieve the desired mix color. We propose to choose a set of combinations that will (approximately) maximize the potential color distortion and spatial deformation.

Given a pixel pair P_1 with $C_1 = (x_1, y_1, Y_1)$ and P_2 with $C_2 = (x_2, y_2, Y_2)$ from a fusion pair, to maximize the recorded color distortion, we need to find out the relation between the distortion and the choice of (x_1, y_1) and (x_2, y_2). The correctly fused color is C with coordinate $(x, y) = \alpha(x_1, y_1) + (1 - \alpha)(x_2, y_2)$, here $\frac{Y_1}{Y_2} = \frac{\alpha}{1-\alpha}$. Recall that color distortion happens when the camera fails to capture the complete light signal, which causes the recorded illuminance ratio of P_1 and P_2 deviates from the correct ratio. Let the recorded ratio be $\frac{Y_1'}{Y_2'} = \frac{\beta}{1-\beta}$, then we have $\alpha \neq \beta$. The distorted color is C' with $(x', y') = \beta(x_1, y_1) + (1 - \beta)(x_2, y_2)$. Then the color distortion is

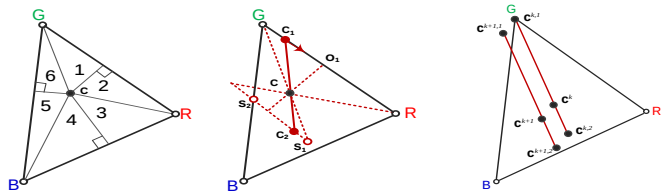
$$D_c(C, C') = |\alpha - \beta| D_c(C_1, C_2).$$

Here α is determined by the original video and the illuminance pollution, while β is determined by the camera's parameters. This shows that, larger $D_c(C_1, C_2)$ can lead to severe potential color distortion. As a result, when we choose two decomposed colors, we need to maximize the distance between them. Note that, the distance is bound by the range of the RGB triangle in the CIE diagram (as shown in Fig. 1). Three vertexes of the RGB triangle are $R = (0.64, 0.33)$, $G = (0.177, 0.712)$ and $B = (0.15, 0.06)$.

Given the color of the original pixel with color $C = (x, y, Y)$, Y_1 and Y_2 of the decomposed pixels's colors C_1 and C_2 are predetermined. Then determination of (x_1, y_1) and (x_2, y_2) is an optimization problem:

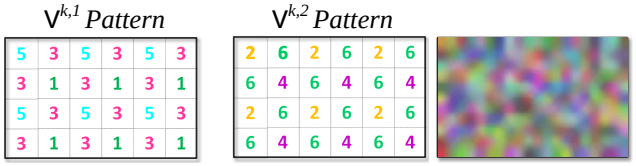
$$\max D_c(C_1, C_2) \quad \text{such that} \tag{5}$$
$$\begin{cases} \frac{D_c(C_1, C)}{D_c(C_2, C)} = \frac{Y_2}{Y_2} \\ \text{both } C_1 \text{ and } C_2 \text{ are within the RGB triangle.} \end{cases}$$

We notice that, the optimal solution must have at least one decomposed color lying on the edge of the RGB triangle. We then propose an algorithm to achieve the optimum with constant time complexity. Our algorithm works as follows. We first divide the RGB triangle into six regions as illustrated in the left figure in Fig.8(a). Then we start to search the local optimum within each region. Within a single region, we find that the optimization objective $D_c(C_1, C)$ changes monotonically (as illustrated in the right figure in Fig.8(a)). Leveraging the monotonicity, one can simply find the optimum in each region using constant computation. We get at most six local optimal solutions. In some regions, there could be no solution. Finally, comparing those six solutions gives us the optimal solution. There is a special case that, three primary colors



(a) (b)

Figure 8: (a) Color space division and solution search, (b) reducing successive pixel decomposition cost.



(a) Patterned Color Selection (b) Mixture

Figure 9: Spatial deformation.

(red, green and blue) cannot be decomposed. So the primary color remains the same in sub-frames.

3.3.3 *Maximize Spatial Deformation*

When determining the decomposed color pair, it is difficult to achieve the tradeoff between maximization of color distortion and spatial deformation. Notice that, unlike visual cryptography, we cannot pick an arbitrary sub-frame and then compute the other sub-frame accordingly s.t. the perceived visual effect is exactly same as the original frame, because the constraints on color additive rule. We propose several simple light-weight methods for achieving a good balance between temporal and spatial quality degradation.

Patterned Color Selection: This is based on the 6 regions divided in Fig. 8, We divide each sub-frame into small grids of certain fixed size. The grid division can be consistent with the illuminance pollution. For one sub-frame of a fusion pair, we assign each grid a region following the left pattern presented in Fig. 9. If a pixel with color C is in a grid labeled Region q, then the decomposed color C_1 will only be searched within the Region q. Thus, the whole sub-frame will have the assigned pattern despite the original shape. The other sub-frame of this fusion pair will use the right pattern in Fig. 9(a) for finding the corresponding color to produce the original color. Fig. 10(b) and (c) illustrate two sub-frames produced for an original frame in Fig. 10(a).

Random Color Selection: For each pixel, we first select a random color, and then compute the corresponding pixel's color in the complement frame. As shown in Fig. 10, although intuitively it will produce a pair of random sub-frames, the actual produced sub-frames still contains a rich shape information. The reason is that the similar color in adjacent pixels may result in similar optimal solution in each region. Fig. 10(d) and (e) show two sub-frames produced.

Mixture of Random and Smoothing: The third method is to use a combination of random choice for each pixel and smoothing among adjacent pixels. First, for each pixel with color C, we randomly select a color C_1 till that there is another color C_2 such that C is produced using color additive rule with C_1 and C_2. Then for each pixel $P^k_{i,j}$, the color of the pixel $P^{k,1}_{i,j}$ in the first sub-frame is an average of the neighboring pixels in this sub-frame. Fig. 9(b) shows an example of chromatic deformation map for the mixture based method. Fig. 10(f) and (g) show two sub-frames produced.

Fig. 10 illustrates the original frame, and a pair of sub-frames produced by these different methods. Note that such randomization and mixture can effectively remove the spatial information in the original frame with different tradeoffs between the viewing experience and anti-piracy ability.

3.3.4 *Reducing the Encoding Cost*

The computational overhead for decomposing the frames of original video into two successive sub-frames pixel by pixel cannot be neglected, especially for high-definition video with 1920×1080 spatial resolution. Leveraging the inherent property of the video, we improve the method to reduce the decoding overhead. Given a normal $30fps$ video, the color for the pixel P_{ij} in both k and $k+1$ frame are $C^k_{ij} = (x^k_{ij}, y^k_{ij}, Y^k_{ij})$ and $C^{k+1}_{ij} = (x^{k+1}_{ij}, y^{k+1}_{ij}, Y^{k+1}_{ij})$ respectively. Thus the color difference,

$$D_c(C^k_{ij}, C^{k+1}_{ij}) = \sqrt{(x^k_{ij} - x^{k+1}_{ij})^2 + (y^k_{ij} - y^{k+1}_{ij})^2},$$

between the same pixel in two successive original video frames could be considered as the chromatic distance in the color space. To reduce the computational overhead, we define ϵ as the threshold of the color difference, and if the difference is less than ϵ the pixel color in the two sub-frame could be calculated directly from the previous one without conducting the color distortion maximization repeatedly.

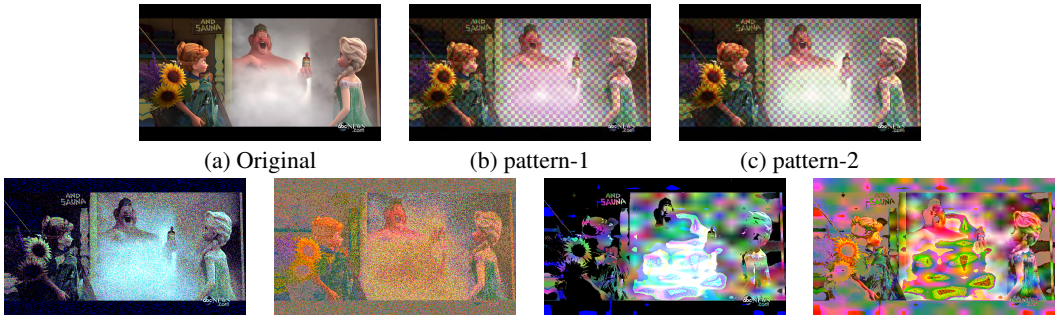
We use a two stage method of complementary color pair determination as illustrated in Fig. 8(b): (1) we draw a parallel line segment to the line segment $C^{k,1}_{ij}C^{k,2}_{ij}$ in the previous frame. (2) the illuminance of pixels in the sub-frames are determined after the illuminance frame pollution. Then we shrink the length of the parallel line segment align with the illuminance ratio so that the adjusted line segment is within the RGB triangle, and determine the coordinate of both $(x^{k+1,1}_{ij}, y^{k+1,1}_{ij})$ and $(x^{k+1,2}_{ij}, y^{k+1,2}_{ij})$ with maximized line segment distance between $C^{k+1,1}_{ij}$ and $C^{k+1,2}_{ij}$.

4. EVALUATION

We now evaluate the performance of KALEIDO via experiments. The prototype of KALEIDO is implemented in C++ with OpenCV library. KALEIDO re-encodes the original video stream into high frame rate video stream, and displays it through regular LCD monitors or projectors. We then evaluate the video quality of both watch-only video and pirate video respectively. We generate the watch-only video through the basic methods mentioned in the previous section, and compare the corresponding pirate video captured by multiple cameras with original video clips to determine whether the content of the original video is protected through some standard video quality assessment metrics. As such objective video quality metrics may not directly reflect the subjective viewing experience by human eyes, we also combine both objective and subjective experiments to measure the effectiveness of the pirate video recording prevention.

4.1 Experiment Settings

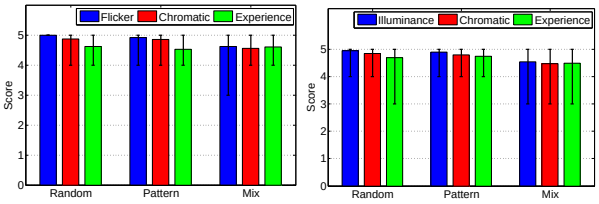
In our evaluation, we use both LCD monitor and projector as the main display. The 27' LCD monitor (AOC G2770PQU) supports 1920×1080 spatial resolution and up to 144Hz refresh rate, while Acer $D600$ projector supports 1280×720 spatial resolution with 120fps frame rate. During the evaluation, we set the frame rate for both two display devices as 120fps. We simulate the working scenarios of both movie and presentation, and verify whether the



(a) Original (b) pattern-1 (c) pattern-2

(d) random-1 (e) random-2 (f) mixed-1 (g) mixed-2

Figure 10: Encoding the subframes for deforming and hiding the spatial information in the frame.



(a) LCD Monitor (b) Projector

Figure 11: Average subjective score.

viewing quality of watch-only video escapes from degradation. On the camcorder side, we use 5 different smartphones (iPhone 5s, iPhone 6, Samsung Note3, Note4 Edge, HTC M8) to capture and record the video projected on the screen.

We also employ a various of video source to examine whether our system could be widely applicable. We select twenty different high-definition (1280×720) video clips with different characteristics on brightness, contrast, and motion. The content of videos are ranged from drama, sports, landscape to animation.

The subjective perception quality of is conducted through users' study. We invite 50 volunteers in aging from 20 to 40 with 31 males and 19 females. All the volunteers have regular visual sensitivity, and one of them is graphic designer with great sensitivity to the video quality.

4.2 Watch-Only Video Quality Assessment

We first evaluate the video quality of the displayed watch-only video. As the original video is re-encoded before displaying, and the content in each frame in the watch-only video is irregular and random, it is difficult to evaluate the quality of the watch-only video objectively by existing standard quality assessment metrics. Thus, we only evaluate its quality by subjective watching experience of volunteers.

In the evaluation, we display both the original video and the watch-only video side by side in two identical display system, and ask them to rate the video quality of the watch-only video by comparing with the original one in three aspects respectively: illuminance level, chromatic correctness, and overall watching experience. Similar to [45], we use score 5 to 1 for each aspect, where 5 indicates the highest quality without any differences in illuminance, chromaticity and the video quality is satisfying; 4 represents the difference being "almost unnoticeable", and 3 to 1 denote "merely noticeable", "evident noticeable", and "strong noticeable or artifact". Since the format of all the selected video clips are high-definition, only scores above 4 indicate the acceptable video quality. We collect the watching experience feedback from all volunteers, and plot the average score in Fig. 11. All the watch-only videos show great smoothness in both projector and LCD monitor, where no jitter is noticed by the audiences. The main subjective difference come from flickers brought by both the illuminance change and the chromatic distortion, which also results in the spacial deformation. The encoding method with random choice of pixel colors provides the best view quality, where the average scores for the first two metrics are both greater than 4.9. 96% volunteers did not even notice they are watch-only video clips. The encoding method with pattern follows with slight drop in performance, because the illuminance and chromatic flicker blocks have larger size than pixels. 92% volunteers did not distinguish them. The encoding method with mixed techniques disturb the original frames mostly, where audiences may experience distortion of both chromaticity and illuminance. Although 38% volunteers noticed that those video clips are re-encoded, but the degradation is acceptable and the average score is above 4.

We also consider other parameters affecting the watching experience, including display devices, different light conditions and different video types. As shown in Fig. 11, LCD monitors have a slightly better performance than projectors, possibly because the projector has a larger display area, which makes illuminance and chromatic flicker easily noticable. Moreover, light condition and video type do not cause significant differences of watching experience.

4.3 Pirated Video Quality Assessment

We then evaluate the performance of our KALEIDO prototype in dealing with piracy camcorder by comparing the pirate video first with the original video clips to present the quality degradation of the pirate watch-only video. However, it is still not easy to determine whether the large amount of quality degradation results from the recording process or the success of frame decomposition. Since multiple factors will lead to quality degradation, and the standard metrics for video quality assessment do not have strong linear correlation to the actual watching experience, it is difficult to compare the definite video quality based on the metric results only. Essentially, the content of the pirate video from regular video is easy to recognize, especially when the recording devices are increasingly powerful. Here we also compare the quality of pirated watch-only video to the pirated video from the original video.

Five smartphones are used to record pirate video, where the capturing rate is $1080p$ in $30fps$ or $60fps$ and $720p$ in $120fps$. The extensive evaluation is conducted in an indoor office with two different light conditions: *nature light condition* representing the p-

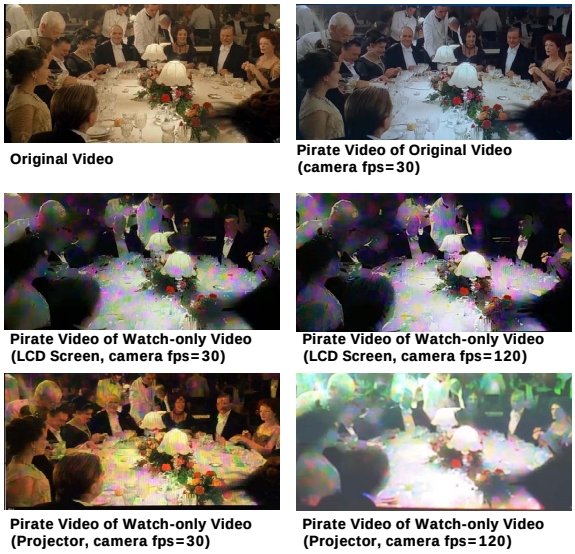


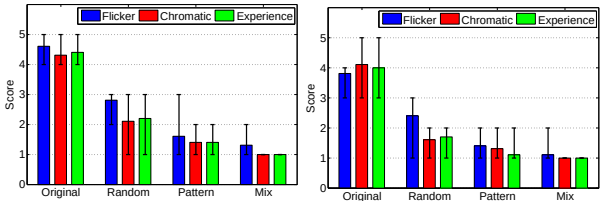
Figure 12: The snapshot of pirate videos with different capturing scenarios.

resentation scenarios and *dark condition* indicating the theater scenarios. The quality of the pirate video will be evaluated both subjectively by watching experience and objectively by standard metrics.

4.3.1 *Subjective Viewing Experience*

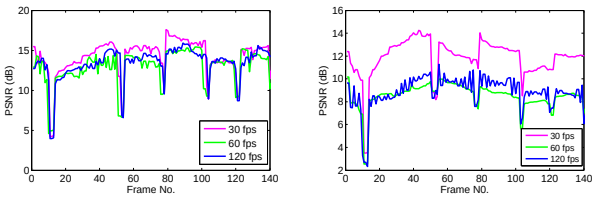
Several facts could affect the pirate video quality, including display device (LCD or projection), camera capture frame rate and light condition. Fig. 12 illustrates the comparison of pirate videos captured using different display devices or fps, where the top-left image is the snapshot of the original video frame. If the pirate video is recorded from the playing of original video, it could still reveal most significant detail of the image, as shown in top-right. When recorded from watch-only video, the content of the frames is difficult to recognize compared with from the original video. For example, the middle two frames come from the watch-only video played in LCD and the last two are displayed by projector. We notice that the pirate video quality degrades with increasing fps, because the flicker frequency down-conversion (as analyzed in Case2 of Section. 3) and more unstable frame interval. Different display devices and light conditions do not cause any significant difference of watching experience. One thing we should keep in mind is that although some of the frames still could be perceived by human eyes, when a sequential of such distorted image frames are played in a regular frame rate, the viewers' viewing experience is significantly affected when playing the pirate video recorded from the watch-only video. Thus, we pay more attention to the overall video quality degradation.

In the subjective assessment, we consider the content of the pirated video, compared to both original video clips and pirated original video clips. We display the original video, pirated video of original video and pirate video of watch-only video side by side in three identical display systems. The rating score for the pirated video is still from 1 to 5 as in previous evaluation. Fig. 13 illustrates the rating for all pirated video clips captured using different display devices (LCD monitor and projector) respectively. The score for pirated original video is about 4, which indicates acceptable quality. Both our watch-only video effectively reduces the quality of



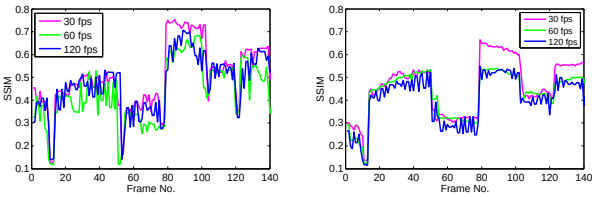
(a) LCD Monitor (b) Projector

Figure 13: Subjective view experiences: pirate original video, pirate watch-only video with various techniques.



(a) LCD Monitor (b) Projector

Figure 14: PSNR in different recording frame rates.



(a) LCD Monitor (b) Projector

Figure 15: SSIM in different recording frame rates.

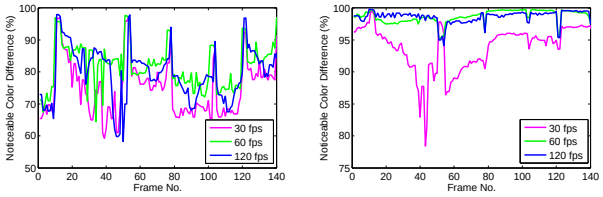
the pirate video from the watch-only video in all tested scenarios. 96% volunteers claim the quality degradation is intolerable, and the average rating score is below 2.

4.3.2 *Objective Measurement*

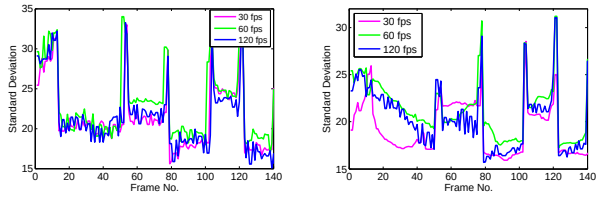
We use five different standard metrics to measure the quality of the pirated video, including PSNR, SSIM, CD, and Histogram. For objective measurement, we setup the evaluation scenario to the finest where the video is being displayed in the screen with largest brightness and the camera is directly facing the screen so that the whole screen could be captured without trapezoid. The usual pirate videotaping scenario would be worse than this ideal testing scenario. Thus, if the quality of the pirate video in this ideal scenario is intolerable, the pirate video taken in worse conditions will experience more severe quality degradation. Due to the disparity between the frame rate of the original video ($30fps$) and the pirate video, we duplicate the frame of original video to align each frame to the captured frames in the pirate video.

PSNR (Peak-Signal-to-Noise-Ratio) is one of the most basic video quality assessment metrics to measure the quality of lossy video compression so as to provide an approximation to human perception of re-encoded video quality. Fig. 14 plots the real-time PSNR for a random selected video clip in different shooting frame rates. The PSNR usually has a value ranging form 30 to 50dB for medium to high quality video [46]. However, the PSNR values fluctuate in a wide range for all pirate video frames, and the values are always below 18dB, indicating a significant quality degradation.

SSIM [47] is proposed as a method to calculate the similarity between two images. The SSIM gets the best value of 1 for two

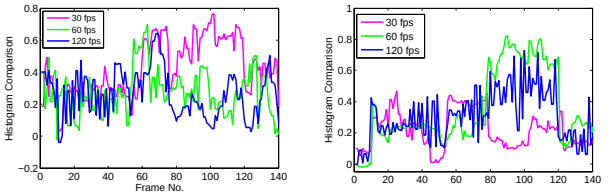


(a) LCD Monitor (b) Projector



(c) LCD Monitor (d) Projector

Figure 16: Color difference in different recording frame rates (proportion and standard deviation).



(a) LCD Monitor (b) Projector

Figure 17: Histogram in different recording frame rates.

identical images, and with the quality decrease, the value of SSIM drops accordingly. The value drops below 0.7 when the image contain large distortion, and the content is difficult to recognized clearly [47]. Our evaluation (Fig. 15) shows that the videos have strong structure distortion when captured by camcorder in three different frame rates from 30fps to 120fps. The average SSIM for LCD are 0.4615, 0.4305 and 0.4070 respectively and 0.5053, 0.4339 and 0.4717 for the projector.

Color Difference (CD, or Chromatic Aberrations CA) is another reliable metric to verify the quality of captured video stream, which usually is generated from a failure of lens to focus all colors to the convergence point. Recording a pirate video will definitely generate color difference, and the value of the color difference determines the amplitude of the color distortion. In this case, we adopt ICEDE2000 [31] to calculate the color difference between the pirate video frame and original video frame on each pixel. When the value of ICEDE2000 exceeds 6, the color difference could be noticed clearly. As the amplitude of color difference usually has nonuniform distribution on the frame, it is ineffective to measure the average color distortion. Instead we calculate the proportion of pixels in each frame that has color difference larger than 6, and compute the variance for those pixels. Based on our evaluation (Fig. 16), such proportion is beyond 70% for most of the video frames and the standard deviation for the color difference is over 21.

We also compare the color histogram of the pirate video to the original video, and plot the correlation for different frame rates in two display systems. As shown in Fig. 17, the value of histogram has no obvious correlation to the frame rate, and all the video shows moderate correlation.

We then evaluate the performance of quality degradation in different decomposition methods (pattern based, random, and mixed) (see Fig. 18). Clearly, all our methods distort the color of original frames, which leads to significant quality degradation in all videos.

We extend our comparison of the video quality degradation in two difference light conditions. The watch-only video is displayed by projector, which is used to simulate the theater and presentation scenarios, and we record the video by two most popular mobile phones in two frame rates. As the results shown in Fig. 19, the pirate video contains similar quality degradation in both environments, and when the camcorder captures the video in lower frame rate, the amplitude of quality degradation is lower than high frame rate.

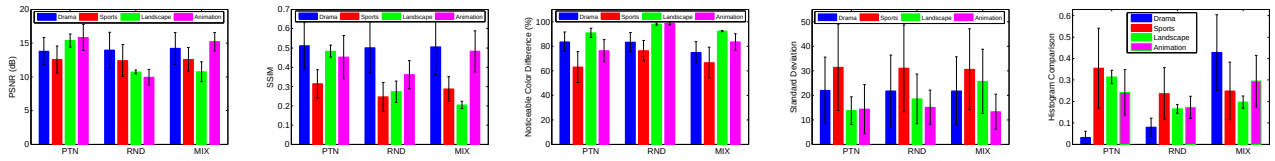
The purpose of the previous experiments is to present the quality degradation of the pirate watch-only video, comparing the original video clips. We now evaluate the same metrics of pirated watch-only video compared with pirated original video (video recorded from playing the original video). As shown in Fig. 20, our results indicate that the pirate watch-only video also has severe quality degradation compared to the pirate original video of non-modified version. Since the pirate original video has already given viewers a unpleasant watching experience, the pirate watch-only video has a much worse quality.

Summary: KALEIDO re-encodes the original video into a watch-only one. The subjective assessment shows that the watch-only video can preserve the viewer's watching experience satisfactorily. And both the subjective and objective evaluation results indicate that the quality of the pirate video from watch-only video is severely degraded in all cases (different combinations of display device, camera fps, video type and light condition) compared to both original video and pirated original video. There is still a room for audience experience optimization, and we will improve it in our future work.

5. DISCUSSION AND OPEN ISSUES

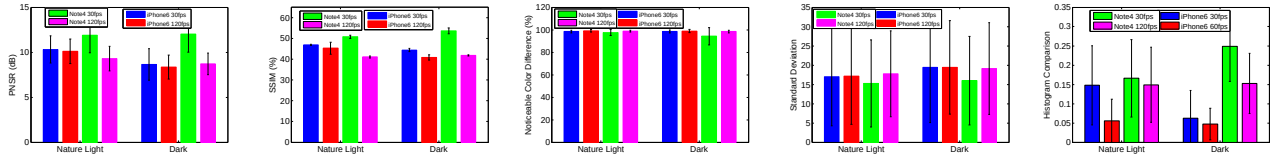
KALEIDO is a first step towards solving piracy problem by generating watch-only videos. While our evaluations demonstrate that KALEIDO is promising, there are some limitations and open problems as discussed below.

System Applicability: In our system design, we leverage the rolling shutter effect to achieve watch-only video against mobile devices. Thus our method may not be working well when facing high-end cameras with global shutter. But, our mechanism could prevent most piracy events caused by current consumer cameras, which is the main focus of this work. On one hand, pirate video captured by personal mobile devices cause great loss to movie industry and severe infringement of copyright. It is easy to prevent high-end professional camcorder from cinema or lecture hall, but it's difficult to forbid attendees to bring personal mobile phones. In MPAA's latest attempt to crack down on piracy, it is pressuring movie theaters to adopt a ban on mobile phones with cameras and certain kinds of eyeglasses, which causes great concern on the security of personal phones and degrades the experience of audiences. On the other hand, the rolling shutter camera dominates the consumer camera market. According to Grand View Research's report about image sensor market [3], by 2013 CMOS image sensors takes 83.8% market share, while CCD image sensors takes only 16.2%. By 2015, CMOS shipments will amount to 3.6 billion units or 97 percent market share, compared to CCD shipments of just 95.2 million, or 3 percent [1]. The majority of CMOS sensors



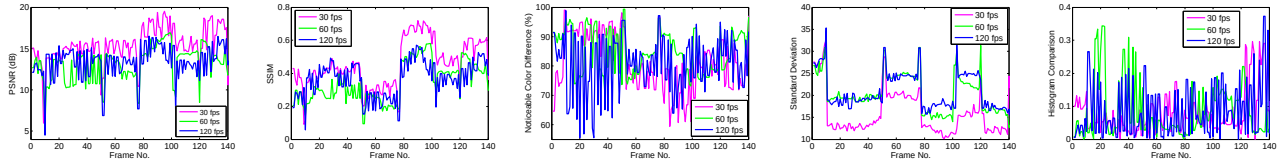
(a) PSNR (b) SSIM (c) CD Proportion (d) CD Std. Dev. (e) Histogram

Figure 18: The quality evaluation for different decomposition methods.



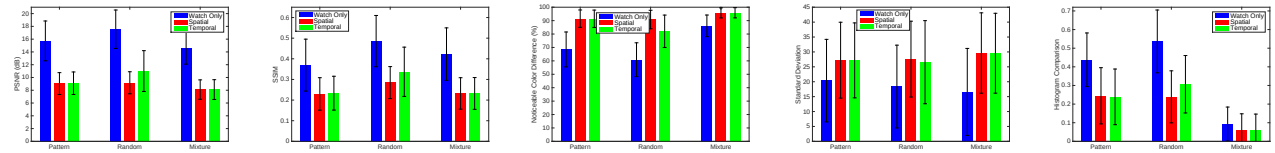
(a) PSNR (b) SSIM (c) CD Proportion (d) CD Std. Dev. (e) Histogram

Figure 19: The video quality assessment in two light conditions.



(a) PSNR (b) SSIM (c) CD Proportion (d) CD Std. Dev. (e) Histogram

Figure 20: Comparing pirate video for both original and watch-only.



(a) PSNR (b) SSIM (c) CD Proportion (d) CD Std. Dev. (e) Histogram

Figure 21: The video quality comparison after noise removal

found in the consumer market utilize a rolling shutter, while only few expensive high-end CMOS sensor can support global shutter, as the global shutter is hard to accomplish in current CMOS designs. We will explore the solution against high-end cameras in our future work.

Post Processing: Since KALEIDO re-encodes the common video into a watch-only one, one of possible attack to our approach is to remove the noise in the video through post processing. Generally, video denoising methods have two different categories: spatial and temporal. Non-local means are the most common spatial video denoising method, which removes the noise at a pixel through certain operations with neighbors within single video frame, such as gaussian weighted average. Although temporal approaches will reduce the noise between frames through tracking blocks along trajectories defined by motion vector and removing the noise of a pixel by taking a number of same pixels from different frames, it is still not suit our watch-only video. In our method we decompose each frame by chromaticity and illuminance in a random manner, and pollute frame temporally and deform frames spatially. As we have described in Section 3.2 due to the rolling shutter effect and unstable inter-frame intervals, there are information loss and distortion in the recorded frames rather than simple Gaussian white noise, and our techniques maximize such loss and distortion. Therefore, it is still difficult to restore original pixels using incomplete and distorted information. To better present that our method is resistent to existing noise removal techniques, we conduct the attack through two mainstream video denoising method: spatial [13] and temporal [12] noise cancelation process, and plot the results in Fig. 21. In this experiment, we compared all the processed videos to the pirate video of original video with standard metrics as before, and we also put the video quality metrics of the watch-only video as comparison. Obviously, the common post denoising process not only cannot recover the original video, but also deteriorate the video quality compared to the watch-only version in all five basic metrics due to recognizing noise incorrectly. Therefore, KALEIDO is insusceptible to common denoising attack, and guarantees the reliable video privacy preservation.

System Overhead: KALEIDO does not generate watch-only video while playing it, but converts the video off-line and loads processed frames to GPU buffer for playing to optimize the watch experience. We evaluate the computation cost for video conversion using a commercial computer with Intel i7-4790 3.6GHZ CPU and

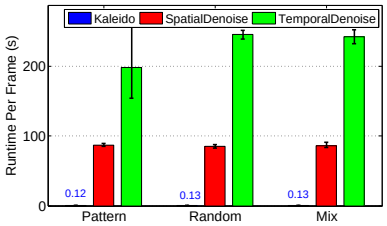


Figure 22: Processing runtime of one video frame for generating watch-only video and denoising. It takes 0.12s for KALEIDO to process one frame.

8G RAM. For software, we employ matrix operations provided OpenCV to achieve optimal performance. As presented in Fig. 22, it takes 0.12s in average to process one 1280×720 video frame, *i.e.*, the process speed is about $8.3fps$. Compared to generating watch-only video, denoising is orders of magnitude slower. Two mainstream video denoising methods cost about 85s and 200s to process one frame separately. For the storage cost, since our method increase the frame rate from $30fps$ to $120fps$, the watch-only video quadruples the file size of the original one.

Watching Experience Degradation: Although KALEIDO can severely reduce the quality of pirate video, we admit that there still has some degradation on watching experience for onsite audiences. Actually, there is a tradeoff between watching experience and piracy prevention. We design our method to maximize the viewing experience and the evaluation results show that the degradation is nearly negligible. There is still a room to improve the viewers' experience, and we will leave this as our future work.

6. RELATED WORK

Screen-Camera Communication: High quality information transmission and decodability maximization [30] are important issues for communication systems. Screen-camera communication systems employ one-way video stream to transmit information [22, 26]. PixNet [35] leverages 2D OFDM to modulate high-throughput 2D barcode frame, and optimizes high-capacity LCDs and camera communication. COBRA [20] achieves real-time phone-to-phone optical streaming by designing a special code layer supporting fast corner detector and blur-resilience technology. VRCodes [50] explores unobtrusive barcode design which is imperceptible to human eyes. Rolling shutter is utilized to simulate high frequency changes of selected color so that only the mixed color is perceived by human eyes. Such idea is also adopted by [23], in addition to address the imperfect frame synchronization. Diversity of cameras will also lead to unsynchronized light-to-camera channel, RollingLight [27] allows a light to deliver information to diverse rolling-shutter cameras while boosting the data rate and the communication reliability. Strata [24] supports wide range of frame capture resolutions and rates so as to deliver information rate correspondingly. Hilight [28, 29] transmit the information by adjusting the hues of the image dynamically. Both InFrame [45] and InFrame++ [44] achieve dual-mode full frame communication between screen and both humans and devices simultaneously. [51] develops an optical communication channel which takes the characteristics of light emittance from the display into account. And [14] proposes a systematic approach to measure the performance of screen-camera communication channel. Our techniques are relevant to screen-camera communication, but also different from it. While existing algorithms try to maximize decodability of screen-camera channel, we seek to maximize the quality degradation of the display-camera channel while retain the quality of the screen-eye channel.

Visual Cryptography: Visual cryptography [33] is a simple but perfectly secure solution for image encryption. Exploiting HVS (Human Visual System) to recognize a secret image from overlapping shares without any additional computation required in traditional cryptography. There are many algorithms to encrypt an image in another image [7, 21, 33, 39]. Rijmen and Preneel [39] propose visual cryptography scheme for color images by expanding each pixel of secret images into a 2×2 block. Hou [21] presents three different methods for visual cryptography of gray-level and color images via exploiting halftone technology and color decomposition. Sozan [7] proposes a different approach by splitting an image into three shares based on three primitive color components.

Image and Video Privacy: Privacy protection is always a broad topic, especially when sharing photos and videos online is becoming increasingly popular [9, 10, 54]. Some efforts have been taken to protect the privacy by concealing a person, blurring faces, masking and mosaicking the selected area of a image [2, 36, 53]. Bo *et al.* [11] proposes a privacy expressing protocol, which requires people to wear a Privacy.Tag to express their privacy desire and the photo sharing services to exert privacy protection by following users' policy expression. By leveraging the sparsity and quality of images to store most significant information in a secret part, P3 [37] extracts and encrypts such small component while preserving the rest in public. Some methods focus on providing privacy preserving image search in a photo database [40, 52]. A number of creative methods [17, 41] were proposed for protecting the privacy of objects in a video. E.g., [34] removes people's facial characteristics from video frame for privacy protection. [42] proposes that denaturing should not only involve content modification but also meta-data modification. [55] designs a protocol to protect portrait privacy when capturing photos. Although existing works could offer privacy protection to the required image/video, they cannot prevent quality pirate video-taping of the displayed video. To the best of our knowledge, our work is the first that allows smooth viewing experience while preventing pirate video-taping.

7. CONCLUSION

In this work we propose a scheme for re-encoding the original video frames such that it can prevent a good quality pirate video-taping of the displayed video using commercial off-the-shelf smart-devices, while do not affect the high-quality viewing experience of live audiences. Our design exploits the subtle disparities between the screen-eye link and the screen-camera link. Extensive evaluations of our implementation demonstrate its effectiveness against pirate video. One remaining work is to improve the encoding efficiency, and reduce the time delay of generating the watch-only video. A more daunting challenge is to design a scheme that can even prevent a good-quality pirate video-taping by high-end professional cameras.

Acknowledgments

This research is partially supported by the National Natural Science Foundation of China under Grant No. 61125202, 61170216, 61228202, 61428203 and CERG-61361166009, and the US National Science Foundation under Grant No. ECCS-1247944, ECCS-1343306, CNS-1319915, CNS-1343355 and CMMI-1436786. We thank all the reviewers and our shepherd for their valuable comments and helpful suggestions.

8. REFERENCES

[1] Cmos image sensors continue march to dominance over ccds. https://technology.ihs.com/389476/cmos-image-sensors-continue-march-to-dominance-over-ccds.

[2] Google street view. http://www.google.com/streetview.

[3] Image sensors market analysis. http://www.grandviewresearch.com/industry-analysis/image-sensors-market.

[4] Pirateeye. http://www.pirateeye.com/.

[5] Preventing movie piracy. http://www.technologyreview.com/news/406063/preventing-movie-piracy/.

[6] Putting a price tag on film piracy. http://blogs.wsj.com/numbers/putting-a-price-tag-on-film-piracy-1228/.

[7] ABDULLA, S. New visual cryptography algorithm for colored image. *arXiv preprint arXiv:1004.4445* (2010).

[8] ANDERSON, S. J., AND BURR, D. C. Spatial and temporal selectivity of the human motion detection system. *Vision research 25*, 8 (1985), 1147–1154.

[9] BADEN, R., BENDER, A., SPRING, N., BHATTACHARJEE, B., AND STARIN, D. Persona: an online social network with user-defined privacy. In *ACM SIGCOMM Computer Communication Review* (2009), vol. 39, ACM, pp. 135–146.

[10] BESMER, A., AND RICHTER LIPFORD, H. Moving beyond untagging: photo privacy in a tagged world. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2010), ACM, pp. 1563–1572.

[11] BO, C., SHEN, G., LIU, J., LI, X.-Y., ZHANG, Y., AND ZHAO, F. Privacy.tag: Privacy concern expressed and respected. In *Proceedings of the 12th ACM Conference on Embedded Network Sensor Systems* (2014), USENIX, pp. 163–176.

[12] BUADES, A., COLL, B., AND MOREL, J.-M. Denoising image sequences does not require motion estimation. In *Proceedings of Conference on Advanced Video and Signal Based Surveillance* (2005), IEEE, pp. 70–74.

[13] BUADES, A., COLL, B., AND MOREL, J.-M. A non-local algorithm for image denoising. In *Proceedings of Conference on Computer Vision and Pattern Recognition* (2005), vol. 2, IEEE, pp. 60–65.

[14] CHEN, C., AND MOW, W. H. A systematic scheme for measuring the performance of the display-camera channel. *arXiv preprint arXiv:1501.02528* (2015).

[15] CISCO, I. Cisco visual networking index: Forecast and methodology, 2013–2018. *CISCO White paper* (2014), 2013–2018.

[16] COX, I. J., KILIAN, J., LEIGHTON, T., AND SHAMOON, T. Secure spread spectrum watermarking for images, audio and video. In *International Conference on Image Processing* (1996), vol. 3, IEEE, pp. 243–246.

[17] DUFAUX, F., AND EBRAHIMI, T. Scrambling for privacy protection in video surveillance systems. *IEEE Transactions on Circuits and Systems for Video Technology 18*, 8 (2008), 1168–1174.

[18] FARRELL, F. Fitting physical screen parameters to the human eye. *Vision and visual dysfunction: The man-machine interface. Boca Raton, FL: CRC* (1991).

[19] GILBERT, G. M. Dynamic psychophysics and the phi phenomenon. *Archives of Psychology (Columbia University)* (1939).

[20] HAO, T., ZHOU, R., AND XING, G. Cobra: color barcode streaming for smartphone systems. In *Proceedings of International conference on Mobile systems, applications, and services* (2012), ACM, pp. 85–98.

[21] HOU, Y.-C. Visual cryptography for color images. *Pattern Recognition 36*, 7 (2003), 1619–1629.

[22] HRANILOVIC, S., AND KSCHISCHANG, F. R. A pixelated mimo wireless optical communication system. *IEEE Journal of Selected Topics in Quantum Electronics 12*, 4 (2006), 859–874.

[23] HU, W., GU, H., AND PU, Q. Lightsync: Unsynchronized visual communication over screen-camera links. In *Proceedings of International Conference on Mobile Computing & Networking* (2013), ACM, pp. 15–26.

[24] HU, W., MAO, J., HUANG, Z., XUE, Y., SHE, J., BIAN, K., AND SHEN, G. Strata: layered coding for scalable visual communication. In *Proceedings of International conference on Mobile computing and networking* (2014), ACM, pp. 79–90.

[25] JIANG, Y., ZHOU, K., AND HE, S. Human visual cortex responds to invisible chromatic flicker. *Nature neuroscience 10*, 5 (2007), 657–662.

[26] JO, K., GUPTA, M., NAYAR, S., LIU, D., GU, J., HITOMI, Y., MITSUNAGA, T., YIN, Q., ISO, D., SMITH, B., ET AL. Disco: Displays that communicate. *IEEE Transactions on Pattern Analysis and Machine Intelligence 99* (2013), 1.

[27] LEE, H.-Y., LIN, H.-M., WEI, Y.-L., WU, H.-I., TSAI, H.-M., AND LIN, K. C.-J. Rollinglight: Enabling line-of-sight light-to-camera communications. In *Proceedings of International Conference on Mobile Systems, Applications, and Services* (2015), ACM, pp. 167–180.

[28] LI, T., AN, C., CAMPBELL, A. T., AND ZHOU, X. Hilight: Hiding bits in pixel translucency changes. *ACM SIGMOBILE Mobile Computing and Communications Review 18*, 3 (2015), 62–70.

[29] LI, T., AN, C., XIAO, X., CAMPBELL, A. T., AND ZHOU, X. Real-time screen-camera communication behind any scene. In *Proceedings of International Conference on Mobile Systems, Applications, and Services* (2015), ACM, pp. 197–211.

[30] LI, Z., DU, W., ZHENG, Y., LI, M., AND WU, D. O. From rateless to hopless. In *Proceedings of International Symposium on Mobile Ad Hoc Networking and Computing* (2015), ACM.

[31] LUO, M. R., CUI, G., AND RIGG, B. The development of the cie 2000 colour-difference formula: Ciede2000. *Color Research & Application 26*, 5 (2001), 340–350.

[32] MASON, R. J., SNELGAR, R. S., FOSTER, D. H., HERON, J. R., AND JONES, R. E. Abnormalities of chromatic and luminance critical flicker frequency in multiple sclerosis. *Investigative ophthalmology & visual science 23*, 2 (1982), 246–252.

[33] NAOR, M., AND SHAMIR, A. Visual cryptography. In *Advances in Cryptology - EUROCRYPT'94* (1995), Springer, pp. 1–12.

[34] NEWTON, E. M., SWEENEY, L., AND MALIN, B. Preserving privacy by de-identifying face images. *IEEE Transactions on Knowledge and Data Engineering 17*, 2 (2005), 232–243.

[35] PERLI, S. D., AHMED, N., AND KATABI, D. Pixnet: Lcd-camera pairs as communication links. *ACM SIGCOMM Computer Communication Review 41*, 4 (2011), 451–452.

[36] POLLER, A., STEINEBACH, M., AND LIU, H. Robust image obfuscation for privacy protection in web 2.0 applications. In *IS&T/SPIE Electronic Imaging* (2012), International Society for Optics and Photonics, pp. 830304–830304.

[37] RA, M.-R., GOVINDAN, R., AND ORTEGA, A. P3: Toward privacy-preserving photo sharing. In *Proceedings of Symposium on Networked Systems Design and Implementation* (2013), pp. 515–528.

[38] RAJAGOPAL, N., LAZIK, P., AND ROWE, A. Visual light landmarks for mobile devices. In *Proceedings of international symposium on Information processing in sensor networks* (2014), IEEE Press, pp. 249–260.

[39] RIJMEN, V., AND PRENEEL, B. Efficient colour visual encryption or ąőshared colors of benettonąŕ. *rump session of EUROCRYPT 96* (1996).

[40] SADEGHI, A.-R., SCHNEIDER, T., AND WEHRENBERG, I. Efficient privacy-preserving face recognition. In *Information, Security and Cryptology*. Springer, 2010, pp. 229–244.

[41] SENIOR, A., PANKANTI, S., HAMPAPUR, A., BROWN, L., TIAN, Y.-L., EKIN, A., CONNELL, J., SHU, C. F., AND LU, M. Enabling video privacy through computer vision. *IEEE Security & Privacy 3*, 3 (2005), 50–57.

[42] SIMOENS, P., XIAO, Y., PILLAI, P., CHEN, Z., HA, K., AND SATYANARAYANAN, M. Scalable crowd-sourcing of video from mobile devices. In *Proceeding of international conference on Mobile systems, applications, and services* (2013), ACM, pp. 139–152.

[43] TYLER, C. W., AND HAMER, R. D. Analysis of visual modulation sensitivity. iv. validity of the ferry-porter law. *JOSA A 7*, 4 (1990), 743–758.

[44] WANG, A., LI, Z., PENG, C., SHEN, G., FANG, G., AND ZENG, B. Inframe++: Achieve simultaneous screen-human viewing and hidden screen-camera communication. In *Proceedings of International Conference on Mobile Systems, Applications, and Services* (2015), ACM, pp. 181–195.

[45] WANG, A., PENG, C., ZHANG, O., SHEN, G., AND ZENG, B. Inframe: Multiflexing full-frame visible communication channel for humans and devices. In *Proceedings of Workshop on Hot Topics in Networks* (2014), ACM, p. 23.

[46] WANG, Y. Survey of objective video quality measurements.

[47] WANG, Z., LU, L., AND BOVIK, A. C. Video quality assessment based on structural distortion measurement. *Signal processing: Image communication 19*, 2 (2004), 121–132.

[48] WOLFGANG, R. B., AND DELP, E. J. A watermark for digital images. In *International Conference on Image Processing* (1996), vol. 3, IEEE, pp. 219–222.

[49] WONG, P. W. A public key watermark for image verification and authentication. In *International Conference on Image Processing* (1998), vol. 1, IEEE, pp. 455–459.

[50] WOO, G., LIPPMAN, A., AND RASKAR, R. Vrcodes: Unobtrusive and active visual codes for interaction by exploiting rolling shutter. In *Proceedings of International Symposium on Mixed and Augmented Reality* (2012), IEEE, pp. 59–64.

[51] YUAN, W., WENGROWSKI, E., DANA, K. J., ASHOK, A., GRUTESER, M., AND MANDAYAM, N. Optimal radiometric calibration for camera-display communication. *arXiv preprint arXiv:1501.01744* (2015).

[52] ZHANG, L., JUNG, T., FENG, P., LIU, K., LI, X.-Y., AND LIU, Y. Pic: Enable large-scale privacy preserving content-based image search on cloud. In *Proceedings of the International Conference on Parallel Processing* (2015), IEEE.

[53] ZHANG, L., JUNG, T., LIU, C., DING, X., LI, X.-Y., AND LIU, Y. Pop: Privacy-preserving outsourced photo sharing and searching for mobile devices. In *Proceedings of the International Conference on Distributed Computing Systems* (2015), IEEE.

[54] ZHANG, L., LI, X.-Y., LIU, K., JUNG, T., AND LIU, Y. Message in a sealed bottle: Privacy preserving friending in mobile social networks. In *IEEE Transactions on Mobile Computing* (2014), IEEE.

[55] ZHANG, L., LIU, K., LI, X.-Y., FENG, P., LIU, C., AND LIU, Y. Enable portrait privacy protection in photo capturing and sharing. *arXiv preprint arXiv:1410.6582* (2014).

[56] ZHAO, J., AND MERCIER, G. Transactional video marking system, Mar. 14 2014. US Patent App. 14/214,366.