A Practical Classifier for Photographs and Non-Photographic Images Based on Local Visual Features

Kei Terayama and Hirohisa Hioki
Graduate School of Human and Environmental Studies
Kyoto University, Kyoto 606-8501, Japan
{terayama, hioki}@i.h.kyoto-u.ac.jp

Abstract

Classification of digital images into photographs and various kinds of non-photographic images has not been sufficiently studied but has many applications such as retrieval of real scene photographs from web sites and image databases. In this paper, we show that the combination of Bag of Visual Words of SURF features and histograms of LBPs for HSV and Luminance components (SURF+LBP(HSVL)) is simple, but works well as visual features for photographs and non-photographic image classification. We found that a classifier trained with SURF+LBP(HSVL) was the best among all the classifiers we tested using various visual features. Our classifier attained an accuracy of 96.8% for our image dataset and outperformed the other state-of-the-art classifiers.

1 Introduction

The classification problem of digital images into photographs (photo) and various kinds of non-photographic (non-photo) images (including paintings, pixel arts, line drawings and CG images) has been presented in [1]. Constructing a classifier for the problem is important for many applications. A classifier for photos and non-photo images enables us to retrieve only real scene photographs from web sites. It is useful to provide such a classifier as a preprocessor for image analysis tasks. If images are properly classified into photos and non-photos at the first stage of analysis, we can select methods suitable for photos or for non-photos adaptively in later stages. If a classifier does not only separate photos from non-photos but is also able to measure the degree of "photorealisticness" or "non-photorealisticness" of images, it is possible to objectively evaluate the performance of a non-photorealistic rendering (NPR) method.

In general, the problem of image classification has been addressed in the fields of scene recognition and image retrieval. Luo and Savakis proposed a method for classifying indoor vs outdoor images [2]. Vailaya et al. developed a method to identify city and landscape images [3]. Several research efforts related to the classification of photos and non-photo images have far been carried out. The classifier in [4] discriminated photographs from CG drawings. Photographs and paintings were classified in [5, 6]. In these methods, visual features derived from color, edge and texture are employed for the purpose of discrimination. For example, texture features are computed by Gabor filter with consideration of the human visual system. Recently, in the field of digital forensics, several classifiers for identifying digital camera images and realistic CG images have been proposed [7, 8].

We have proposed a classification method for photographs of real scenes and various kinds of non-photo images, including paintings, pixel arts, line drawings, CG images, and even hand-drawn pictures taken by camera [1]. We employed several visual features by considering the limitations of hand drawn actions and achieved 95% accuracy with the collected photos and non-photos dataset in [1]. However, it is not certain that the visual features we had employed are effective, because we did not compare the classification method with others sufficiently.

In this study, we prepared a variety of classifiers, including our previous one, and compared them through experiments. For constructing classifiers, instead of introducing our own visual features, we select local visual features that are already used in the field of image processing successfully, because such features are known to work effectively and seem promising for our classification task. Among various features, we find that the combination of Bag of Visual Words (BoVW) of SURF features and histograms of LBPs for HSV and Luminance components (SURF+LBP(HSVL)) is simple but works well practically.

For classification experiments, we collected images from three image-sharing websites, Flickr [9], deviantART [11] and Pixiv [10]. A total of 25,000 photos and 25,000 non-photo images has been retrieved. For the images, we calculated various local visual features that were proposed in related work and basic ones such as LBPs and BoVW of SIFT and SURF features and their combinations. After that, by selecting training image sets from our dataset, various classifiers were constructed with non-linear SVMs using three kernel functions: RBF, chi squared and histogram intersection kernels. We found that a classifier trained with SURF+LBP(HSVL) was the best among all the classifiers we tested. Our classifier attained an accuracy of 96.8% for our image dataset and outperformed the other state-of-the-art classifiers.

In the rest of this paper, the detail of our image dataset is described in Section 2. We introduce various visual features we tested and explain how to construct classifiers in Section 3. Section 4 gives our experimental results. Finally we summarize this paper and give directions of future work in Section 5.

2 Dataset

For our experiments we collected photos and non-photo images from three web sites: the photo sharing site Flickr [9], an illustration sharing site called Pixiv [10] and the digital art sharing site deviantART (dART) [11] where we were able to find both photograph and digital art image categories. We picked im-

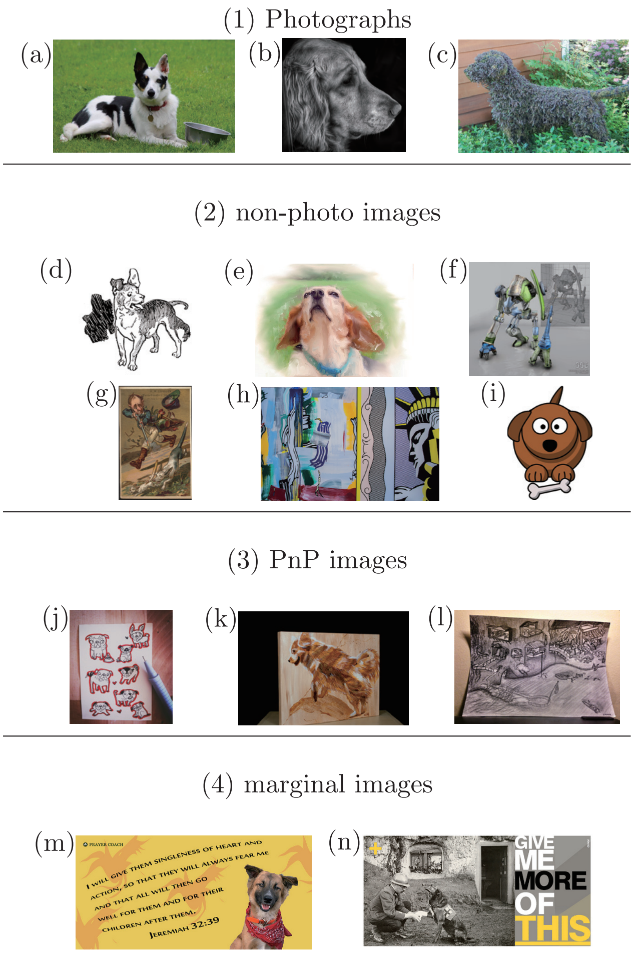


Figure 1. Examples of four groups:(1) photographs, (2) non-photo images, (3) PnP images and (4) marginal images. (a) by Jonny Green, (b) by psyberartist, (c) by InAweofGod'sCreation, (d) by CircaSassy, (e) by Lisa Yarost, (f) by Veronica Electronica, (g) by Boston Public Library, (h) by Frank Kovalchek, (j) by gokce yavas onal, (k) by Ben Mason, (l) and (n) by danaor shtruzman and (m) by Kevin Shorter from Flickr. (i) public domain image.

ages up from the following five categories: dog, landscape, people, train and beach.

The images collected from Flickr and the photograph category of dART mainly consist of photos, while the majority of images from Pixiv and the digital art category of dART are non-photo images. We, however, also found exceptions. Non-photo images were found in Flickr and in the photograph category of dART. Photographs were collected from Pixiv and the digital art category of dART. In addition, we found images which were neither real photographs nor non-photo images created from scratch. For example, there were retouched photos, images of paintings taken by camera, and composite images that have both photographic and non-photographic regions.

As we just show examples above, in general, it is not obvious for us to divide images into two distinct groups of photos and non-photo images. For this problem, in [1], four image groups are defined: (1) photographs, (2) non-photo images, (3) photographs of non-photo (PnP) images and (4) marginal images. A typical image in group (3) is a photograph which captures a hand-drawn picture or a CG print. Images in (4) include those that can be hardly distinguishable either as a photograph or a non-photo image and those that have both the photo and non-photo regions. Figure 1 shows examples of the four groups.

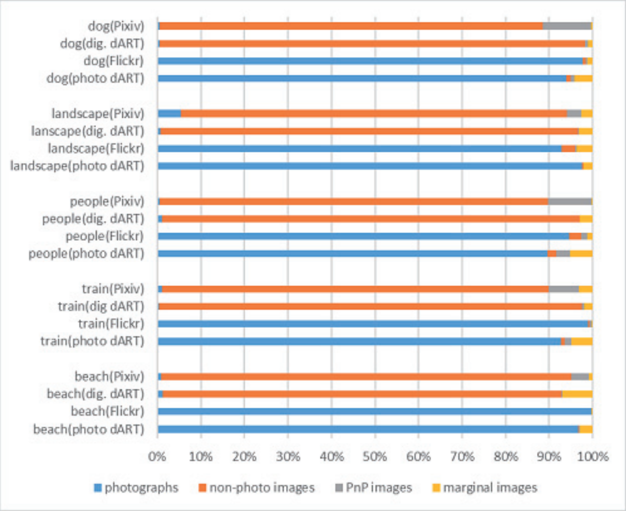


Figure 2. The composition ratios of four image groups of collected images

For the images we collected from web sites, we manually classified them into the above four groups. Figure 2 shows the composition ratios of each group of the collected images.

In this study, we did not take images in groups (3) and (4) into our dataset, which means that we concentrated on classifying images in groups (1) and (2). It is not obvious whether PnP images should be classified as photos or non-photos. A PnP image is physically a photograph, but what we find there is a hand-drawn picture or a CG image, i.e., we do not see a real scene in a PnP image. In this sense, a PnP image is a non-photo image. How we should classify PnP images hence depends on the purpose of classification. We therefore exclude them because we aim to construct a classifier useful for general purposes. It is not easy to determine whether a retouched photo image should be classified as a photo or non-photo. The situation is the same for composite images that have both photo and non-photo regions. We do not take such images into our dataset in order to make our classification steady.

For each of the five image categories mentioned above (i.e. dog, landscape, people, train and beach), we collected 2,500 photos from Flickr and the photograph category of dART respectively. In addition, 2,500 non-photo images were obtained from Pixiv and the digital art category of dART respectively. We therefore have 10,000 images for each image category and our dataset consists of a total of 50,000 images.

3 Visual Features

In this section, we introduce the visual features used for classification experiments. We first present several basic visual features that are shown to be powerful for image classification or texture classification tasks.

We then describe the visual features proposed in [1,6,8] that are related to the photo and non-photo image classification. How we construct classifiers from the visual features is briefly described subsequently.

3.1 Basic visual features

For our image classification task, we examined the following basic visual features: local binary pattern (LBP) [12], BoVW of SIFT [13] features, BoVW of SURF features [14] and higher-order local autocorrelation (HLAC) [15]. All of these features can be computed locally, i.e., they are local visual features.

LBP was proposed in [12] and has been widely used for various purposes such as image classification and face detection. Many variants and extensions of LBP have been devised such as [16]. We compute standard LBP values of images in HSV and YCbCr color space and use their histograms as visual features.

The BoVW approach has been successfully applied to problems such as image classification, image retrieval and object recognition. We derive visual features of BoVW with SIFT [13] and SURF [14] features. We selected about 50 images from photos and non-photo images respectively in our dataset to build a codebook, with which we obtain histograms of the codewords as feature vectors of images.

We also examined HLAC [15] as a visual feature. HLAC is a powerful feature that satisfies additive and position invariant properties. HLAC is used in various applications such as texture classification and face recognition. In this study, we calculate standard HLAC features whose dimension is 35 for each image and use it as a visual feature.

3.2 Visual features of related work

Hammoud et al. proposed image filters that classify images either as photographs of real-scene or as art painting [5,6]. In their studies, they developed feature vectors consisting of several visual features including ones derived from colors, edges and texture information measured by Gabor filters. In [6], their method is applied to a dataset that consists of 10,000 photographs and 10,000 paintings. A feature vector called Receptive Field Profiles (RFPs), which is of dimension 72 calculated by Gabor filter, showed the best classification performance (93%) for their image dataset.

Li et al. proposed a method distinguishing CG images from photo images using uniform LBP [16] with the help of SVM [8]. They first extracted Y and Cb components of each image in YCbCr color space and computed prediction-error images of the two components. They then compute uniform LBP features from the four images: the Y and Cb components and their prediction-errors images. They constructed a database which consists of 2,455 CG images and 2,455 photo images. Their method with 236 visual features achieved higher classification accuracy compared to state-of-the-art works at that time. We call their visual feature LBP(uniform) in this paper.

We proposed a method for classifying photos and various kinds of non-photo images based on visual features derived by considering the limitations of hand-drawing actions in our previous work [1]. We collected a dataset consisting of over 130,000 images including photos, non-photo images, PnP images and marginal images and introduced feature vectors of 371 dimensions. The accuracy of the SVM classifier achieved over 95% where PnP images were regarded as a kind of non-photo images. As is already stated, the features employed in [1] have not been evaluated enough and it is not clear how the features have contributed to improve classification accuracies. Because the visual feature based on visual characteristics of hand-drawn images, we call it FHD in this paper.

3.3 Construction of classifier from visual features

We construct classifiers with non-linear SVM. The RBF kernel is commonly used for non-linear SVM. For image classification and object recognition tasks, SVM with the chi squared and histogram intersection kernels has been employed successfully under the BoVW approach [17,18]. We hence constructed various classifiers as SVM trained with our dataset using RBF, chi squared and histogram intersection kernels. For constructing classifiers, we first provided each visual feature described above separately. We also prepared classifiers using a combination of different visual features, which appear promising for improving classification accuracy. Among them, SURF+LBP(HSVL), i.e. the combination of BoVW of SURF features and histograms of LBPs for HSV and Luminance components, was found to be best through our experiments. Details are shown in the next section.

4 Experimental Results

We present experimental results in this section. We first applied various classifiers to the images in the dog category and compared 10-fold cross-validation accuracies of the classifiers. Here, we define accuracy by the following equation:

$$\text{accuracy} = \frac{\text{PP} + \text{NN}}{\text{total images}}$$

where PP and NN respectively represent the numbers of correctly identified images of photos and non-photo images. In the first experiment, images larger than a threshold were shrunk. We then further shrank images to smaller sizes and performed the second experiment to examine how resolutions of images affect accuracies of classifiers. Finally, in the third experiment, the classification results of our method for all the five image categories in our dataset are shown.

4.1 Experiment 1

Table 1 shows the classification results for the images in the dog category with various classifiers. In this experiment, we shrank images with preserving aspect ratios and the longer side of each image became not longer than a threshold of 1280 pixels. The images whose longer sides are shorter than the threshold were used as they were. Among all the classifiers we have tested, the one with SURF+LBP(HSVL) with chi squared kernel showed the best result: the accuracy was about 98%.

Table 1. Classification results for dog category

visual feature	feature size	kernel	CV
SURF+LBP(HSVL)	2024	χ^2	0.9812
SURF+LBP(HSVL)	2024	RBF	0.9788
SURF	1000	χ^2	0.9744
SURF	1000	RBF	0.9722
LBP(HSVL)	1024	χ^2	0.9668
LBP(HSVL)	1024	RBF	0.9667
LBP(YCbCr)	768	χ^2	0.9614
LBP(YCbCr)	768	RBF	0.9610
LBP(L)	256	χ^2	0.9599
LBP(L)	256	RBF	0.9595
LBP(uniform) [8]	236	χ^2	0.9594
LBP(uniform) [8]	236	RBF	0.9589
FHD [1]	371	χ^2	0.9499
FHD [1]	371	RBF	0.9532
SIFT	1000	χ^2	0.9458
SIFT	1000	RBF	0.9352
RFPs [6]	72	χ^2	0.9034
RFPs [6]	72	RBF	0.9053
HLAC	35	χ^2	0.8160
HLAC	35	RBF	0.8204

In our experiments, the chi squared kernel showed better performance than the RBF kernel in many cases. We have also experimented with the histogram intersection kernel, but the accuracies were lower than those of the other two kernels and we thus omitted the results from Table 1.

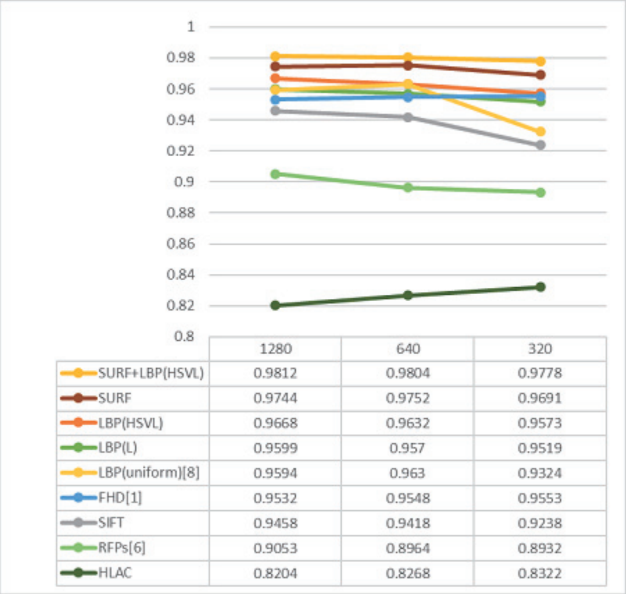
4.2 Experiment 2

In the second experiment, we further shrank images to smaller sizes and examined how the resolutions of images affect accuracies of classifiers. The results of Experiment 1 (the size threshold was 1280 pixels) were compared with the results where thresholds of 640 pixels and 320 pixels were used. Note that the images were shrunk as in the case of Experiment 1. The RBF kernel was employed for the features of RFPs, FDH and HLAC whereas the chi squared kernel was employed for the others.

Figure 3 shows the results. Although the accuracies decreased in the most cases as the sizes of the images decreased, the classifier with SURF+LBP(HSVL) was found to be the best for all sizes and achieved about 97.8% accuracy even when the size threshold was 320 pixels.

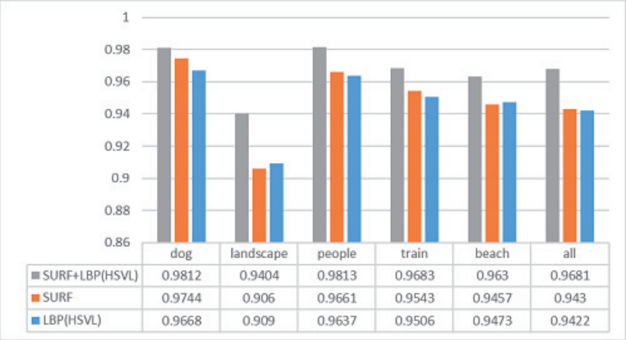
4.3 Experiment 3

From the results of first and second experiments, the classifier with SURF+LBP(HSVL) was found to be effective. We thus further examined this classifier by applying it to each of the five image categories and the set consisting of all the images in five categories in our dataset. In this experiment, the classifier with SURF and the classifier with LBP(HSVL) were also tested for comparison. The chi squared kernel was used for the three classifiers.



	1280	640	320
SURF+LBP(HSVL)	0.9812	0.9804	0.9778
SURF	0.9744	0.9752	0.9691
LBP(HSVL)	0.9668	0.9632	0.9573
LBP(L)	0.9599	0.957	0.9519
LBP(uniform)[8]	0.9594	0.963	0.9324
FHD[1]	0.9532	0.9548	0.9553
SIFT	0.9458	0.9418	0.9238
RFPs[6]	0.9053	0.8964	0.8932
HLAC	0.8204	0.8268	0.8322

Figure 3. Classification results for three sizes



	dog	landscape	people	train	beach	all
SURF+LBP(HSVL)	0.9812	0.9404	0.9813	0.9683	0.963	0.9681
SURF	0.9744	0.906	0.9661	0.9543	0.9457	0.943
LBP(HSVL)	0.9668	0.909	0.9637	0.9506	0.9473	0.9422

Figure 4. Classification results for all categories

Figure 4 shows the results. Although the accuracy for the landscape category was lower than those for the other categories, when we employed SURF+LBP(HSVL), the average accuracy for the categories achieved about 96.7% and the accuracy for all the categories was about 96.8%. These results were better than the results for the classifier with SURF and the classifier with LBP(HSVL).

5 Conclusion

In this paper, the classification problem of photo and non-photo images has been addressed. We have shown that the classifier with SURF+LBP(HSVL) was the best among all the classifiers we tested using various local visual features. The classification accuracy of our classifier attained an accuracy of 96.8% for our dataset of 50,000 images. Our classifier is based on a simple combination of well-known visual features, but the experimental results indicate that our classifier is practically useful.

Our future work includes the improvement of classification performance by introducing other variants of LBP and other keypoint descriptors such as ORB and D-BRIEF. Analyzing why our classifier effectively works will help us to understand what is “photorealisticness” and “non-photorealisticness”. It is also worth considering employing a training framework for classifiers other than non-linear SVM such as deep learning.

References

[1] K. Terayama and H. Hioki, "A method classifying digital images into photos and non-photos based on visual characteristics of hand-drawn images," in *The 17th Meeting on Image Recognition and Understanding*, 2014.

[2] J. Luo and A. Savakis, "Indoor vs outdoor classification of consumer photographs using low-level and semantic features," in *ICIP*, vol.2, pp.745-748, 2001.

[3] A. Vailaya, A. Jain, and H. J. Zhang, "On image classification: city vs. landscape," in *Content-Based Access of Image and Video Libraries*, pp.3-8, 1998.

[4] V. Athitsos, M. J. Swain, and C. Frankel, "Distinguishing photographs and graphics on the world wide web," in *Content-Based Access of Image and Video Libraries*, pp.10-17, 1997.

[5] F. Cutzu, R. Hammoud, and A. Leykin, "Estimating the photorealism of images: Distinguishing paintings from photographs," in *CVPR*, pp.305-312, 2003.

[6] R. Hammoud, "Color texture signatures for art-paintings vs. scene-photographs based on human visual system," in *ICPR*, vol.2, pp.525-528, 2004.

[7] D. Chen, J. Li, S. Wang, and S. Li, "Identifying computer generated and digital camera images using fractional lower order moments," in *4th IEEE Conference on Industrial Electronics and Applications*, pp.230-235, 2009.

[8] Z. Li, J. Ye, and Y. Q. Shi, "Distinguishing computer graphics from photographic images using local binary patterns," in *Digital Forensics and Watermarking*, pp.228-241, Springer, 2013.

[9] Flickr. http://www.flickr.com/.

[10] Pixiv. http://www.pixiv.net/.

[11] deviantART. http://www.deviantart.com/.

[12] T. Ojala, M. Pietikainen, and D. Harwood, "Performance evaluation of texture measures with classification based on kullback discrimination of distributions," in *ICPR*, vol.1, pp.582-585, 1994.

[13] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. Journal of Computer Vision*, vol.60, no.2, pp.91-110, 2004.

[14] H. Bay, T. Tuytelarrs, and L. V. Gool, "SURF: Speeded up robust features," in *ECCV*, pp.404-417, 2006.

[15] N. Otsu and T. Kurita, "A new scheme for practical flexible and intelligent vision systems," in *IAPR Workshop on Computer Vision*, pp.431-435, 1988.

[16] T. Ojala, M. Pietikainen, and T. Maenpaad, "Multiresolution gray-scale and rotation invariant texture classification with local binary patters," *IEEE Trans. on PAMI*, vol.24, no.7, pp.971-987, 2002.

[17] F. Barla A. Odone and A. Verri. "Building kernels from binary string for image matching," *IEEE Trans. on Image Processing*, vol.14, no.2, pp.169-180, 2005.

[18] A.C. Berg S. Maji and J. Malik. "Classification using intersection kernel support vector machines is efficient," in *CVPR*, pp.1-8, 2008.