Columbia Photographic Images and Photorealistic Computer Graphics Dataset

Tian-Tsong Ng, Shih-Fu Chang, Jessie Hsu, Martin Pepeljugoski*
{ttng,sfchang,yfhsu}@ee.columbia.edu, martinpep@yahoo.com
Department of Electrical Engineering
Columbia University

ADVENT Technical Report #205-2004-5
Feb 2005

Abstract

Passive-blind image authentication is a new area of research. A suitable dataset for experimentation and comparison of new techniques is important for the progress of the new research area. In response to the need for a new dataset, the Columbia Photographic Images and Photorealistic Computer Graphics Dataset is made open for the passive-blind image authentication research community. The dataset is composed of four component image sets, i.e., the *Photorealistic Computer Graphics* Set, the *Personal Photographic Image* Set, the *Google Image* Set, and the *Recaptured Computer Graphics* Set. This dataset, available from `http://www.ee.columbia.edu/trustfoto`, will be for those who work on the photographic images versus photorealistic computer graphics classification problem, which is a subproblem of the passive-blind image authentication research. In this report, we describe the design and the implementation of the dataset. The report will also serve as a user guide for the dataset.

1 Introduction

Digital watermarking [1] has been an active area of research since a decade ago. Various fragile [2, 3, 4, 5] or semi-fragile watermarking algorithms [6, 7, 8, 9] has been proposed for the image content authentication and the detection of image tampering. In addition, authentication signature [10,

*This work was done when Martin spent his summer in our research group

11, 12, 13] has also been proposed as an alternative image authentication technique. Both digital watermarking and authentication signature are considered active image authentication techniques. They respectively requires a known signal to be embedded into an image or the content features to be extracted from an image before the image can be authenticated. Recently, a passive-blind image authentication approach was proposed [14, 15]. The passive-blind image authentication approach does not require any prior knowledge from an image for content authentication and tampering detection. The passive-blind technique would be useful in the situation where the opportunity for embedding an active authentication signal on an image does not present itself, as often the case for various image forensics situations. These techniques are important for application such as criminal investigation, trustworthy journalistic reporting, and intelligence analysis.

We expect to see a plethora of proposed new techniques for tackling the related open issues in the passive-blind image authentic research. To assess the merits of the proposed techniques by various researchers, there should be a way to measure how well the technique has solved its intended problem. In situation where there is a lack of a good mathematical model for the authentication object such as fake images, an empirical model can be realized through a dataset, e.g., an image-splicing detection algorithm can be evaluated on a dataset with spliced images. On the other hand, the performance of a technique evaluated on a specific dataset may be very different from the results obtained using another different dataset, due to the possible bias within the different datasets. This points to the importance of proper dataset design and the need for having a common dataset for a fair comparison of various proposed techniques. The availability of such a common and proper dataset would help to expedite the progress of a thriving research area.

There are a number of image dataset available for various types of image processing research; content-based image retrieval community commonly uses Corel image dataset, digital watermarking community has a dataset put together by Fabien Petitcolas [16], face recognition research has Yale face database [17] and other general image processing research can use USC-SIPI Image Database [18]. The issue is whether we can reuse one of the available image dataset for the research of passive-blind image authentication. Evaluation of the passive-blind image authentication techniques requires fake images as well as authentic images with a reliable origin, and these images, particularly the fake images, are not readily available. Therefore, an effort is needed for collecting suitable datasets for the passive-blind image authentication research. For instance, the problem of image splicing [15] is a new problem in need of a dataset with spliced images. In response to the need, we have earlier released an open dataset, the Columbia Image Splicing Detection Evaluation Dataset [19], for the image splicing problem.

Another problem identified under passive-blind image authentication research is the classification of photographic images (PIM) and photorealistic computer graphics (PRCG) [20, 21]. Working on the PIM and PRCG classification problem requires a dataset containing PRCG of high photorealism, and PIM of reliable sources and with diversity in terms of image content and the image acquisition factors such as the types of camera being used and the photographing styles and techniques. Such dataset is not readily available and we have collected such a dataset, namely the Columbia Photographic Images and Photorealistic Computer Graphics Dataset, during the process of working of the PIM versus PRCG classification problem. We are making this dataset available to the research community. This report describes the design and the implementation of the dataset.

Section 2 will discuss the requirements for a dataset which caters for the PIM versus PRCG classification problem. In Section 3, we give an overview of the Columbia Photographic Images and Photorealistic Computer Graphics Dataset. Then, the subsequent sections are dedicated for the detailed description of the dataset components, i.e., the *PRCG*, the *Personal*, the *Google* and the *Recaptured PRCG* image sets. Then, Section 8 provides a guide for downloading the respective components of the dataset. Finally, we conclude with Section 9.

2 Dataset Requirements for the PIM versus PRCG Classification Problem

While the problem of classifying PIM and the general computer graphics (including both the photorealistic and non-photorealistic computer graphics) has been studied for the purpose of improving video retrieval [22] and other applications [23], the PIM versus PRCG classification problem in the passive-blind image authentication settings is a new problem. It emphasizes on highly photorealistic PRCG rather than normal or non-photorealistic computer graphics, such as the cartoon-like images seen on television. In general, a passive-blind PIM versus PRCG classifier would be evaluated in the following aspects:

1. The discrimination rate/accuracy of the classifier.

2. The robustness of the classifier to various image processing operation

on images, such as JPEG compression, resizing, the various in-camera image processing operations for PIM, and so on.

3. The robustness of the classifier to various computer graphics techniques such as the simulated camera depth-of-field (DoF) effects, soft shadow and so on.

4. The robustness of the classifier to various adversarial attacks. When the algorithm of a classifier is known, the attacker may be able to pre-process a PRCG such that it is classified as a photographic image.

5. The sensitivity of the classifier to image content, in particular for those ambiguous content such as that of the recaptured PRCG or paintings, PRCG of natural scene, PIM of artificial objects and so on.

Apart from facilitating the evaluation of the PIM versus PRCG classifier according to the above-listed aspects, a good dataset for the PIM and PRCG classification problem in the passive-blind image authentication settings should also model the authentic and the fake images well. Hence, we have to ensure the reliable authenticity of the PIM besides that the PRCG are from reliable sources and are of high photorealism.

The concern of high photorealism of PRCG is due to the fact that only PRCG of high photorealism will be used to fake PIM in realistic situation. Unfortunately, PRCG of high photorealism are not readily available in abundance in the Internet. There are many computer graphics in Internet but many of them are not truly photorealistic, so a conscious effort is needed to select only PRCG with high photorealism.

Besides that, we also need to make sure that the content of the PRCG is comparable to that of the PIM. The concern of content compatibility between PIM and PRCG is to ensure that we are comparing apple to apple. Otherwise, a trained classifier may overfit to the content discrepancy between the two image sets, for example, this can happen if the dataset contains mainly PIM of buildings and PRCG of forest. There are two ways to ensure the matching of the content. First way is to narrowly restrict the image content in both the PIM and the PRCG sets, e.g., we can restrict the dataset to have only images of vegetation. The second way is to define a broader scope for the content but ensure the content diversity within the scope, in order to lower the likelihood of content mismatch. In our case, we follow the second way; we define the content scope to be natural scene and ensure the content diversity within the defined scope.

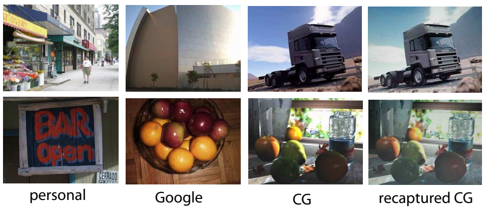
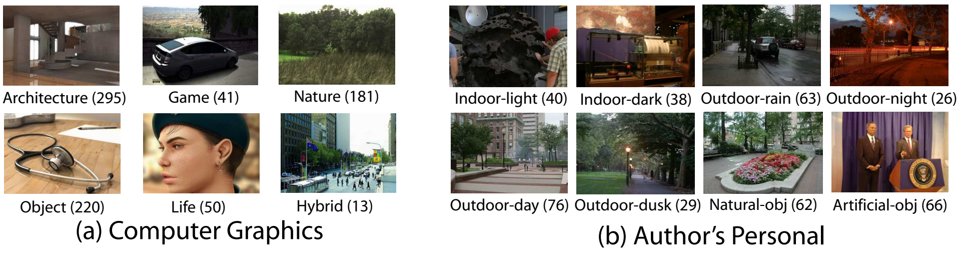


Figure 1: Examples from our image sets. Note the photorealism of all images.



(a) Computer Graphics

(b) Author's Personal

Figure 2: (a) Subcategories within the *PRCG* image set and (b) Subcategories within *Personal* image set, the number is the image count.

3 A Overview of the Columbia Photographic Images and Photorealistic Computer Graphics Dataset

We have designed and implemented the Columbia Photographic Images and Photorealistic Computer Graphics Dataset in accordance to the criteria mentioned in Section 2. The dataset is used in our work for the classification of PIM and PRCG [21]. The dataset consists of four sets of images, as shown in Figure 1 and briefly described below. A detailed description would be given in the subsequent sections.

1. **800 PRCG images from the Internet (*PRCG*)**: These images are categorized by content into architecture, game, nature, object and life, see Figure 2(a). The PRCG are mainly collected from 40 3D-graphics websites, such as `www.softimage.com`, `www.3ddart.org`, `www.3d-ring.com` and so on. The rendering software used are such as 3ds MAX, softimage-xsi, Maya, Terragen and so on. The geometry modelling tools used include AutoCAD, Rhinoceros, softimage-3D and so on. The high-end rendering techniques used include global illumination with ray tracing or radiosity, simulation of the camera depth-of-

field effect, soft-shadow, caustics effect (i.e., the specular light pattern seen near a glass when the glass is illuminated), and so on.

2. **PIM images from the personal collections (*Personal*)**: The *Personal* set consists of two parts, i.e., 800 images from the authors' personal collections (*Personal Columbia*) and 400 images from the personal collection of Philip Greenspun (*Personal Greenspun*). The reason for including images from Greenspun's collection is to increase the diversity of the *Personal* set in terms of the image content, the camera models and the photographer styles. The *Personal Greenspun* set are mainly travel images with content such as indoor, outdoor, people, objects, building and so on. Whereas the *Personal Columbia* set are acquired by the authors using the professional single-len-reflex (SLR) Canon 10D and Nikon D70. It has content diversity in terms of indoor or outdoor scenes, natural or artificial objects, and lighting conditions of day time, dusk or night time. See Figure 2(b).

3. **800 PIM from Google Image Search (*Google*)**: These images are the search results based on the keywords that match the categories within the *PRCG* set. The keywords are such as architecture, people, scenery, indoor, forest, statue and so on.

4. **800 photographed PRCG (*Recaptured CG*)**: These are the photograph of the screen display of the images from the *PRCG* set. Computer graphics are displayed on a 17-inch (gamma linearized) LCD monitor screen with a display resolution of 1280×1024 and photographed by a Canon G3 digital camera. The acquisition is conducted in a dark room in order to reduce the reflections from the ambient scene.

The rationale of collecting two different sets of PIM is the following: the *Google* set has a diverse image content and involves more types of cameras, photographer styles and lighting conditions but the ground truth may not be reliable, whereas the *Personal* set has reliable sources but it has limited diversity in camera and photographer style factors. On the other hand, based on the two-level definitions of image authenticity, i.e., the imaging-process authenticity and the scene, as introduced in [21], we should be able to restore the imaging-process authenticity of the PRCG by recapturing them using a camera. Therefore, we produce the *Recaptured PRCG* image set for evaluating how much the scene authenticity can be captured by a classifier.

The list below gives a summary of how the design criteria mentioned in Section 2 are fulfilled in the Columbia Photographic Images and Photorealistic Computer Graphics Dataset:

1. **PIM of a reliable source**: The *Personal* set are from the authors' and Greenspun's personal collection, therefore the images are certain to be photograph.

2. **PRCG from reliable sources**: Only PRCG from trustable websites are downloaded. These trustable websites requires the PRCG to be submitted with the name and the contact information of the creator.

3. **The content match for PIM and PRCG**: The content of images in the PIM and PRCG sets are limited to natural scene, which is defined as scenes commonly encountered by human. PIM shown in Figure 5 and PRCG shown in Figure 4 demonstrate the content match between the *PRCG* set and the *Personal Columbia* set.

4. **PIM with diversity in terms of the image content, the camera models and the photographer styles**: Although the *Personal* set has PIM of reliable source, it has limited diversity in terms of the image content, the camera model and the photographer style. This shortcoming is overcome by having the *Google* set, which is obtained from the Internet and is supposedly diverse in terms of the image content, the camera model and the photographer style.

5. **PRCG with diversity in terms of the image content and the rendering techniques**: Ensuring a variety of PRCG content categories such as human, animal, building, scenery, indoor, outdoor, objects and so on. The PRCG are also generated using different high-end computer graphics techniques (e.g., global illumination, soft shadowing, environment map) and software (e.g., SoftImage, Maya and so on).

6. **PRCG of high photorealism and from high-quality rendering techniques**: From the available rendering information of the *PRCG* set, the PRCG are mainly generated with high-quality rendering effects such as global illumination, caustics effect, simulation of camera depth of field and so on. The PRCG are also subjectively evaluated to be highly photorealistic.

7. **For evaluating the robustness of the classifier with respect to the image processing operations on PIM**: The RAW format of

the *Personal Columbia* images are recorded. RAW format image is the direct output from the CCD sensor and therefore can be used to generate images with different in-camera image processing operations such as white-balancing, sharpening, contrast adjustment and nonlinear transformation, as well as image compression. See Subsection 5.1.1 for more details.

8. **For evaluating the robustness of the classifier with respect to the computer graphics rendering techniques**: When the algorithm of the classifier is known, adversarial attack specific to the algorithm can be designed to manipulate the output of the classifier. For example, if the classifier is known for recognizing PRCG by the high color saturation, attackers can reduce the color saturation of a PRCG such that it is classified as a photographic image. Due to the dependency of an adversarial attack upon the specific algorithm, it is hard for our dataset to meet the specific needs from evaluating the classifier robustness against such attack.

9. **For evaluating sensitivity to ambiguity in content and camera effects**: In *Personal Columbia* set, there are images of artificial objects such as drawing, sculpture, wax figures, fake fruit for decoration, and so on. In the *PRCG* set, we have images of natural scenes such as forest, ponds, seaside as well as images of building. Furthermore, the simulation of the camera effects such as the depth-of-field (DoF) effect in the PRCG also results in an ambiguity in terms of camera effects.

4 The *PRCG* Set

This section will describe the acquisition process of the 800 PRCG in our dataset. The following list shows the properties of the *PRCG* set:

1. **Reliable Sources**: The images are downloaded from the professional 3D artist websites on the Internet such as `www.softimage.com`, `www.3ddart.org`, `www.3d-ring.com` and so on. Unfortunately, there are websites where people submit photographic images and claim to be PRCG. One characteristic of these unreliable websites is that the submitted PRCG are often from anonymous persons or by people identify themselves by nicknames. Therefore, we only download the trustable PRCG, submitted together with the creator's name, contact information. Table 1 shows the major online sources for the *PRCG* set.

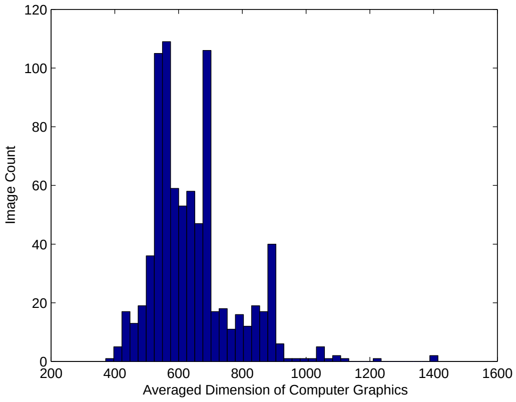


Figure 3: The histogram of the average dimension of the *PRCG* set

2. **Natural secne content scope**: Similar to the PIM image sets, the content of the images is limited to natural scene only. By setting this scope, many of the abstract or fantastic scene PRCG are excluded.

3. **Image size**: The size of the PRCG varies between 400 pixels and 1400 pixels in average dimension, $(width + height)/2$, with the mean at 645 pixels. Figure 3 shows the histogram of the average dimension of the PRCG set.

4. **High photorealism**: The images are visually inspected by the authors to ensure their high photorealism. Incidentally, from the available rendering information of the *PRCG* set, we can see that the PRCG images often have high quality rendering effect such as global illumination, caustics effects, or camera effect like depth of field (DoF).

5. **Record of the rendering information**: The associated rendering information for the PRCG is recorded during the collection process whenever it is available. See Section 4.1 for more details.

6. **Category**: The PRCG are categorized according to its content type. More details in Section 4.2.

7. **Post-processing**: The white border or text at the corners are cropped out and the PRCG duplicates are removed.

http://www.realsoft.fi	http://www.3ddart.org
http://www.softimage.com	http://www.3dshop.com
http://www.realsoft.com	http://www.3dlinks.com
http://www.realtimeuk.com	http://www.exchange3d.com
http://www.accurender.com	http://www.npowersoftware.com
http://www.marlinstudios.com	http://www.conceptvisualization.com
http://screenshots.teamxbox.com	http://www.psxextreme.com
http://www.sitexgraphics.com	http://www.greenworks.de
http://www.xfrogdownloads.com	http://www.marlinstudios.com
http://www.flamingo3d.com	http://www.cg3d.org
http://www.designcommunity.com	http://www.surrealstructures.com
http://www.archimodel.com	http://www.evs3d.com
http://www.artifice.com	http://www.mentalimages.com
http://www.abvent.com	http://www.pandromeda.com
http://www.cgchannel.com	http://www.3d-community.com
http://www.renderaid.com	http://artgallery.novatek-unlimited.com
http://www.highend3d.com	http://www.3dtotal.com
http://www.caligari.com	http://www.learning-maya.com
http://www.maxon.net	http://raph.com
http://www.3d-ring.com	http://www.digitalrepose.com

Table 1: Major online sources for the *PRCG* set. The column partitioning has no significance but just for compact display of information.

3ds max, viz	SOFTIMAGE—XSI
Real3D	V-ray
Lightwave 3D	Maya, paint FX
Terragen	Vue d'Esprit
AccuRender	Brazil R/S
Final Render	Shake
Vue 4 Pro	Discreet Combustion and Inferno
Alias ImageStudio	Blender
sasquatch	Messiah studio
Gaffer	Bryce
Hash Animation Master	Pixels 3D studio
Carrara	Flamingo
Strata 3D	Pov-ray/megapov
Mojoworld	Ray Dream Studio

Table 2: Rendering software or tools. The column partitioning has no significance but just for compact display of information.

4.1 Rendering Information of PRCG

Out of the 800 PRCG, 280 of them does not have any associated rendering information, the remaining 520 images has rendering information of various extent; some images contain a more complete rendering description while some contain a very mininal information of only the rendering software used. The rendering information includes the rendering software, the geometric modeling techniques, the rendering techniques, and the post-rendering processing. The information is extracted from the description of the PRCG found at the original webpage.

Table 2 lists the rendering software or tools used in the *PRCG* set. Most of them are the commercial computer graphics rendering software. Table 3 lists the geometric modeling software or tools used in the *PRCG* set and Table 4 lists the software used for the post-rendering image composition or for the creation of image texture. Finally, Table 5 lists the geometric modeling and the rendering techniques, where many of them are for producing high-photorealistic rendering effects.

4.2 Content Category of PRCG

The PRCG are categorized into the following content categories:

1. arch - images of building, architectural structure or building interior.

SOFTIMAGE—3D	Rhinoceros
AutoCAD	Facade
archT	Solid Works
formZ	Amapi
Poser	Nichimen Geometry

Table 3: Geometry modeling software or tools. The column partitioning has no significance but just for compact display of information.

Adobe Photoshop and After Effect	Universal Image Creator
Painter	Photopaint

Table 4: Software used for the post-rendering image composition or the creation of image texture

Geometric modeling using polygon, simple spline, NURBS[1], subdivision surfaces, solids, or meshsmooth	Global illumination using ray tracing or radiosity
Local illumination	Final gathering (for SOFTIMAGE—XSI)
Caustics effect	Simulation of depth of field, vignette, or motion blur
Texture modeling by image, hand-drawn, or procedural textures	Texture representation - image map, bump map, scope map or displacement map
Mapping high-dynamic range/radiance image to the final images	Using third party human, plant, hair or flu models
Area lighting, single/many light, skylight, environmental map	Adding procedural noise
Soft shadow	

Table 5: geometric modeling and rendering techniques. The row and column partitioning has no significance but just for compact display of information.

PRCG Category	**Count**
arch	295
game	41
nature	181
object	220
poeple	50
hybrid	13
Total	800

Table 6: The count of the PRCG in each category

2. game - images from computer games.
3. nature - images of natural scenery such as forest, plants, beaches, ponds and also general outdoor scene.
4. object - images of objects such as books, tables, watches and so on.
5. people - images of human, animal or insects.
6. hybrid - images of combined computer graphics and camera images.

Figure 4 shows the example images for the content categories within the *PRCG* set. Table 6 shows the count for the PRCG of the different content category.

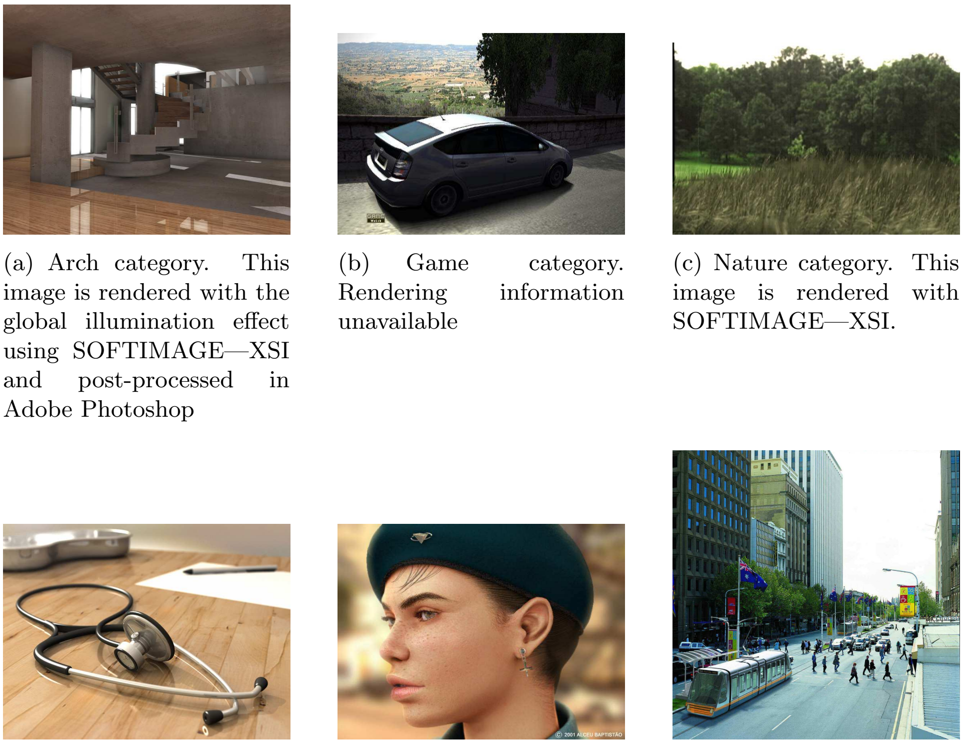
5 The *Personal* Set

The *Personal* set consists of two parts, i.e., 800 images from the authors' personal collections (*Personal Columbia*) and 400 images from the personal collection of Philip Greenspun (*Personal Greenspun*). The reason for including images for Greenspun's collection is to increase the diversity of the *Personal* set in terms of the image content, the camera models and the photographer styles.

5.1 The *Personal Columbia* Set

The list below describes the characteristics of the *Personal Columbia* set:

1. **Reliable sources**: The images are captured by the authors in order to ensure its reliable authenticity. These images are captured in New



(a) Arch category. This image is rendered with the global illumination effect using SOFTIMAGE—XSI and post-processed in Adobe Photoshop

(b) Game category. Rendering information unavailable

(c) Nature category. This image is rendered with SOFTIMAGE—XSI.

(d) Object category. This image is rendered with the final gathering effect (a simulated effect of secondary illumination) using SOFTIMAGE—XSI

(e) People category. This image is rendered using Maya rendering software and post-processed in Adobe Photoshop

(f) Hybrid category. This image is rendering using the Realsoft 3D rendering software.

Figure 4: Example images from the different PRCG content categories and their associated rendering information. Note that rendering information is not available for the game category and the amount of rendering varies for different PRCG.

York City and Boston during the summer and early autumn of year 2004, using two models of professional single-lens reflex (SLR) camera which are Canon 10D and Nikon D70.

2. **Format**: The images are stored simultaneously in both RAW and JPEG format. While RAW images provide a control on the in-camera image processing operations, including the compression option, on the final image, the JPEG images provide images indisputably processed by the camera. More information is in Section 5.1.1.

3. **Camera parameters**: Camera parameters are kept in the EXIF meta-data format and stored with the RAW and JPEG images. More information is in Section 5.1.2.

4. **Natural scene content**: Like the *PRCG* set, the image content is limited to natural scene only. Figure 5 shows the examples images from the *Personal Columbia* set which match the content of the PRCG shown in Figure 4.

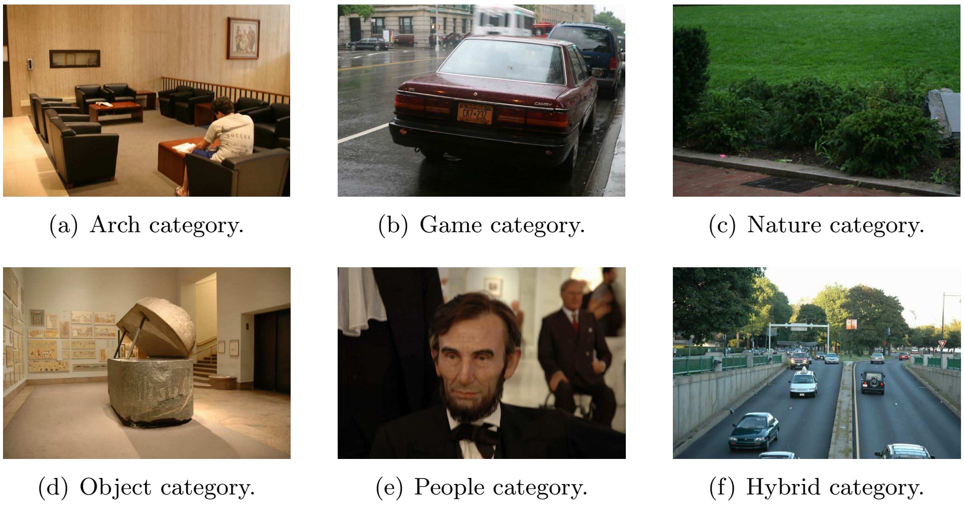
5. **Image size**: The original size of the PIM are 3072 × 2048 pixels for Canon 10D images and 3008 × 2000 pixels for Nikon D70 images.

6. **Diversity in the photographic process**: There is a diversity in terms of the major variables within a photographic process, i.e., the illumination sources, the objects, the camera, and the photographer. More explanation is in Section 5.1.3.

5.1.1 RAW Format of PIM

RAW images, often regarded as the negative of digital images, are the direct output of the camera CCD sensors. RAW format is proprietary to the camera manufacturers. Canon associates its RAW format images with the .CRW file extension, while Nikon RAW image file has a .NEF file extension. The settings of the in-camera operation, such as white-balancing, radiometric transformation, sharpening effect and so on, are recorded in the RAW-format images, but not applied to the raw sensors data. Therefore, the in-camera settings can be changed later using any external RAW processing software. Although camera manufacturers provide their own RAW processing software, there are also third party RAW processing software, such as Adobe Photoshop and BreezeBrowser Pro[2].

[2] see `http://www.breezesys.com/BreezeBrowser/`



(a) Arch category. (b) Game category. (c) Nature category.

(d) Object category. (e) People category. (f) Hybrid category.

Figure 5: Example images from the PIM set that resemble the respective PRCG shown in Figure 4 in terms of the content.

From RAW images, we can generate images of different in-camera image processing operations such as white-balancing, sharpening, contrast adjustment, as well as compression. As a result, both uncompressed or lossy-compressed images can be obtained from RAW images. It is possible that the built-in image processing operation in a camera may be different from that of the RAW processing software. We assume that the operations in the software supplied by the camera manufacturer are not too different from those built into the camera, even if they are not identical.

5.1.2 EXIF Metadata of PIM

The EXIF metadata records the camera settings for which an image is captured which include camera model, shooting date and time, shooting mode (e.g. auto or manual mode), shutter speed, aperture value, exposure compensation, metering mode, ISO Speed, lens focal range, focal length, image quality, flash setting, white balance and so on. The EXIF format is a template and the amount of information actually provided is camera dependent. EXIF format metadata can be extracted using the camera manufacturer software or freeware such EXIF Reader[3].

[3] see `http://www.takenet.or.jp/~ryuuji/minisoft/exifread/english/`

Image Category for the *Personal Columbia* Set	**Count**
indoor-light	74
indoor	68
outdoor-day	277
outdoor-night	34
outdoor-rain	63
outdoor-dawn-dusk	31
natural	111
artificial	142
Total	800

Table 7: Image count for the categories within the *Personal Columbia* Set

5.1.3 Diversity of the *Personal Columbia* Set

The PIM are designed to have diversity in the major variables of a photographics process:

1. Illumination source/lighting condition: There are content categories of indoor-dark, indoor-light, outdoor-dawn-dusk, outdoor-day, outdoor-night and outdoor-rain.

2. Object - There are content categories of natural and artificial.

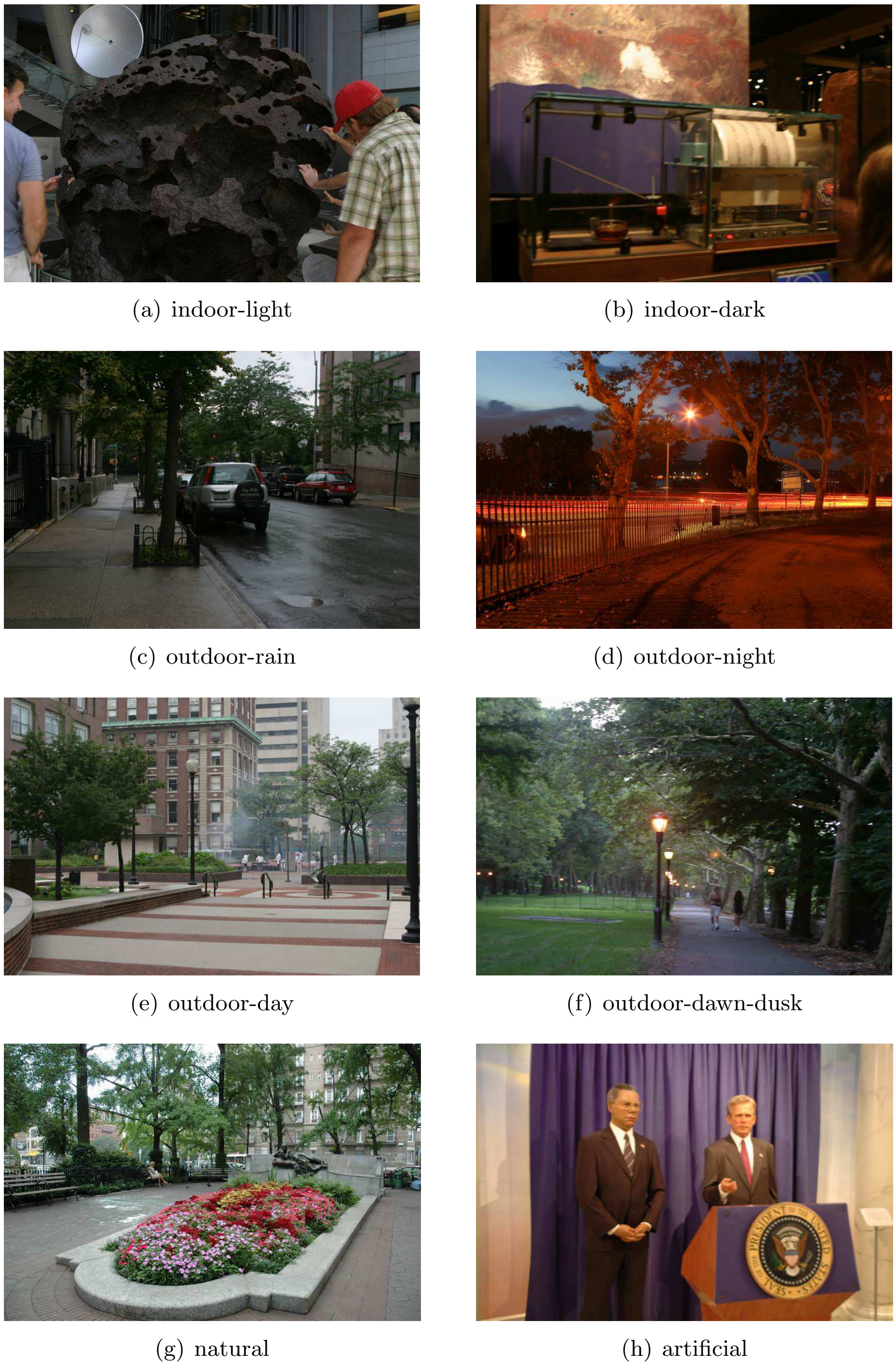
3. Camera - The images are taken by two models of SLR cameras, i.e., Canon 10D and Nikon D70, which are known to have different makes of in-camera image processor.

4. Photographer/personal photographing style: We have three photographers identified as 'martin', 'jessie' and 'tian-tsong' in the dataset meta-inforation.

Figure 6 shows the example images for different lighting and object categories mentioned above. Table 7 shows the count of the images in each category.

5.2 The *Personal Greenspun* Set

The *personal Greenspun* set is from the personal collection of Philip Greenspun, which is accessible from `http://www.photo.net/philip-greenspun/photos/digiphotos/`.



(a) indoor-light

(b) indoor-dark

(c) outdoor-rain

(d) outdoor-night

(e) outdoor-day

(f) outdoor-dawn-dusk

(g) natural

(h) artificial

Figure 6: Example images for different lighting and object categories within *Personal Columbia* image set

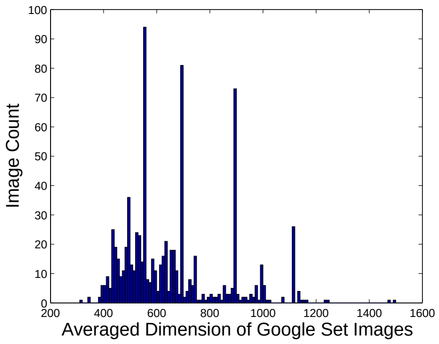


Figure 7: The histogram of the average dimension of the *Google* set.

6 The ***Google*** Set

The *Google* set is from the search results based on the keywords that matches the *PRCG* set. The keywords are architecture, nature scene, landscape, animal, building, people, scenery, indoor, object, machine, insect, interior, plant, forest, vehicle, fruit and statue. The images are filtered subjectively to include only PIM of size larger than 300 pixels for the width or the height, whichever smaller. The size of the *Google* images varies between 315 pixels and 1500 pixels in average dimension, $(width + height)/2$, with the mean at 660 pixels. Figure 7 shows the histogram of the average dimension of the *Google* set.

7 The ***Recaptured PRCG*** Set

The *recaptured PRCG* set consists of the photographed screen display of the 800 images from the *PRCG* set. PRCG are displayed on a 17-inch LCD monitor screen with a display resolution of 1280×1024 pixels and photographed by a Canon G3 digital camera mounted on a tripod at a distance of four feet in front of the screen. The captured images have a resolution 1024×768 pixels and of a high quality JPEG compression. The camera operates at the time-priority mode where the shutter speed is set to 1/4 seconds while the aperture size and focal length are set automatic. The acquisition is conducted in a dark room in order to reduce the reflections from the ambient scene.

8 A guide for Downloading the Columbia Photographic Images and Photorealistic Computer Graphics Dataset

The download and other additional information for the dataset is described in the dataset website, accessible from the Trustfoto website: `http://www.ee.columbia.edu/dvmm/trustfoto`. Except for the *Personal Columbia* set, we are not able to diseeminate the original images from the *PRCG*, *Personal Greenspun*, *Google* and *Recaptured PRCG* image sets, due to copyright constriants. We have webpages for each of the image sets, which show the thumbnails from the image sets. The *PRCG* and the *Google* thumbnail webpages provide the hyperlinks and the URLs to the original location of the images. Therefore, researchers can follow the URLs to access the original web pages containing the images and download the images if they want. The following list describes how to obtain the image sets of the Columbia Photographic Images and Photorealistic Computer Graphics Dataset.

- *PRCG* set: The *PRCG* images can be downloaded by following the hyperlinks or the URLs listed in the *PRCG* thumbnails page, accessible from the dataset website. The thumbnail page also contains meta-information including the source URL, the content category, the software or tools used, the rendering techniques, the geometric modeling techniques and the post-rendering processing.

- *Personal* set:

    1. *Personal Columbia* set: The *Personal Columbia* images come in three versions: the downsized (about 700x500 pixels) JPEG version (55MB), the original-size (about 3000x2000 pixels) JPEG version (1GB), the original-size RAW-format version (4.9GB). The downsized version are used in our work for experiments [21]. The downsized and the original-size JPEG version of the *Personal Columbia* set can be downloaded from the dataset website, while the original-size RAW-format version will be distributed in DVD and can be obtained upon request. The *Personal Columbia* thumbnail webpage shows the thumbnails of the *Personal Columbia* set as well as the meta-information of the images such asthe filename, the lighting/object category, the photographer and the camera model.

2. *Personal Greenspun* set: The *Personal Greenspun* images can be downloaded from `http://www.photo.net/philip-greenspun/photos/digiphotos/`. The thumbnails as well as the filename of the 400 *Personal Greenspun* images can be found in the *Personal Greenspun* thumbnail page.

- *Google* set: The *Google* images can be downloaded by following the hyperlinks or the URLs listed in the *Google* thumbnails page, accessible from the dataset website.

- *Recaptured PRCG* set: As the recaptured PRCG are the reproduction of the PRCG, the distribution of these images is limited by the copyright. As a result, the *Recaptured PRCG* will not be distributed.

9 Conclusions

This technical report explains the needs for a new dataset catering for the PIM versus PRCG classification problem of the passive-blind image authentication research. During the process of working on the mentioned problem, we have designed and implemented a dataset that specifically takes into account the unbiased diversity of content, reliability of sources, and advanced CG rendering techniques. This Columbia Photographic Images and Photorealistic Computer Graphics Dataset is to be made available to the research community and this technical report will serve as a user guide for the dataset. We make the data set open and publicly accessible in order to promote active interest and broad collaboration in this exciting area.

10 Acknowledgements

This project is supported in part by NSF CyberTrust program (IIS-04-30258) and Singapore A*STAR scholarship (for the first author). The authors thank Philip Greenspun for granting us the permission of using his personal images and Lexing Xie for her kind help.

References

[1] I. J. Cox, M. L. Miller, and J. A. Bloom, *Digital Watermarking*, ser. The Morgan Kaufmann Series in Multimedia Information and Systems. Morgan Kaufmann, 2002.

[2] M. Yeung and F. Mintzer, “An invisible watermarking technique for image verification,” in *IEEE International Conference on Image Processing*, vol. 2, 1997, pp. 680 – 683.

[3] M. Wu and B. Liu, “Watermarking for image authentication,” in *IEEE International Conference on Image Processing*, vol. 2, 1998, pp. 437–441.

[4] J. Fridrich, M. Goljan, and A. C. Baldoza, “New fragile authentication watermark for images,” in *IEEE International Conference on Image Processing*, Vancouver, Canada, 2000.

[5] P. W. Wong, “A watermark for image integrity and ownership verification,” in *IS&T Conference on Image Processing, Image Quality and Image Capture Systems*, Portland, Oregon, 1998.

[6] E. T. Lin, C. I. Podilchuk, and E. J. Delp, “Detection of image alterations using semi-fragile watermarks,” in *SPIE International Conference on Security and Watermarking of Multimedia Contents II*, vol. 3971, San Jose, CA, 2000.

[7] C.-Y. Lin and S.-F. Chang, “A robust image authentication method surviving jpeg lossy compression,” in *SPIE Storage and Retrieval of Image/Video Database*, ser. IS&T/SPIE Symposium on Electronic Imaging: Science and Technology, San Jose, 1998.

[8] J. Fridrich, “Image watermarking for tamper detection,” in *IEEE International Conference on Image Processing*, Chicago, 1998.

[9] N. Memon and P. Vora, “Authentication techniques for multimedia content,” in *SPIE Multimedia Systems and Applications*, Boston, MA, 1998.

[10] C.-Y. Lin and S.-F. Chang, “A robust image authentication method distinguishing jpeg compression from malicious manipulation,” *IEEE Transactions on Circuits and Systems for Video Technology*, 2000.

[11] M. Schneider and S.-F. Chang, “A robust content based digital signature for image authentication,” in *IEEE International Conference on Image Processing*, vol. 3, Lausanne, Switzerland, 1996, pp. 227–230.

[12] S. Bhattacharjee, “Compression tolerant image authentication,” in *IEEE International Conference on Image Processing*, Chicago, 1998.

[13] E.-C. Chang, M. S. Kankanhalli, X. Guan, H. Zhiyong, and W. Yinghui, “Image authentication using content based compression,” *ACM Multimedia Systems*, vol. 9, no. 2, pp. 121–130, 2003.

[14] H. Farid and S. Lyu, “Higher-order wavelet statistics and their application to digital forensics,” in *IEEE Workshop on Statistical Analysis in Computer Vision*, Madison, Wisconsin, 2003.

[15] T.-T. Ng, S.-F. Chang, and Q. Sun, “Blind detection of photomontage using higher order statistics,” in *IEEE International Symposium on Circuits and Systems*, Vancouver, Canada, 2004.

[16] “Image database for digital watermarking.” [Online]. Available: http://www.petitcolas.net/fabien/watermarking

[17] A. Georghiades, P. Belhumeur, and D. Kriegman, “From few to many: Illumination cone models for face recognition under variable lighting and pose,” *IEEE Transactions in Pattern Analysis Machine Intelligence*, vol. 23, no. 6, pp. 643–660, 2001.

[18] “Usc-sipi image database,” University of Southern California. [Online]. Available: http://sipi.usc.edu/services/database/Database.html

[19] T.-T. Ng and S.-F. Chang, “A data set of authentic and spliced image blocks,” Columbia University, ADVENT Technical Report 203-2004 -3, June 8 2004.

[20] S. Lyu and H. Farid, “How realistic is photorealistic?” *IEEE Transactions on Signal Processing*, vol. 53, no. 2, pp. 845–850, 2005.

[21] submitted to a conference for blind review.

[22] T. Ianeva, A. de Vries, and H. Rohrig, “Detecting cartoons: A case study in automatic video-genre classification,” in *IEEE International Conference on Multimedia and Expo*, vol. 1, 2003, pp. 449 – 452.

[23] J. R. Smith and S.-F. Chang, “Visually searching the web for content,” *IEEE Multimedia*, vol. 4, no. 3, pp. 12–20, 1997.