CURA: A Framework for Quality-retaining Power Saving on Mobile OLED Displays

CHUN-HAN LIN, National Taiwan Normal University
CHIH-KAI KANG, Academia Sinica
PI-CHENG HSIU, Academia Sinica

Organic light-emitting diode (OLED) technology is regarded as a promising alternative to mobile displays. In this paper, we introduce the design, algorithm, and implementation of a novel framework called CURA for quality-retaining power saving on mobile OLED displays. First, we link human visual attention to OLED power saving and model the OLED image scaling optimization problem. The objective is to minimize the power required to display an image without adversely impacting the user's visual experience. Then, we present the algorithm used to solve the modeled problem, and prove its optimality even without an accurate power model. Finally, based on the framework, we implement two practical applications on a commercial OLED mobile tablet. The results of experiments conducted on the tablet with real images demonstrate that CURA can reduce significant OLED power consumption while retaining the visual quality of images.

Categories and Subject Descriptors: C.3 [**Special-Purpose and Application-Based Systems**]: Real-Time and Embedded Systems

General Terms: Algorithms, Design, Experimentation, Human Factors

Additional Key Words and Phrases: OLED displays, low power, human visual attention, mobile systems

1. INTRODUCTION

Organic light-emitting diode (OLED) technology is deemed a promising display alternative for the emerging genre of mobile devices because, compared to conventional *liquid crystal display* (LCD) technology, it provides brighter colors, wider viewing angles, and faster response times [Shin et al. 2011]. It also allows fabrication on flexible plastic substrates. Unlike LCD displays whose power consumption is dominated by external light sources, OLED displays are self-emissive and their power consumption is highly dependent on the image content (more precisely, the pixel values). The power consumption of OLED displays is almost zero when a black image is presented, but they may use more than twice the power consumed by LCD displays when a white image is shown [Dong and Zhong 2011]. In other words, OLED power consumption increases dramatically with the pixel values of the displayed image. This property

A preliminary version [Lin et al. 2014b] of this paper appears in the Proceedings of the IEEE/ACM Design Automation Conference (DAC) 2014.
This work was supported in part by the Ministry of Science and Technology of the Republic of China under grants 102-2221-E-001-007-MY2 and 104-2628-E-001-003-MY3.
Authors' Addresses: C.-H. Lin, Department of Computer Science and Information Engineering, National Taiwan Normal University, Taipei 116, Taiwan; email: chlin@csie.ntnu.edu.tw; C.-K. Kang, Research Center for Information Technology Innovation, Academia Sinica, Taipei 115, Taiwan; email: akaikang@citi.sinica.edu.tw; P.-C. Hsiu, Research Center for Information Technology Innovation (CITI), and Institute of Information Science (IIS), Academia Sinica, Taipei 115, Taiwan; email: pchsiu@citi.sinica.edu.tw.

DOI: http://dx.doi.org/10.1145/0000000.0000000

presents new opportunities for, as well as challenges to, optimizing the display power consumption of mobile systems.

Displays account for a large proportion of the total power consumption on mobile devices [Dong and Zhong 2012]. A number of approaches based on *backlight scaling*, some of which incorporate *image compensation*, have been proposed to dim the LCD backlight while limiting the distortion and/or maintaining the fidelity of the displayed images [Chang et al. 2004; lranli and Pedram 2005; Bartolini et al. 2009a; Xiao et al. 2013; Lin et al. 2014a]. However, they cannot be applied seamlessly for OLED displays, which obviate external lighting due to their self-emissive nature. In recent years, researchers and vendors have explored various low-power techniques for OLED displays [Dong et al. 2009; Betts-LaCroix 2010; Shin et al. 2011]. An intuitive approach is to darken the portions of the displayed content that are not of interest to the user, referred to as *partial display disabling/dimming* [Betts-LaCroix 2010]. The challenge thus lies in identifying the unimportant parts because they will become invisible or dark to the user [Tan and Balan 2012; Tan et al. 2013; Chen et al. 2014]. Motivated by the observation that different colors require drastically different amounts of power, Dong et al. [Dong et al. 2009] proposed the *color remapping* technique, which changes the original colors of an image into colors that consume less power. The technique can be utilized by graphical user interfaces and applications that do not deal with natural images [Dong and Zhong 2011]. To retain the original visual experience, Shin et al. [Shin et al. 2011] developed a technique called *OLED dynamic voltage scaling*, which tries to reduce the supply voltage of each pixel's circuit to the minimum that is sufficient to maintain the pixel's (luminance) value. As the technique requires hardware support and incurs extra circuit costs [Shin et al. 2013], a display panel is normally partitioned into a limited number of rectangular regions [Chen et al. 2012; Zhao et al. 2013].

In this paper, we introduce an alternative low-power technique called *image pixel scaling*, which leverages the flexibility provided by OLED technology to scale down the pixel values of different-shaped regions. The technique allows us to incorporate *human visual attention* into the quality-retaining power-saving design for mobile OLED displays. Our notion is based on the following interesting findings of psychological experiments [Lin and Kuo 2011]: 1) the regions of an image receive varying degrees of visual attention; 2) different regions can tolerate different degrees of image distortion; and 3) not every change is noticeable. The psychological findings also provide the foundations for our realization of the notion. In addition to the above notion, the technical contributions of this paper are as follows.

To realize the above notion, we designed a framework called *CURA*, which tries to *Catch UseR Attention* for quality-retaining OLED power saving. First, we employ mature techniques developed in the *computer vision* field as the links between human virtual attention and OLED power saving. The design of CURA involves *attention region segmentation*, *region distortion assessment*, and *boundary effect elimination*. Second, we model the fundamental problem of carrying out CURA as an *OLED image scaling optimization problem*. The objective is to minimize the power required to display an image, without violating some constraints inspired by the psychological findings to retain the image's visual quality. Third, we propose an efficient algorithm to solve the formulated problem optimally, even if the OLED display's accurate power model is unknown; that is, the optimal solution derived by CURA can be applied to any OLED device. Fourth, we address some technical implementation issues and utilize CURA in two practical applications, namely an image converter and a power-saving mode, for commercial mobile devices. We also elaborate on how the two applications are implemented as a stand-alone app and in the Android kernel on a Samsung Galaxy Tab 7.7, respectively. Finally, to obtain further insights, we conduct extensive experiments on

the Samsung tablet using images with different characteristics, and compare CURA with the closely related, state-of-the-art approach proposed in [Chen et al. 2012]. The results demonstrate that CURA can achieve substantial OLED power savings between 38% and 42% with an acceptable computational overhead, while retaining satisfactory visual quality.

The remainder of this paper is organized as follows. In Section 2, we provide background information and discuss the rationale behind image pixel scaling. Then, we present the CURA framework's design details, fundamental algorithm, and system implementation in Section 3; report the experiment results in Section 4; and review related work in Section 5. Section 6 contains our concluding remarks.

2. BACKGROUND AND DESIGN CHALLENGES

In this section, we provide some background knowledge about OLED power models and human visual attention. We also discuss the rationale behind the image pixel scaling technique and point out some major design challenges.

2.1. OLED Power Models

Usually, an organic light-emitting diode comprises two organic layers, namely an *electron* transport layer and a *hole* injection layer, embedded between the cathode and the anode. When a voltage is applied across the diode, electrostatic forces make electrons leave the cathode while holes are transported from the anode in the opposite direction. Then, the electrons and holes recombine to form a photon with a frequency in the visible region of light [Burrows et al. 1997]. As the power applied to the diode is transformed into light, the diode is emissive without external lighting. Furthermore, different materials and dopants can be used to produce colored diodes that emit red, green, or blue light. Different combinations of the diodes can be used to create various colors.

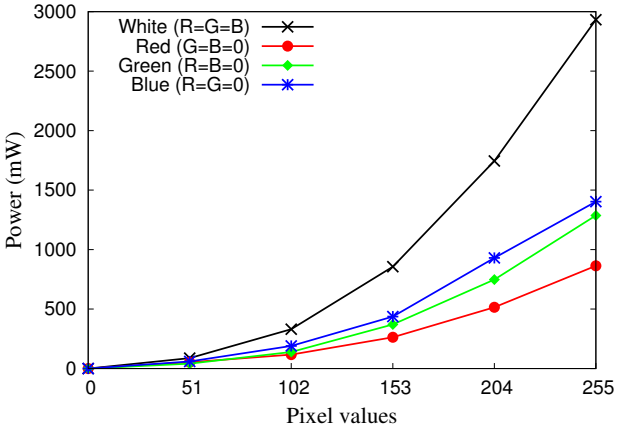
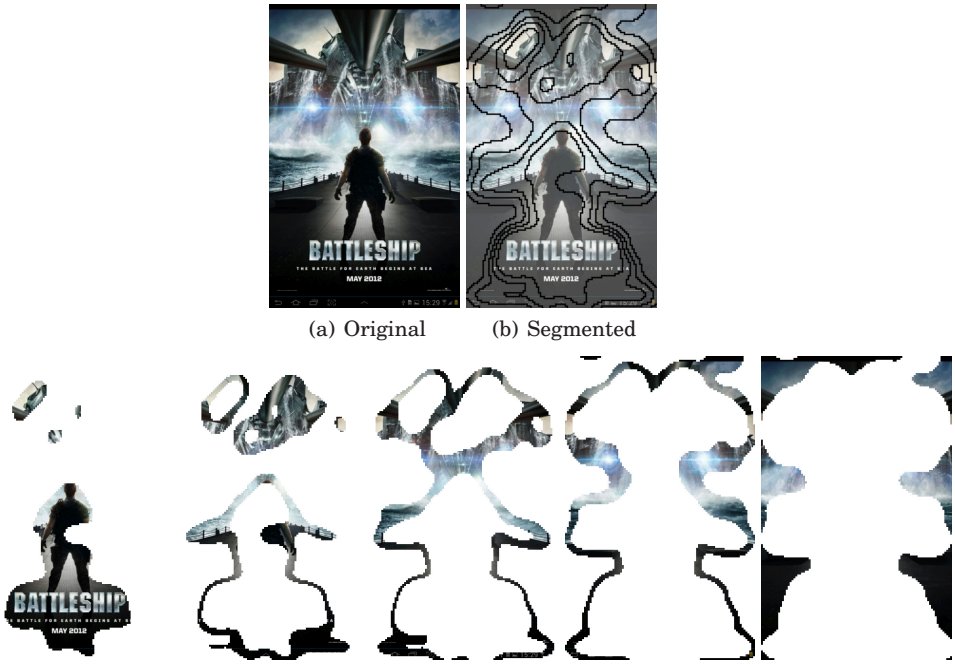


Fig. 1. An OLED power model

An OLED display comprises a matrix of pixels, each of which contains three diodes that emit red, green, and blue light respectively. The diodes have different luminance efficacy. The current of electrons that flows through a diode is determined by its pixel value, and is in the range 0 to 255. Consequently, the amount of power that a diode consumes varies significantly depending on its value and color. In other words, the power required to display an image depends primarily on the image content. As modeled in

[Dong and Zhong 2011; 2012; Chen et al. 2012], the power consumption of an image on an OLED display is the sum of the power consumption of all the pixels; and the power consumed by each pixel is the sum of the power consumed by its RGB subpixels. Therefore, the power model of an OLED display can be expressed by a strictly increasing function, denoted by $P(x)$, of the pixel value x; however, the increasing slopes may vary between OLED displays. Figure 1 shows the power model measured based on a Samsung Galaxy Tab 7.7, where the power consumption increases dramatically with the pixel value and differs from color to color.

2.2. Human Visual Attention



(a) Original (b) Segmented

(c) 5 Attention Regions (from high to low)

Fig. 2. Five attention regions of an image segmented based on human visual attention

Primates have a remarkable ability to interpret complex scenes in real time. To reduce the complexity of scene analysis, the *human visual system* selects a subset of the available sensory information for more detailed processing. The selection appears to be implemented in the form of a spatially circumscribed region of the visual field, called the *focus of attention* [Itti et al. 1998]. Psychological experiments have demonstrated that not every pixel or region in an image receives the same level of attention [Lin and Kuo 2011]. This is because the human visual system's selectivity initially responds to the most attractive parts of a scene, with subsequent eye movements from one fixed location to another. Thus, based on visual attention, an image can be segmented into a set of disjoint attention regions with different attention levels. Note that an attention region may comprise several non-adjacent subregions. Figure 2 shows an image with five attention regions segmented according the level of human visual attention.

Because the regions of an image receive varying degrees of attention, the same magnitude of change to different regions may yield different perceptual effects [Lin and

Kuo 2011]. In other words, the regions can tolerate different degrees of *image distortion* in inverse proportion to the level of attention they receive. Thus, the magnitude of change applied to each region should be limited and reflect the region's tolerable distortion. Importantly, any change in a pixel's value will lead to a change in the *chrominance* and/or the *luminance* of the pixel. Changes in chrominance are much more noticeable than those in luminance. To maintain the chrominance of a pixel when its value is changed, the same magnitude of change must be applied to its RGB subpixels. Moreover, the pixel values of each region should be scaled down by the same magnitude; otherwise, the phenomenon of *luminance gradients* may occur within a region if the region's pixel values are scaled down by a variety of magnitudes.

Applying different magnitudes of change to different regions may result in sharp edges between adjacent regions that will severely impact the visual experience [Chen et al. 2012]. However, not every change in an image is noticeable to the human visual system [Lin and Kuo 2011]. The *just noticeable difference* is the minimum amount by which the stimulus intensity must be changed in order to produce a noticeable variation in sensory experience [lranli and Pedram 2005]. Therefore, to ensure that the region boundaries are too indistinct to be discerned by the human eye, we can restrict the magnitudes of change applied to two adjacent regions.

2.3. Rationale and Design Challenges

The image pixel scaling technique allows a trade-off of image quality to save OLED power. Because the OLED power model is a strictly increasing function of the pixel value, lowering an image's pixel values (by simply multiplying each pixel's RGB values by a scaling ratio in the range 0 to 1) is an effective way to reduce its power consumption on an OLED display. However, this can adversely impact the visual experience if the pixel values are not scaled down appropriately. Motivated by some interesting findings of psychological experiments, we study the practicality of introducing human visual attention into OLED power saving.

Realizing the above notion on mobile systems raises three technical challenges. The first is *how to link human visual attention to OLED power saving*. Changing an image's pixel values inappropriately may result in serious distortion of the image; thus, it is essential to define and impose appropriate *scaling constraints* on changes made to the pixel values. The objective is to ensure that users will be unaware of, or at least have difficulty recognizing, whether the image has been altered by the image pixel scaling technique. To this end, we segment an image into a set of attention regions. We also need to determine the maximum change that each region can tolerate and the maximum difference between the changes applied to two adjacent regions.

The second challenge is *how to minimize OLED power consumption and retain the image quality*. This involves determining an appropriate scaling ratio for each attention region without violating the defined scaling constraints. Mobile devices may utilize OLED displays with different power models; hence, the power-saving version of an image optimized for one OLED display will not necessarily consume minimum power on other mobile devices. We will prove that the proposed algorithm is optimal in terms of power savings under the scaling constraints, even without an accurate power model measured in advance. This property is particularly important if the power-saving image is intended for use on any OLED display and thus requires portability.

Finally, we need to determine *how to apply the image pixel scaling technique to commercial mobile devices*. We delineate two possible application scenarios, namely an image converter and a power-saving mode. In the first scenario, after an image has been converted on a personal computer, its power-saving version can be displayed on any OLED mobile device. The second scenario is more challenging because, in addition

to the issue of compatibility with existing mobile operating systems, the computational overhead incurred by processing a screen scene should be acceptable on mobile devices.

3. QUALITY-RETAINING OLED POWER SAVING

Next we present the CURA framework. CURA exploits existing techniques in the field of computer vision to characterize an image as a set of parameters and defines two scaling constraints that retain the image's visual quality (Section 3.1). Then, CURA uses the proposed algorithm, which takes the above parameters as the input, to optimize the power required to display the image on any OLED display while satisfying the constraints (Section 3.2). Finally, some technical implementation issues are addressed so that CURA can be utilized in two practical applications (Section 3.3).

3.1. Links Between Virtual Attention and Power Saving

3.1.1. Attention Region Segmentation. Not every pixel or region in an image receives the same amount of attention. *Visual saliency* is the distinct subjective perceptual quality that makes some pixels stand out from their surroundings and become the focus of human attention. Feature integration theory and behavioral studies [Treisman and Gelade 1980; Desimone and Duncan 1995] show that important visual features include intensity, color, and orientation; and different features will be more significant in some image scales than in others. Usually, a number of feature maps of various scales are created to capture the magnitudes of different visual features. The maps are then combined by specially-designed normalization operators to form a *saliency map* that represents the conspicuousness of scene locations. Several visual attention models have been developed [Borji and Itti 2013]. If a model produces saliency maps, it can be exploited in CURA for attention region segmentation.

Given an image, we explain how we segment it into a set R of attention regions with an adjacent matrix A. In our implementation, we exploit *Itti's model* [Itti et al. 1998], which was specially designed for still images, to evaluate human visual attention. For the given image, Itti computes a saliency map that quantitatively evaluates each pixel's conspicuousness in the range 0 to 1, with 1 representing the highest attention level. Pixels with similar saliency values have similar strength to attract visual attention. Therefore, we divide the range evenly into N subranges, and classify all the pixels into N corresponding attention regions according to their saliency values. Consequently, each region's shape is formed by the pixels with saliency values in the same subrange.

To create the adjacent matrix A that corresponds to the set R, all the $N \times N$ entries of A are initialized at 0 first. Then, the entry $A[i, j]$ is updated to 1 if two adjacent pixels in the image belong to different regions r_i and r_j. The process involves scanning all the pixels in the image from left to right and top to bottom in sequence. During the scan, for each pixel x, we examine the pixel to the right of x and the pixel below x to check whether either of them belongs to another region (i.e., does not belong to the region that x belongs to). Note that R and A can be constructed simultaneously by scanning the saliency map once sequentially. Thus, the time required to segment an image based on its saliency map is linear to the number of pixels.

3.1.2. Region Distortion Assessment. The regions of an image segmented based on human visual attention can tolerate different degrees of image distortion, which is normally defined as the resemblance between the original image and the pixel-scaled image [Wang et al. 2004]. It is known that the human visual system processes achromatic and chromatic signals separately, and visual signals are differentiated by their frequency and orientation. Thus, both the original image and the pixel-scaled image are usually decomposed into different colors, spatial channels, and temporal channels

to facilitate the detection of features that are common to them. The features' values are then pooled and reduced to a single number to represent the quality score. A number of visual quality metrics have been proposed to assess the distortion of an image [Lin and Kuo 2011].

Given a set R of attention regions, for each region $r_i \in R$, we explain how to derive a *critical scaling ratio* $c(i)$, which represents the lowest ratio that will not violate the tolerable distortion. This *distortion constraint* is defined to restrict the scaling ratio applicable to each region and thus limit its distortion. Our implementation utilizes the *structural similarity* (SSIM) index [Wang et al. 2004], a metric specially designed to comply with the perception of the human eye and widely used in related studies [Chen et al. 2012; Lin et al. 2014a], to assess image distortion. The resultant SSIM score is a decimal value between -1 and 1, where 1 is only achievable in the case of two identical images. Let the SSIM index required for the region with the lowest attention level be set at s. We assign each of the N regions an SSIM requirement uniformly distributed over the range s to 1 based on their attention levels; that is, the SSIM index required for the region with the ith highest attention level is set at $1 - \frac{i \times (1-s)}{N}$. The reason for the uniform assignment of SSIM requirements is that the decline in visual quality is roughly linear to the decrease in the SSIM score, as indicated in [Bartolini et al. 2009b]. Then, for each region $r_i \in R$, we find its critical scaling ratio $c(i)$, which satisfies the SSIM requirement assigned to r_i, by performing a binary search on all 256 possible scaling ratios (in the range 0 and 1 with an interval of $\frac{1}{255}$). This means that eight SSIM scores have to be computed to derive a region's critical scaling ratio. Moreover, our implementation derives a region's critical scaling ratio by applying SSIM to the smallest rectangle that covers all the pixels of the region, instead of the any-shaped region, because SSIM assesses similarity by considering the structural information in rectangular regions.

3.1.3. Boundary Effect Elimination. The human visual system cannot notice every change in an image because of visual recognition processes in the brain. The just noticeable difference refers to the smallest perceptible difference between two intensity levels of a particular sensory stimulus. For many sensory modalities, such as the brightness of lights, the just noticeable difference between two stimuli is proportional to the magnitude of the stimuli. Weber's Law states that the ratio of the just noticeable difference to the stimulus intensity is a constant [Weber 1834; Acharya and Ray 2005]. Specifically, $\frac{\Delta I}{I} = d$, where I is the original intensity, ΔI is the smallest perceptible addition to or subtraction from I, and d is a constant.

If the critical scaling ratios of two adjacent regions are applied directly, sharp edges may occur and be perceived by the human eye. We exploit the just noticeable difference to eliminate the boundary effect caused by image segmentation in the image pixel scaling technique. A sharp edge occurs because the pixels on the two sides of a region boundary are similar (in terms of their values), but they become very different after the scaling ratios are applied to them. Thus, based on Weber's law, we define the *differential constraint*, which requires that the difference between the scaling ratios applied to two adjacent regions is not greater than either scaling ratio multiplied by a *differential constant* d. The rationale behind this constraint is that if two pixels are similar and adjacent in the original image, the difference between their scaled-down values should be limited to the just noticeable difference.

3.2. OLED Image Scaling Optimization

3.2.1. Problem Definition. We have explained how to segment an image into a set R of attention regions with an adjacent matrix A; and why it is necessary to determine a critical scaling ratio $c()$ for each region and a differential constant d for two adjacent

regions. Next we consider how to assign an appropriate scaling ratio to each region so that the power required for an image on any OLED display is minimized. For every region $r_i \in R$, the determination of a scaling ratio $\sigma(i)$, in the range 0 to 1, is called a *scaling assignment*. Such an assignment is *feasible* if both the distortion constraint (i.e., $\sigma(i) \geq c(i)$) and the differential constraint (i.e., $|\sigma(i) - \sigma(j)| \leq d \times \sigma(j)$ if $A[i,j] = 1$) are satisfied, $\forall r_i, r_j \in R$. We formally define the OLED image scaling optimization problem as follows.

Instance: A set of attention regions $R = \{r_1, r_2, r_3, ..., r_N\}$ and its adjacency matrix A, where each region $r_i \in R$ is associated with a critical scaling ratio $c(i)$; and a differential constant d for any two adjacent regions.

Objective: A feasible scaling assignment σ such that the total power consumption, $\sum_{r_i \in R} \sum_{x_k \in r_i} P(\lceil x_k \sigma(i) \rceil)$, is minimized, where $P(x)$ is any OLED power model.

Algorithm 1

Input: A region set R with an adjacency matrix A, as well as critical scaling ratios $c()$ and a differential constant d

Output: A feasible assignment σ

1: $\sigma \leftarrow c$
2: $L \leftarrow R$
3: Sort L in nonincreasing order
4: **while** $L \neq \emptyset$ **do**
5: $r_i \leftarrow$ the first region in L
6: $\hat{r} \leftarrow$ find the last region whose key $\geq \frac{\sigma[i]}{1+d}$ in L
7: **for all** r_j behind $\hat{r}$ in L **do**
8: **if** $A[i,j] = 1$ **then**
9: $\sigma[j] \leftarrow \frac{\sigma[i]}{1+d}$
10: Move r_j to immediately behind $\hat{r}$
11: Remove r_i from L
12: **return** σ

3.2.2. Algorithm Description. Given an image represented by a set R of regions with an adjacency matrix A, as well as the critical scaling ratios $c()$ and a differential constant d, Algorithm 1 determines a feasible scaling assignment σ such that the power required for the image on any OLED display is minimized. Initially, each region's scaling ratio is set at its critical scaling ratio and stored in an array σ (Line 1). Throughout the algorithm, we maintain a linked list L that initially contains all the regions in R, keyed by their current scaling ratios (Line 2). The list is sorted in non-increasing order according to the keys (Line 3). Then, we examine all the regions in L and remove them in sequence until L is empty (Line 4). Let r_i be the first region in L currently (Line 5), and let $\hat{r}$ be the last region whose key is larger than or equal to $\frac{\sigma[i]}{1+d}$ (Line 6). All the regions after $\hat{r}$ in L are examined to determine whether their scaling ratios should be updated (Line 7). Let r_j be any region that is after $\hat{r}$ in L and adjacent to r_i in the image (Line 8). Note that $\sigma[j]$ is smaller than $\frac{\sigma[i]}{1+d}$ because r_j is behind $\hat{r}$. To satisfy the differential constraint, the difference between $\sigma[i]$ and $\sigma[j]$ cannot be larger than either ratio (or, equivalently, the smaller ratio $\sigma[j]$) multiplied by the differential constant d. Thus, region r_j should be assigned a scaling ratio of at least $\frac{\sigma[i]}{1+d}$ (Line 9). Moreover, to maintain the non-increasing order of L, r_j is moved to the position immediately after $\hat{r}$ (Line 10). When all the regions after $\hat{r}$ have been examined, region r_i is removed from

L (Line 11). Then, its scaling ratio is determined and will not change thereafter. In the last step, the scaling assignment σ is returned (Line 12).

3.2.3. Properties. Next, we analyze the time complexity of Algorithm 1 and prove that it solves the OLED image scaling optimization problem optimally without the information about the OLED power model.

LEMMA 1. *The time complexity of Algorithm 1 is* $O(N^2)$.

PROOF. The initialization of array σ and linked list L in Lines 1 and 2 can be completed in $O(N)$ time. Sorting L in Line 3 takes $O(N \ln N)$ time. The **while** loop in Lines 4 to 11 is executed exactly N times. Within the loop, assigning the first region in L to r_i in Line 5 costs $O(1)$ time, while finding the region $\hat{r}$ in Line 6 requires at most $O(N)$ time. Because there are at most $N-1$ regions after $\hat{r}$, the **for** loop in Lines 7 to 10 will be iterated no more than $N-1$ times. For each iteration, an update to $\sigma[j]$ and a move in L (if necessary) both take O(1) time. Finally, removing r_i from L in Line 11 also costs $O(1)$ time. It is obvious that the **for** loop dominates the **while** loop in terms of the running time. Thus, the time complexity of Algorithm 1 is $O(N^2)$. □

LEMMA 2. *Throughout Algorithm 1, the scaling ratio of any region will not decrease after initialization.*

PROOF. After being initialized in Line 1, the scaling ratio of a region can only be updated by the assignment in Line 9. The scaling ratio will not decrease if we can prove that $\sigma[j] \leq \frac{\sigma[i]}{1+d}$ immediately before the assignment. Because $\hat{r}$ is the last region whose key is not smaller than $\frac{\sigma[i]}{1+d}$ in list L and r_j is behind $\hat{r}$, we can derive that $\sigma[j]$ must be smaller than $\frac{\sigma[i]}{1+d}$. Thus, the lemma follows. □

LEMMA 3. *The scaling assignment σ returned by Algorithm 1 is feasible.*

PROOF. We prove this lemma by showing that σ satisfies the distortion and differential constraints. The scaling ratio of each region r_j is initialized as its critical scaling ratio and will not decrease (according to Lemma 2); thus, the scaling assignment σ satisfies the distortion constraint. Moreover, the scaling ratio of r_j can be only updated when an adjacent region r_i is to be removed from L. After being updated, the scaling ratio of r_j will be fixed at $\frac{\sigma[i]}{1+d}$ and will not change throughout the algorithm. The reason is that a subsequent update to the scaling ratio would reduce the ratio (and violate Lemma 2) because L is maintained in non-increasing order. Furthermore, the scaling ratio of r_i is determined and will not change once r_i is removed from L. Thus, the scaling ratios determined for two adjacent regions, r_j and r_i, in σ are subject to the differential constraint. □

LEMMA 4. *The power consumption under σ returned by Algorithm 1 is no higher than that under any feasible assignment.*

PROOF. We prove this lemma by contradiction. Suppose there exists a feasible assignment σ' that requires less power consumption than σ. As $P()$ is a strictly increasing function, there must be at least one region whose scaling ratio in σ' is lower than that in σ. Let r_j be the first region with $\sigma'[j] < \sigma[j]$ during the execution of Algorithm 1. Immediately before r_j is removed from L, we should have $\sigma'[j] < \sigma[j]$ and $\sigma'[i] \geq \sigma[i], \forall r_i \notin Q$. We delineate two possible cases, depending on whether any adjacent regions of r_j have been removed from L.

(1) If no adjacent region of r_j has been removed from L, its scaling ratio has not been updated yet. Thus, $\sigma'[j] < \sigma[j] = c(j)$.

(2) Alternatively, some adjacent regions of r_j have been removed from L. Let r_i be an adjacent region whose scaling ratio is the maximum among the ratios of all the adjacent regions that have been removed from L. Then, $\sigma'[j] < \sigma[j] = \frac{\sigma[i]}{1+d} \leq \frac{\sigma'[i]}{1+d}$. In other words, $\sigma'[i] - \sigma'[j] > d \times \sigma'[j]$.

To conclude, $\sigma'[j]$ violates either the distortion constraint in Case 1 or the differential constraint in Case 2, which contradicts the assumption that σ' is a feasible assignment. □

THEOREM 1. *Algorithm 1 yields an optimal solution for the OLED image scaling optimization problem.*

PROOF. The theorem follows directly from Lemmas 1, 3, and 4. □

3.3. System Implementation

Using the CURA framework, we implemented two practical applications on a Samsung Galaxy Tab 7.7, equipped with a 1.4 GHz dual-core processor and a 1280×800 AMOLED display. The related hardware and software specifications are detailed in Table I. In this section, we describe our system implementation and discuss some technical issues associated with it.

Table I. Specifications of Samsung Galaxy Tab 7.7

Hardware	
CPU	1.4 GHz Dual-Core Exynos 4210
Memory	1 GB LPDDR2
Screen	1280x800 Super AMOLED Plus
Storage	16GB SD 2.0 compatible
Battery	5100 mAh
Software	
OS	Android 4.0.4 Linux Kernel 3.0.15

3.3.1. An Image Converter. First, we exploited CURA as an image converter to generate power-saving versions of images for mobile OLED displays. Figure 3 shows the flow chart of the image converter, which was developed as a stand-alone Android app. It provides a friendly GUI for the user to adjust the three tunable parameters, with SSIM, N, and d set respectively at 0.94, 5, and 2% by default. The converter is comprised of four components, each of which implements the approach or algorithm presented earlier. An image is first segmented into N attention regions with an adjacent matrix A based on its saliency map, as described in Section 3.1.1. Region segmentation for a 1280×800 image takes approximately 3.6 seconds[1] on the Samsung tablet. Then, for each region, a critical scaling ratio $c()$ is determined based on the designated SSIM requirement (Section 3.1.2), and a differential constraint d is assigned (Section 3.1.3). Critical ratio determination requires approximately 4.2 seconds if the image is segmented into five regions. After the above analysis, the image is transformed into several intermediate images associated with the parameters. Next, the proposed algorithm computes a scaling assignment for the image based on the parameters, as explained in Section 3.2.2. Our algorithm costs less than 1 millisecond when $N = 5$. The scaling assignment is represented by a *scaling map* that indicates each pixel's scaling ratio in the range

[1]A step's running time was the difference between its start and end times acquired via the `gettimeofday()` function provided in Linux.

0 and 1. Finally, the image converter generates a power-saving version by multiplying the original image by the scaling map in a pixel-by-pixel manner. Applying image pixel scaling to the image takes only 0.2 seconds. Overall, it requires approximately 8 seconds to process a 1280×800 image with five attention regions on the tablet. Note that the power-saving image minimizes the amount of OLED power required while retaining the desired visual quality, as shown in Section 3.2.3; and this property holds for any OLED display. In other words, a power-saving image once generated can be displayed on many OLED mobile devices.

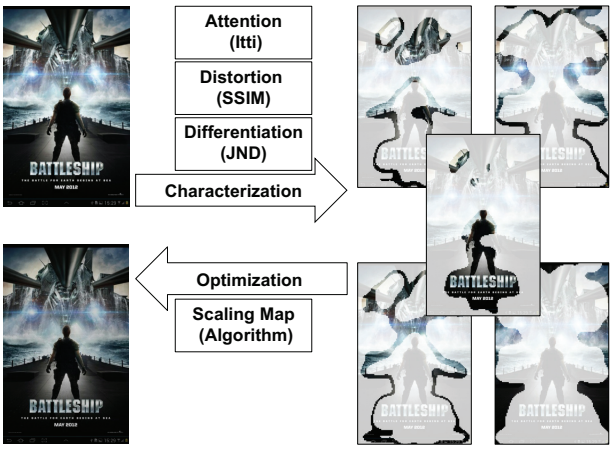


Fig. 3. The flow chart of the image converter

3.3.2. A Power-Saving Mode. In the second scenario, we implemented a screen power-saving mode in the Android mobile system. Like the design of existing power-saving modes, a simple GUI is provided so that the user can easily switch the power-saving mode on or off. To this end, we use `DEVICE_ATTR()`, a macro provided in Linux, to define and implement a new device attribute which maintains an on/off flag in the display driver[2], and use function `device_create_file()` to add the attribute as a file in Linux's file system[3]. Accordingly, our GUI can set or reset the flag in the kernel space from the user space. When the flag is on, the scene currently displayed on the screen will be processed and replaced with a power-saving version in a trade-off of the visual quality for power saving. Figure 4 shows the system architecture, which comprises CURA as well as an *image fetcher* in the user space and a *scene updater* in the kernel space. In Android, each scene to be displayed on the screen has to be written into a frame buffer in the kernel space. When any user interaction, like touching or swiping the screen, changes the current scene, the image fetcher will be triggered to copy the scene from the frame buffer into its own buffer in the user space via system calls provided in Linux. Basically, it utilizes `open()` to open the display device[4], `ioctl()` to obtain the necessary information about the device, and `read()` to read out the scene. Then, the scene is passed to CURA and processed to derive a scaling map, in the same

[2]We implemented our device attribute along with a number of device attributes provided by Samsung for the display model in source directory "`drivers/video/samsung/S3cfb_main.c`".
[3]The file was created under directory "`/sys/devices/platform/samsung-pd.2/s3cfb.0/`"
[4]The display device is mapped to "`/dev/graphics/fb0`" on the Samsung tablet.

way as that described in the first application scenario. Next, the scaling map needs to be copied from CURA to the buffer of the scene updater. Because Linux dose not provide the corresponding system calls, we create a tunnel between the kernel and the user space by defining and implementing, in the display driver[5], a new `ioctl` call that utilizes function `copy_from_user()` provided in Linux to copy data from the user space to the kernel space. The `ioctl` call allows CURA to copy the scaling map to the scene updater's buffer. Finally, the scene stored in the frame buffer is updated by the scene updater based on the scaling map and displayed on the screen.

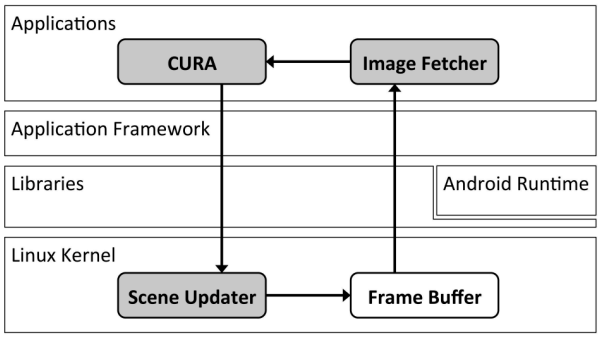


Fig. 4. The system architecture of the power-saving mode

Decisions on when to trigger the above procedure arise with this scenario. The Samsung tablet adopts a *double buffering* mechanism, one frame buffer for the scene being displayed on the screen and the other for the scene in preparation. Once the next scene is ready for display, a function named `s3cfb_pan_display()`, implemented by Samsung in the display driver, will be called to switch the frame buffer. Accordingly, we implement image pixel scaling in this function so that the scene once changed will be replaced with its power-saving version. At first, we decided to have the whole procedure of processing a scene (including fetching the scene, generating its scaling map, and updating the scene) triggered whenever the function was called. However, we observed that the whole procedure was triggered frequently because even a slight change to the scene, like signal strength or clock updating in the status bar, will trigger the function. To avoid unnecessary computations, a slight change will not trigger the whole procedure in our implementation; instead, the scaling map derived based on a previous scene is simply applied to the new scene whenever the function is called, unless the user touches the screen and then the scene remains unchanged for one second. To this end, we capture screen touching by waiting events originated from the corresponding input device[6], and reset a one-second timer once a touching event is captured. When the timer expires, the whole procedure is triggered to derive a scaling map based on the new scene, and the derived map will be applied to the subsequent scenes until the user touches the screen again. Moreover, the above procedure will be terminated and restarted if another interaction event changes the scene before the procedure is completed. Inevitably, the computation is in vain if the user always changes

[5]We implement our `ioctl` call in function `s3cfb_ioctl()`, which contains a number of ioctl calls provided by Samsung, in source directory "`drivers/video/samsung/S3cfb_ops.c`".
[6]The screen's input streams are directed to "`/dev/input/event2`" on the Samsung tablet.

the scene immediately after the replacement. However, whether the OLED energy reduced by CURA can compensate for the computation overhead incurred by CURA depends highly on user behavior. In our implementation, the whole procedure is triggered only when the scene remains unchanged for one second, because a mobile user usually swipes the screen quickly to look for some application and then takes a look at the scene for a while. Our prototype may not fit well with all user behavior but serves as a proof-of-concept implementation of CURA.

Table II. Average running times (in seconds) required for major steps

Step	Scenario 1	Scenario 2
Memory Copy	-	0.02
Image Resizing	-	0.05
Region Segmentation	3.6	0.29
Critical Ratio Determination	4.2	0.17
Algorithm 1	< 0.001	< 0.001
Image Pixel Scaling	0.2	0.2
Total	8	0.73

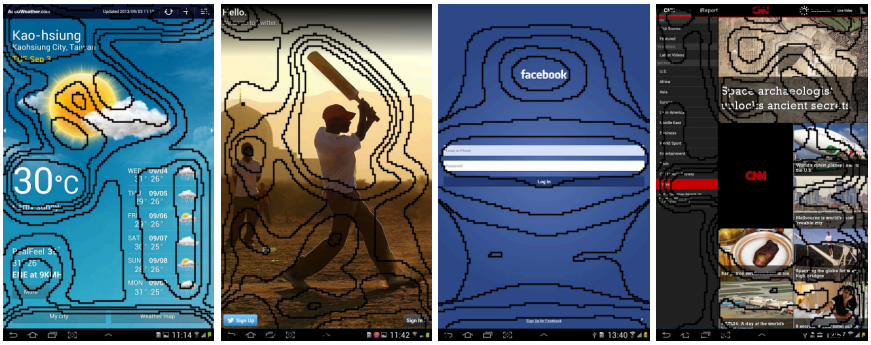
A major difference between the two application scenarios is that the computational overhead should be acceptable in the second scenario. Because both Itti and SSIM involve pixel-level image processing, the native design of CURA requires a few seconds to process a 1280×800 image on the Samsung tablet. This is unacceptable for an online scenario that requires timely responses. Thus, after the image fetcher copies a scene from the image buffer into its own buffer in the user space, *Lanczos resampling* [Lanczos 1970] is used to scale down the resolution of the 1280×800 scene to 320×200 for speed-up. Then, the scaling map, determined by CURA based on the scaled resolution, is copied to the scene updater's buffer in the kernel space. Note that the same scaling ratio is applied to each grid of 4×4 pixels of the original scene (not the downscaled scene) in the frame buffer. In consequence, when CURA incorporates image resizing, it can process a scene in a few hundred milliseconds while preserving the original resolution in the power-saving version. As shown in Table II, after the scene is downscaled by $\frac{1}{4\times4}$ in 0.05 seconds, it takes only 0.29 seconds, 0.17 seconds, less than 1 millisecond, and 0.2 seconds for region segmentation, critical ratio determination, our algorithm, and image pixel scaling, respectively. Moreover, unlike the first scenario, the second scenario requires approximately 0.02 seconds for two extra memory copies between the kernel and user spaces. Overall, it takes approximately 0.73 seconds (and thus consumes only little energy) to process a 1280×800 scene with five regions on the tablet. Image resizing reduces the computational overhead significantly; however, the OLED power consumption is slightly higher than that in the first scenario. We evaluate the trade-off in the experiments described in Section 4.

Our CURA implementation could be further speeded up by carrying out some routine computations in parallel. For example, the step of image pixel scaling, which multiplies an image by a scaling map in a pixel-by-pixel manner, is highly parallelizable. Thus, we can divide the image equally into several sub-images and create multiple *threads* (via `pthread` APIs [Nichols et al. 1996] in the user space or `kthread` APIs [Love 2010] in the kernel space) to handle the sub-images independently. The threads will then be distributed by the CPU scheduler over the available CPU cores for parallel execution. CURA could also benefit from hardware accelerators to speed up image pixel scaling, e.g., by exempting the computations from the CPU to the GPU on the graphics card. On Android (version 4.2 or later), we can implement a function to scale down a pixel's value with `RenderScript` APIs [Guihot 2012] and load the image into an *allocation* with a pixel specified as a basic *element*. Then, the RenderScript runtime will

automatically parallelize the computations, with one execution of the function per element in the allocation, across the available GPU cores. Note that the RenderScript framework currently supports user-space applications only.

4. PERFORMANCE EVALUATION

4.1. Experiment Setup



(a) Weather (b) Twitter (c) Facebook (d) CNN

$$\begin{array}{c} \\ r_1 \\ r_2 \\ r_3 \\ r_4 \\ r_5 \end{array}\!\!\begin{array}{c} \begin{array}{ccccc} r_1 & r_2 & r_3 & r_4 & r_5 \end{array} \\ \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \end{array}$$

(e) Weather

$$\begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

(f) Twitter

$$\begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

(g) Facebook

$$\begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

(h) CNN

Fig. 5. Investigated images with attention regions and adjacent matrixes

Table III. Critical and optimal scaling ratios of investigated images under default parameters

	Weather		Twitter		Facebook		CNN	
	Critical	Optimal	Critical	Optimal	Critical	Optimal	Critical	Optimal
r_1	0.87	0.87	0.88	0.88	0.86	0.86	0.87	0.87
r_2	0.82	0.85	0.82	0.86	0.81	0.85	0.83	0.85
r_3	0.78	0.84	0.78	0.84	0.77	0.83	0.79	0.84
r_4	0.75	0.82	0.75	0.83	0.74	0.82	0.76	0.82
r_5	0.73	0.80	0.73	0.81	0.71	0.80	0.74	0.80

We conducted extensive experiments on the Samsung Galaxy Tab 7.7 with four snapshot images of popular mobile applications (namely Weather, Twitter, Facebook, and CNN), as shown in Figure 5. The images have different characteristics in terms of the average *luminance* and the average saliency value. In the CIELAB space, luminance is a single primary of the pixel value, and a pixel's luminance can be derived by converting its RGB values to its grayscale representation between 0 and 255. The average luminance values of Weather and Twitter (120 and 112 respectively) are higher than those of Facebook and CNN (87 and 79 respectively). On the other hand, a pixel's saliency value, in the range 0 to 1, quantifies its conspicuousness in the image. The average saliency values of Twitter and CNN (0.31 and 0.29 respectively) are higher than those of Weather and Facebook (0.21 and 0.17 respectively). Note that a larger luminance value implies higher power consumption on OLED displays, while a larger

saliency value implies there are more pixels that attract the user's attention. Based on our measurements[7], Weather, Twitter, Facebook, and CNN require 990, 805, 591, and 474 mW of OLED power respectively on the Samsung tablet, and the pixels whose saliency values are above 0.5 account for 11%, 18%, 9%, and 15% respectively of all the pixels in each image. In addition, the images have an identical adjacent matrix, in which r_i is only adjacent to r_{i-1} and/or r_{i+1}. Such a matrix forms when two adjacent pixels always locate in either the same region or two regions with adjacent attention levels. However, as the number of regions increases, a region can be adjacent to more than two regions if the difference between the saliency values of some adjacent pixels is larger than the saliency subrange designated for each region. Table III shows the critical scaling ratios of each investigated image, as well as the image's optimal scaling ratios that satisfy the distortion and differential constraints, when the three tunable parameters were set at their default values.

Table IV. Visual quality of different SSIM ranges [Bartolini et al. 2009b]

SSIM Range	Visual Quality
$1 - 0.98$	High
$0.98 - 0.96$	Medium
$0.96 - 0.94$	Low
≤ 0.94	Unacceptable

To better understand the properties of, and gain insights into, our framework, we conducted three sets of experiments to evaluate CURA from different perspectives. First, we investigated the impact of the three tunable parameters: 1) the impact of the SSIM requirement in the range 1 to 0.88; 2) the impact of the number of regions, N, in the range 1 to 9; and 3) the impact of the differential constant, d, in the range 0% to 4%. Unless otherwise stated, the default values of the SSIM requirement, the number of regions N, and the differential constant d were set at 0.94, 5, and 2% respectively. The settings were based on the following observations. The power consumption was saturated when an image was segmented into five or more regions; thus, we set N at 5 to allow a trade-off between the computational overhead and the visual quality. In [Bartolini et al. 2009b], the impact of various SSIM scores on the visual similarity of original and distorted images was assessed in a side by side comparison by human users. Table IV shows four SSIM ranges and the corresponding visual quality. Accordingly, the SSIM requirement was set at 0.94 so that the region that received the least attention still had acceptable visual quality. We also observed that the region boundaries were difficult to discern when d was less than 2%.

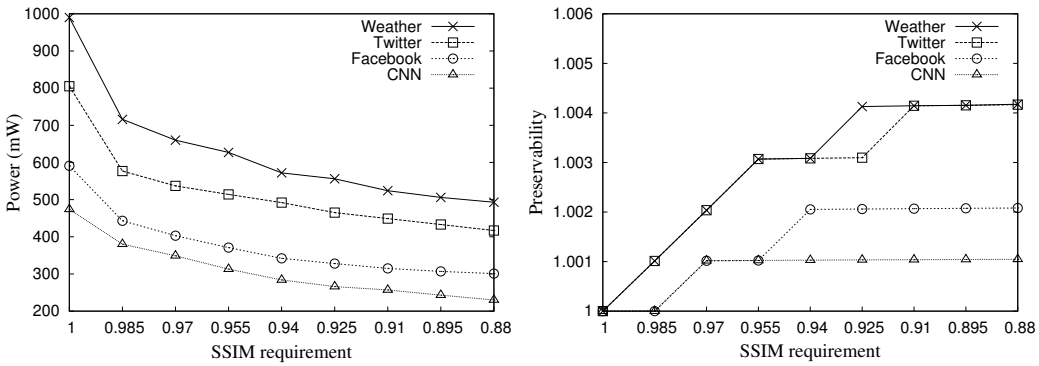
CURA trades off image quality in low-attention regions for power saving. Thus, the OLED power consumption was adopted as an essential performance metric. Besides, to validate whether higher quality is preserved for high-saliency pixels than for low-saliency pixels, we used a metric called *attention-oriented preservability*. The preservability was defined as the ratio of an image's Itti-weighted SSIM score to its SSIM score. Actually, the SSIM score given for an image is the mean value of a matrix with an SSIM value for each pixel, and the Itti-weighted SSIM score is the mean value of the resultant SSIM matrix with each pixel's SSIM value weighted by the pixel's saliency value. Generally, higher preservability implies a larger difference between the SSIM values of high-saliency and those of low-saliency pixels. In the second set

[7]The tablet's transient power (mW) is acquired by multiplying two battery parameters, namely `voltage_now` (mV) and `current_now` (mA), which are updated every 30 seconds in "`/sys/class/power_supply/battery/`". Then, the OLED power required for an image is derived by subtracting the tablet power when a black image (that consumes nearly zero OLED power) is shown on the tablet from the power when the image is displayed.

of experiments, we investigated the impact of image resizing under various resizing ratios. Lanczos resampling was used to scale down every grid of 2×2, 3×3, and 4×4 pixels to one pixel in each case for speed-up. Thus, the entire execution time taken by CURA to process an image was considered an additional performance metric. The results of the two experiment sets are reported in Sections 4.2 and 4.3, respectively.

Finally, we compared CURA with GRID modified based on the grid-based approach proposed in [Chen et al. 2012]. In GRID, an image is partitioned into multiple rectangles of the same size. Then, each rectangular region is given an initial threshold between 0 and 255 based on its pixel values and a predefined *sacrificed luminance ratio*. Any pixel value that exceeds the threshold is simply truncated. Next, GRID computes the SSIM score of every rectangular region and the score of the whole image. If the image's SSIM score is higher (resp. lower) than the SSIM requirement, the threshold of the region with the highest (resp. lowest) SSIM score is decreased (resp. increased) by 1. The process is repeated until the image's SSIM score is the same as the SSIM requirement. Following [Chen et al. 2012], we set the sacrificed luminance ratio at 0.1. To ensure a fair comparison, the number of rectangles was set at 5, and the SSIM requirement was set at 0.94. Note that GRID does not define a differential constraint to smooth the sharp edges between adjacent rectangles. We measured the entire execution time taken to process an image, as well as the OLED power required for and the SSIM scores achieved by the processed image, on the Samsung tablet. The results are reported in Section 4.4.

4.2. The Impacts of the Tunable Parameters



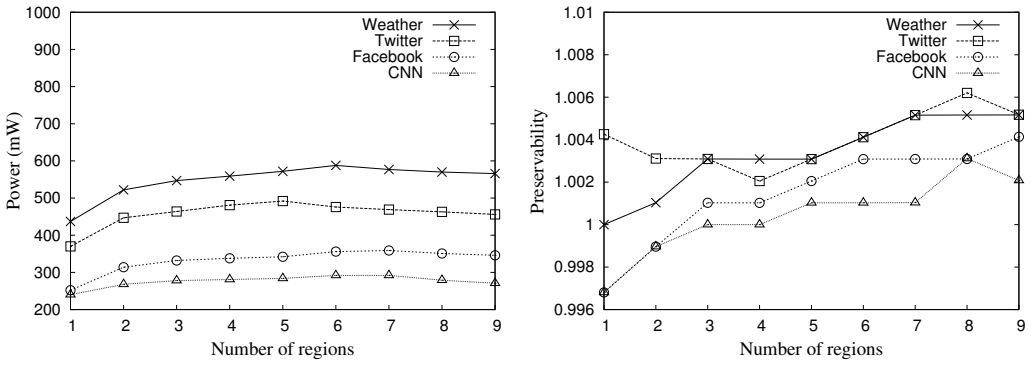
(a) OLED power consumption (b) Attention-oriented preservability

Fig. 6. The impact of the SSIM requirement

Figure 6(a) shows the impact of the SSIM requirement on the OLED power consumption of each image processed by CURA. As expected, the power consumption declines as the SSIM requirement decreases. The reason is that a lower SSIM requirement leads to lower critical scaling ratios. This in turn implies reduced visual quality, but lower power consumption. We observe that the power consumption drops abruptly when the SSIM requirement decreases from 1 to 0.985. This is because the power consumed by an OLED pixel drops dramatically as its value decreases, as shown in Figure 1. Thus, OLED image scaling (if applied appropriately) can achieve a significant reduction in power consumption at a cost of a slight decline in the visual quality. In addition, the power consumption is higher when an image's luminance is larger; for example, the luminance of the Weather image is the largest among the four images, as mentioned

previously. The results show that when the SSIM requirement was set at 0.94, CURA can achieve power savings between 38% and 42%, depending on the image's characteristics.

Figure 6(b) shows the impact of the SSIM requirement on the attention-oriented preservability in each image processed by CURA. The preservability is always not smaller than 1 for all the images under various SSIM requirements. Because high-saliency pixels are given larger weights than low-saliency pixels, the result implies that the SSIM values of high-saliency pixels are generally larger than the average SSIM value. In other words, the visual quality of high-saliency pixels are preserved in general; meanwhile, the power reduction, as shown in Figure 6(a), mainly comes from the reduced quality of low-saliency pixels. When the SSIM requirement was set at 1, the preservability is 1 because the SSIM values given for all pixels are 1 when the processed and original images are identical. Then, the preservability increase as the SSIM requirement decreases. The reason is that a lower SSIM requirement usually leads to a larger difference in the SSIM values between high-saliency and low-saliency pixels. The preservability in an image is saturated when the difference between the scaling ratios applied to (and thus the SSIM values of) high-saliency and how-saliency pixels is limited by the differential constant. Note that the preservability is usually small because high-saliency pixels typically account for only a small proportion of all the pixels in an image, although their SSIM values can be notably larger than those of saliency pixels.



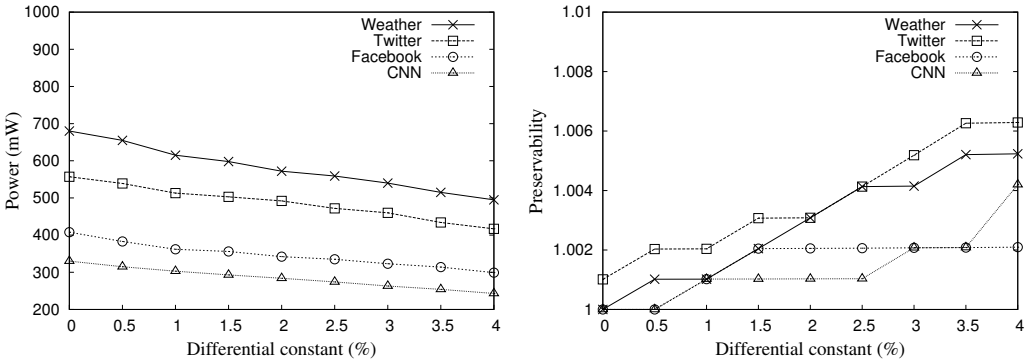
(a) OLED power consumption (b) Attention-oriented preservability

Fig. 7. The impact of the number of regions (N)

Figure 7(a) shows the impact of the number of regions on the OLED power consumption achieved by CURA. The power consumption increases as the number of regions increases from 1 to 5. This is because the average of the SSIM requirements assigned to different regions increases with the number of regions, which implies higher critical scaling ratios. Moreover, in general, the higher the number of regions, the denser will be the adjacent matrix. As a result, the scaling ratios determined by CURA are relatively high. However, the power consumption is saturated when $N = 5$. The reason is that the increments in the average SSIM requirement and the density of the adjacent matrix are reduced and become negligible when $N = 5$. Therefore, in our implementation, we set N at 5 to allow a trade-off between the visual quality and the computational overhead. Note that when the number of regions increases by 1, the number of visual quality assessments (i.e., the number of computed SSIM scores) increases by 8. The processing time of the proposed algorithm also increases slightly.

We observe that the power consumption declines slightly when an image is segmented into 9 regions. The cause of this interesting phenomenon is that the influence of the differential constraint on adjacent regions decreases as the number of regions increases. Consequently, a lower scaling ratio could be applied to the region that receives the least attention, which usually accounts for a large proportion of an image.

Figure 7(b) shows the impact of the number of regions on the attention-oriented preservability achieved by CURA. Interestingly, the preservability in some images, like Facebook and CNN, is smaller than 1 when $N = 1$ or 2. This phenomenon may happen when a segmented region is large, because the scaling ratio depends on not only those high-saliency pixels but also other low-saliency pixels in the same region. When $N = 1$, for example, all the pixels of an image are treated equally and applied with the same scaling ratio; consequently, whether the resultant SSIM values of high-saliency pixels will be relatively large or small is "nondeterministic". This result also provides evidence for the efficacy of attention region segmentation. As the number of regions increases, the preservability increases in general but decreases sometimes. The reason is that after the scaling ratios are redetermined and reapplied in consequence of resegmenting the image into more regions, the SSIM values of some pixels may become closer or farther to the average SSIM value. Thus, the preservability may increase or decrease, depending on whether the SSIM values of most pixels become farther or closer to the average.



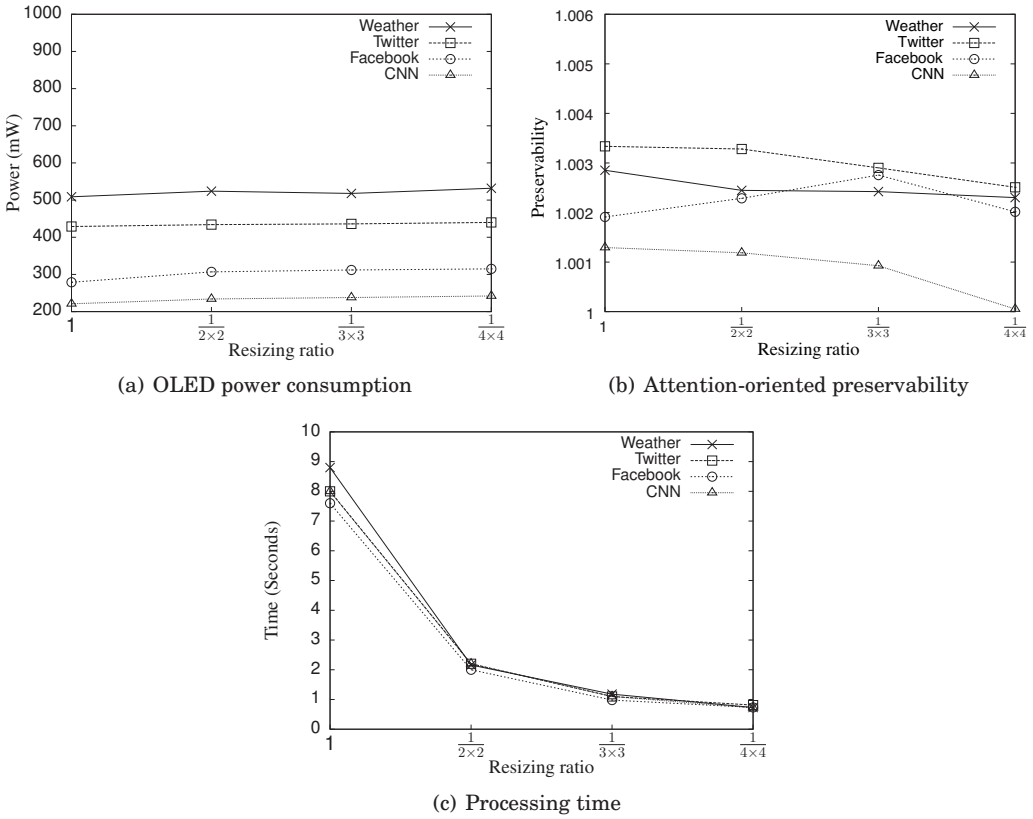
(a) OLED power consumption (b) Attention-oriented preservability

Fig. 8. The impact of the differential constant (d)

Figure 8(a) shows the impact of the differential constant on the OLED power consumption under CURA. The power consumption decreases as the differential constant increases. This result is as expected because a higher differential constant allows the pixel values of a region to be scaled down by a higher magnitude. Thus, the power consumption of a region can be reduced until the scaling ratio determined by CURA reaches its critical scaling ratio. The characteristic is especially beneficial for images whose average saliency values are low to save power. As mentioned previously, Weather and Facebook contain more pixels with low saliency values than Twitter and CNN. This explains why the reduction in power consumption is more manifest for Weather and Facebook than for Twitter and CNN. Finally, comparison of the impacts of the three parameters on the power consumption shows that the impact of varying the differential constant is greater than the impact of varying the number of regions. However, neither of the impacts is as significant as the impact of varying the SSIM requirement.

Figure 8(b) shows the impact of the differential constant on the attention- preservability under CURA. Since the preservability is not smaller than 1 for all the cases, the visual quality is better preserved for high-saliency pixels than for the low-saliency pixels, with the same reason explained earlier in Figure 6(b). As the differential constant increases, the preservability never decreases, because a higher differential constant can only lead to a larger difference in the SSIM values between high-saliency and low-saliency pixels. Moreover, the increment of the preservability depends mainly on the difference in the decrements of the SSIM values between high-saliency and low-saliency pixels. Finally, comparison of the impacts of the three parameters on the preservability shows that the impacts of varying the SSIM requirement and varying the differential constant are similar, but different from the impact of varying the number of regions.

4.3. The Impact of Image Resizing



(a) OLED power consumption

(b) Attention-oriented preservability

(c) Processing time

Fig. 9. The impact of image resizing

Figure 9(a) shows the impact of image resizing on the OLED power consumption under CURA. Clearly, image resizing does not have a significant impact in this case. In general, an image's attention regions will remain similar after resizing. Moreover, a grid of pixels is usually scaled down to one representative pixel whose value is close to the average value of the pixels in the grid. Thus, a region's critical scaling ratio will be similar to that of the corresponding region in the original image. This explains

why CURA derives similar scaling assignments for different-sized images with the same content. However, the power consumption still increases slightly as the resizing ratio decreases. The reason is that, in terms of image similarity, a change is usually more significant in a low-resolution image than in its original version. Thus, to achieve the same visual quality (i.e., the same SSIM requirement), a region's critical scaling ratio derived based on the low-resolution image will be slightly larger than that of its counterpart in the original image. The results show that, with image resizing, the power saving achieved by CURA based on the original image can may increased by 2% to 6%, depending on the image's characteristics.

Figure 9(b) shows the impact of image resizing on the attention-oriented preservability under CURA. As the resizing ratio decreases, the preservability is still kept above 1, so the power reduction shown in Figure 9(a) remains mainly from the reduced quality of low-saliency pixels. It is because image resizing may change pixels' saliency values but generally not change the relative levels of attention they receive in a saliency map. We observe that the preservability usually decreases but sometimes increases, depending on the respective increments of the scaling ratios applied to different regions of the original image. Because the scaling ratios usually increase (as the resizing ratio decreases) more largely in low-attention regions than in high-attention regions, the preservability decreases in most cases; yet, the overall visual quality of the power-saving image usually increases as a consequence.

Figure 9(c) shows the impact of image resizing on the processing time required by CURA. As expected, the processing time decreases with the resizing ratio. This is because the techniques we use for image segmentation and distortion assessment (i.e., Itti and SSIM) involve analyzing a large number of image pixels; consequently, the processing time is proportional to the size of the image. We observe that the decrease is more dramatic when the resizing ratio is reduced from 1 to $\frac{1}{2\times2}$ than when the ratio is increased from $\frac{1}{2\times2}$ to $\frac{1}{4\times4}$. This phenomenon occurs because the original image's size is reduced by as much as $\frac{3}{4}$ when the resizing ratio was set at $\frac{1}{2\times2}$, but it is further reduced by only $\frac{3}{16}$ when the ratio was set at $\frac{1}{4\times4}$. Interestingly, the processing times required for the four images (which are the same size but have different characteristics) are almost the same. This finding implies that CURA's processing time is independent of the image's content and very stable. The results show that, with image resizing, the time required to process an image can be reduced to only seven or eight hundred milliseconds. Note that the processing time includes the time required for Lanczos resampling if it is employed to scale down the resolution. Thus, CURA's computational overhead and energy cost are acceptable on mobile devices.

4.4. Comparison with the State of the Art

Table V. OLED power consumption (mW) of GRID and CURA

Scenario 1	Weather	Twitter	Facebook	CNN
GRID	648	340	237	343
CURA	572	492	342	284
Scenario 2	Weather	Twitter	Facebook	CNN
GRID	797	451	362	357
CURA	595	503	378	305

Table V shows the OLED power consumption of GRID and CURA in the two scenarios, i.e., the image converter and the power-saving mode. In both scenarios, CURA requires less power for Weather and CNN, while GRID requires less power for Twitter

and Facebook. The reasons are as follows. CURA scales down pixel values based on human attention, so it usually saves more power when the low-attention regions contain a large number of bright pixels. In contrast, GRID treats all pixels equally and may save more power by darkening bright pixels in the high-attention regions. Note that a brighter pixel implies higher power consumption on OLED displays. However, we observe that the difference between the power consumption of the two scenarios is more significant under GRID than under CURA. This phenomenon implies that GRID is more sensitive to image resizing. The reason is that, to maintain the required visual quality, the pixel values scaled down in Scenario 1 have to be increased slightly in Scenario 2 because the latter is based on low-resolution versions. CURA distributes the increment over all the pixels, while GRID only increases the pixel values that exceed the threshold. As a result, the power consumption of GRID in Scenario 2 increases significantly due to the characteristics of the OLED power model.

Table VI. Processing times (in seconds) required by GRID and CURA

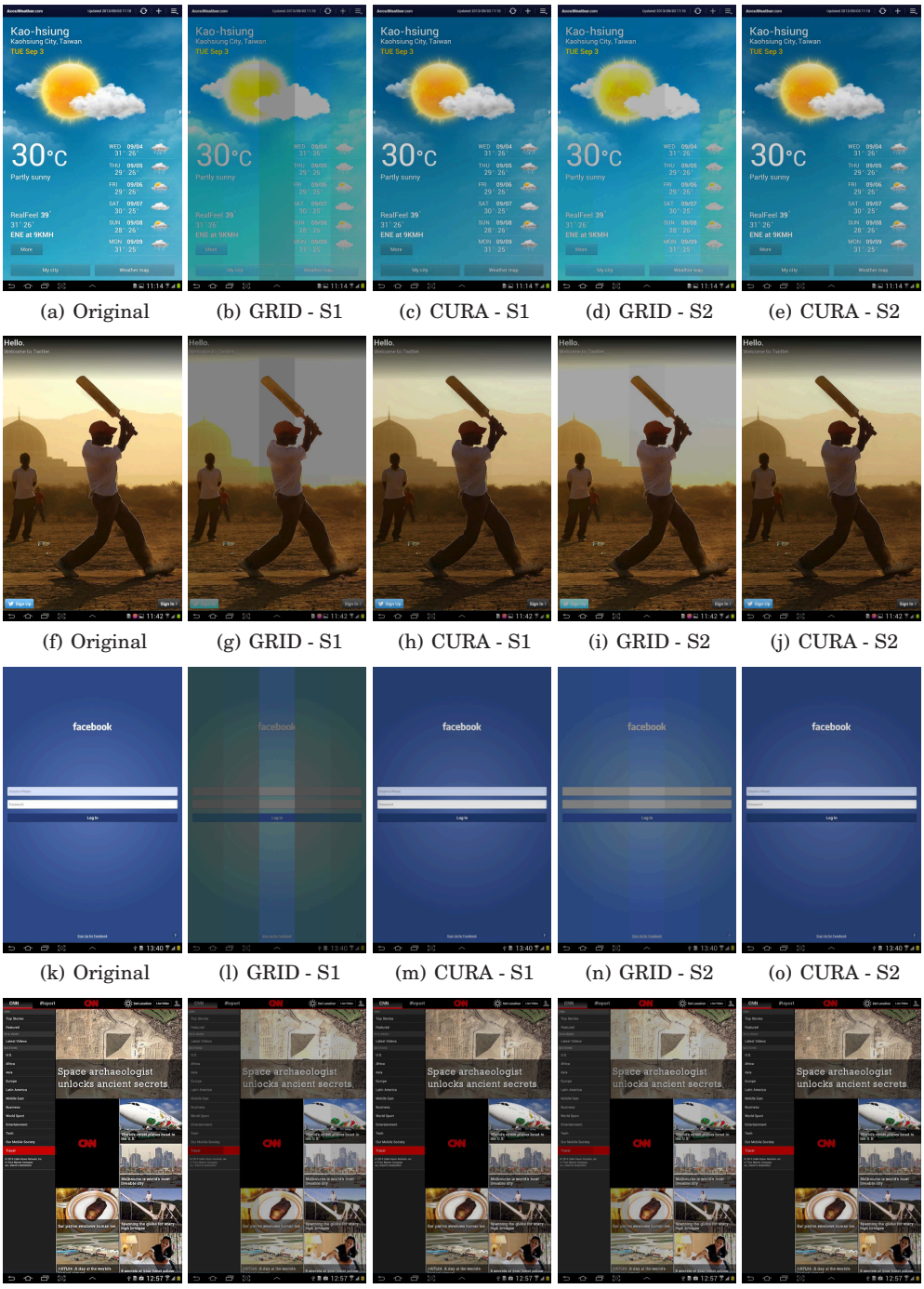
Scenario 1	Weather	Twitter	Facebook	CNN
GRID	28	219	193	27
CURA	8.8	8	7.6	8
Scenario 2	Weather	Twitter	Facebook	CNN
GRID	4.2	4.7	1	3.4
CURA	0.7	0.8	0.7	0.7

Table VI shows the times required by GRID and CURA under the two scenarios. CURA outperforms GRID in terms of the processing time in all cases because GRID has to compute a much larger number of SSIM scores. With image resizing, the times required by GRID and CURA, especially GRID, become much shorter under Scenario 2, compared with those under Scenario 1. This is because the SSIM computation dominates GRID's processing time and, consequently, image resizing greatly benefits GRID. In addition to Itti and SSIM computations, CURA relies on Algorithm 1 to determine an optimal scaling assignment. The algorithm's time complexity is not dependent on the image size, but the number of segmented regions. To use CURA in an online application like Scenario 2, the algorithm's computational overhead is relevant and should be justifiable. We observe that our algorithm only accounts for a very small portion of the processing time when N is small. In other words, CURA's processing time could be further reduced if image segmentation and distortion assessment can be done more efficiently. Moreover, the time required by CURA is very stable in the same scenario, but that of GRID varies significantly. This is because GRID's processing time is highly dependent on accurate prediction of the initial threshold.

Table VII. SSIM scores achieved by GRID and CURA

Scenario 1	Weather	Twitter	Facebook	CNN
GRID	0.94	0.94	0.94	0.94
CURA	0.973	0.973	0.974	0.972
Scenario 2	Weather	Twitter	Facebook	CNN
GRID	0.969	0.973	0.972	0.945
CURA	0.976	0.976	0.983	0.976

Table VII shows the SSIM scores achieved by GRID and CURA under the two scenarios. In both scenarios, CURA achieves higher SSIM scores than GRID does for all the images. This is mainly because GRID attempts to reduce OLED power as much as possible, while just satisfying the designated SSIM requirement. In contrast, CURA applies the designated SSIM requirement to the lowest-attention region, but it increases



(a) Original (b) GRID - S1 (c) CURA - S1 (d) GRID - S2 (e) CURA - S2

(f) Original (g) GRID - S1 (h) CURA - S1 (i) GRID - S2 (j) CURA - S2

(k) Original (l) GRID - S1 (m) CURA - S1 (n) GRID - S2 (o) CURA - S2

(p) Original (q) GRID - S1 (r) CURA - S1 (s) GRID - S2 (t) CURA - S2

Fig. 10. Images processed by GRID and CURA when the SSIM requirement was set at 0.94

the SSIM requirements of the other regions to different degrees based on their attention levels. Of course, GRID can also increase the SSIM requirement to improve the quality of the produced images, but the resultant images will consume more power as a consequence. Interestingly, a higher SSIM score does not necessarily imply less power saving, as shown by comparison of the Weather (or CNN) images processed by CURA and GRID. The result demonstrates that human visual attention exploited by CURA is very effective in reducing the power consumption while retaining the visual quality of images. Moreover, although both GRID and CURA achieve higher SSIM scores under scenario 2 than under scenario 1, the increase in SSIM scores is much more obvious for the images processed by GRID than those processed by CURA. This is because GRID is more sensitive to image resizing. With image resizing, the power consumption of GRID in Scenarios 2 increases significantly, as explained earlier for the results in Table V. Generally, the increased power consumption will be of benefit to the image quality, which is reflected on the SSIM scores.

Figure 10 shows the four original images and those processed by GRID and CURA when the SSIM requirement was set at 0.94. We observe that the images processed by GRID often contain chrominance changes and sharp edges, but such phenomena do not occur in the images processed by CURA. It appears that the human eye can easily notice these problems in a plain image, such as Figure 10(l). This result indicates that GRID, which uses a global SSIM requirement to limit the distortion of the whole image, could not avoid sharp edges when the requirement is set too low. The reason is that the edges between regions only occupy a small number of image pixels, so they do not significantly affect the score given by SSIM for image distortion assessment. By contrast, CURA can eliminate the boundary effect even when the SSIM requirement was set at 0.94, because it uses the just noticeable difference to ensure that the region boundaries are too indistinct to be discerned by the human eye. Based on the experiment results in [Chen et al. 2012], the adverse phenomena will not appear when the SSIM requirement is set at 0.98, which indicates a high visual quality. However, CURA can tolerate lower SSIM requirements. For example, sharp edges remain visible in 10(i), but not in 10(j), although they have similar SSIM scores. Moreover, although the images processed by CURA look slightly darker than and somewhat different from the original images (because their average luminance is lowered to reduce power consumption), the incurred image distortion does not impact the visual experience significantly.

5. RELATED WORK

Displays are among the largest power-consuming components on mobile devices [Dong and Zhong 2012]. Much of the early research on low-power display techniques, especially backlight scaling, focused on LCD displays [Chang et al. 2004; lranli and Pedram 2005; Bartolini et al. 2009a; Xiao et al. 2013; Lin et al. 2014a]. However, the latest mobile devices increasingly feature the emerging OLED displays, and the design of low-power techniques for OLED displays has generated more research interest. The organization of OLED and LCD displays is similar, but their power models are very different. Comprehensive measurements of OLED power modeling have helped researchers understand the characteristics of OLED power consumption [Dong and Zhong 2012]. They have also facilitated the empirical design of various low-power display techniques for OLED displays, such as partial display disabling/dimming [Betts-LaCroix 2010], color remapping [Dong et al. 2009], and OLED dynamic voltage scaling [Shin et al. 2011] introduced in Section 1, as well as the image pixel scaling technique proposed in this work.

Based on the above techniques, a number of effective power-saving methodologies and algorithms have been developed for a variety of mobile applications on OLED displays. In particular, motivated by the observation that a user usually focuses on just

half of the screen for most (but not all) interactive applications, Tan et al. [Tan et al. 2013] proposed an approach that exploits the partial dimming technique to dim the screen's top or bottom portions, which contain content that is relatively unimportant to the user. Chen et al. [Chen et al. 2014] proposed to dim the screen areas covered by user fingers when a user interacts with a smartphone via the touch screen. As the attention is often concentrated on the center of the screen for moving scenes, Tan et al. [Tan and Balan 2012] reduced the OLED power required to play mobile games by modeling the locus of attention with a stack of different-sized dimming boxes; accordingly, dimming (from the locus linearly to the edge of the screen) is performed incrementally as the user initiates movements and vice versa as the user stops moving. Meanwhile to make web browsing more power-efficient on OLED displays, Dong and Zhong [Dong and Zhong 2011] developed a color-remapping methodology that renders web pages with power-optimized color schemes on mobile devices. Finally, Chen et al. [Chen et al. 2012] designed an algorithm based on the OLED dynamic voltage scaling technique to reduce the supply voltage while retaining the visual quality of video streaming applications. They also extended and applied the algorithm to online scenarios [Zhao et al. 2013]. The algorithm divides the OLED panel into multiple rectangular regions and optimizes the voltage of each region under a given quality requirement. It is a closely related approach that also tries to retain the image quality. Thus, we modified it to comply with the proposed image pixel scaling technique for comparison, as described in Section 4.

6. CONCLUDING REMARKS

In this paper, we investigate how to reduce the power consumption of OLED displays by exploiting human visual attention. We also present a framework called CURA for quality-retaining power saving on mobile OLED displays. To retain an image's visual quality, CURA segments the image into regions based on the level of visual attention they receive. It also defines scaling constraints to limit the image distortion of each region and avoid sharp edges between adjacent regions. CURA utilizes an algorithm to derive a scaling assignment that optimizes the OLED power required for an image under the defined constraints and applicable to any OLED display. To validate CURA's practicality, we implemented it in two practical applications on a commercial Samsung Galaxy Tab 7.7 tablet, and conducted a number of experiments with some snapshot images of popular mobile apps. The results show that CURA can achieve a significant reduction in power consumption, while maintaining the high visual quality of the regions that dominate the user's attention. The reduction is particularly obvious when an image has a large luminance value and a small saliency value. Moreover, with image resizing, the time required by CURA to process an image is only a few hundred milliseconds, irrespective of the image content, and it is very stable. Because of the above properties, CURA is suitable for many mobile applications.

CURA aims at still images currently. In our future research, we will extend and adapt CURA for video streaming applications by exploiting the structural similarity of consecutive video frames and the human eye's persistence of vision. Moreover, the visual focus may move when the user interacts with the mobile device. It would also be interesting to extend CURA for handling visual movements and study how the power saving is distributed as the visual focus changes.

REFERENCES

ACHARYA, T. AND RAY, A. K. 2005. *Image Processing: Principles and Applications*. John Wiley & Sons, Inc., 38–39.

BARTOLINI, A., RUGGIERO, M., AND BENINI, L. 2009a. HVS-DBS: Human Visual System-Aware Dynamic Luminance Backlight Scaling for Video Streaming Applications. In *Proc. of the ACM EMSOFT*. 21–28.

BARTOLINI, A., RUGGIERO, M., AND BENINI, L. 2009b. Visual Quality Analysis for Dynamic Backlight Scaling in LCD Systems. In *Proc. of IEEE/ACM DATE*. 1428–1433.

BETTS-LACROIX, J. 2010. Selective Dimming of OLED Displays. US Patent 0149223 A1.

BORJI, A. AND ITTI, L. 2013. State-of-the-Art in Visual Attention Modeling. *IEEE Trans. on PAMI 35,* 1, 185–207.

BURROWS, P. E., GU, G., BULOVIC, V., SHEN, Z., FORREST, S. R., AND THOMPSON, M. E. 1997. Achieving Full-color Organic Light-emitting Devices for Lightweight, Flat-panel Displays. *IEEE Trans. on Electron Devices 44,* 8, 1188–1203.

CHANG, N., CHOI, I., AND SHIM, H. 2004. DLS: Dynamic Backlight Luminance Scaling of Liquid Crystal Display. *IEEE Trans. on VLSI Systems 12,* 8, 837–846.

CHEN, X., NIXON, K. W., ZHOU, H., LIU, Y., AND CHEN, Y. 2014. FingerShadow: An OLED Power Optimization based on Smartphone Touch Interactions. In *Proc. of USENIX HotPower*. 6:1–6:5.

CHEN, X., ZHENG, J., CHEN, Y., ZHAO, M., AND XUE, C. J. 2012. Quality-retaining OLED Dynamic Voltage Scaling for Video Streaming Applications on Mobile Devices. In *Proc. of IEEE/ACM DAC*. 1000–1005.

DESIMONE, R. AND DUNCAN, J. 1995. Neural Mechanisms of Selective Visual Attention. *Annual Review of Neuroscience 18,* 1, 193–222.

DONG, M., CHOI, Y. K., AND ZHONG, L. 2009. Power-saving Color Transformation of Mobile Graphical User Interfaces on OLED-based Displays. In *Proc. of IEEE/ACM ISLPED*. 339–342.

DONG, M. AND ZHONG, L. 2011. Chameleon: A Color-adaptive Web Browser for Mobile OLED Displays. In *Proc. of ACM MobiSys*. 85–98.

DONG, M. AND ZHONG, L. 2012. Power Modeling and Optimization for OLED Displays. *IEEE Trans. on Mobile Computing 11,* 9, 1587–1599.

GUIHOT, H. 2012. *Pro Android Apps Performance Optimization*. Apress, Chapter 9: RenderScript, 231–263.

ITTI, L., KOCH, C., AND NIEBUR, E. 1998. A Model of Saliency-based Visual Attention for Rapid Scene Analysis. *IEEE Trans. on PAMI 20,* 11, 1254–1259.

LANCZOS, C. 1970. *The Variational Principles of Mechanics*. Univ. of Toronto Press.

LIN, C.-H., HSIU, P.-C., AND HSIEH, C.-K. 2014a. Dynamic Backlight Scaling Optimization: A Cloud-based Energy-saving Service for Mobile Streaming Applications. *IEEE Trans. on Computers 63,* 2, 335–348.

LIN, C.-H., KANG, C.-K., AND HSIU, P.-C. 2014b. Catch Your Attention: Quality-retaining Power Saving on Mobile OLED Displays. In *Proc. of IEEE/ACM DAC*. 1–6.

LIN, W. AND KUO, C. J. 2011. Perceptual Visual Quality Metrics: A Survey. *Jour. of Visual Communication and Image Representation 22,* 4, 297–312.

LOVE, R. 2010. *Linux Kernel Development* 3rd Ed. Addison-Wesley Professional, Chapter 3: Process Management, 35–36.

LRANLI, A. AND PEDRAM, M. 2005. DTM: Dynamic Tone Mapping for Backlight Scaling. In *Proc. of IEEE/ACM DAC*. 612–616.

NICHOLS, B., BUTTLAR, D., AND FARRELL, J. P. 1996. *PThreads Programming: A POSIX Standard for Better Multiprocessing*. O'Reilly Media, Chapter 1: Why Threads?, 13–15.

SHIN, D., KIM, Y., CHANG, N., AND PEDRAM, M. 2011. Dynamic Voltage Scaling of OLED Displays. In *Proc. of IEEE/ACM DAC*. 53–58.

SHIN, D., KIM, Y., PEDRAM, M., AND CHANG, N. 2013. Dynamic Driver Supply Voltage Scaling for Organic Light Emitting Diode Displays. *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems 32,* 7, 1017–1030.

TAN, K. W. AND BALAN, R. K. 2012. Adaptive Display Power Management for OLED Displays. In *Proc. of ACM MobiGames*. 25–30.

TAN, K. W., OKOSHI, T., MISRA, A., AND BALAN, R. K. 2013. FOCUS: A Usable & Effective Approach to OLED Display Power Management. In *Proc. of ACM UbiComp*. 573–582.

TREISMAN, A. AND GELADE, G. 1980. A Feature Integration Theory of Attention. *Cognitive Psychology 12,* 1, 97–136.

WANG, Z., BOVIK, A. C., SHEIKH, H. R., AND SIMONCELLI, E. P. 2004. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Trans. on Image Processing 13,* 4, 600–612.

WEBER, E. H. 1834. *De Pulsu, Resorpitione, Auditu et Tactu: Annotationes Anatomicae et Physiologicae*.

XIAO, Y., IRICK, K., NARAYANAN, V., SHIN, D., AND CHANG, N. 2013. Saliency Aware Display Power Management. In *Proc. of DATE*. 1203–1208.

ZHAO, M., ZHANG, H., CHEN, X., CHEN, Y., AND XUE, C. J. 2013. Online OLED Dynamic Voltage Scaling for Video Streaming Applications on Mobile Devices. In *Proc. of CODES+ISSS*. 1–10.