21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Recognizing Surface Qualities from Natural Images based on Learning to Rank

Takashi Abe, Takayuki Okatani, and Koichiro Deguchi
Graduate School of Information Sciences, Tohoku University, Japan
{tabe,okatani}@fractal.is.tohoku.ac.jp

Abstract

This paper proposes a method for estimating the quantitative values of some attributes associated with surface qualities of an object, such as glossiness and transparency, from its image. Our approach is to learn functions that compute such attribute values from the input image by using training data given in the form of relative information. To be specific, each sample of the training data represents that, for a pair of images, which is greater in terms of the target attribute. The functions are learned based on leaning to rank. This approach enables us to deal with natural images, which cannot be dealt with in previous works, which are based on CG synthesized images for both training and testing. We created data sets using the Flickr Material Database for four attributes of glossiness, transparency, smoothness, and coldness, and learn the functions representing the values of these attributes. We present experimental results that the learned functions show very promising performances in the estimation of the attribute values.

1. Introduction

In this paper, we consider the problem of recognizing the surface quality of an object from its imagery. We will use the term *surface quality* to mean a group of sensations that humans receive from the surface of an object through vision and touch. It is firstly dependent on the physical properties of the surface, such as its material (e.g., metal, glass, plastic etc.), surface finish, softness, temperature, etc. It is also dependent on the properties that might be difficult to physically define, such as "feel" or "touch" of the object surface. Although such surface quality is the most closely related to tactile sensations, humans appear to be able to fairly accurately sense it only from visual inputs. This ability is considered to play an important role in everyday life.

The mechanism of how these sensations are processed in human brains remains mostly unclear, and is being studied in the field of vision sciences [1, 2, 3]. Studying the surface quality is also important from an engineering (application) point of view, as it might help, for example, analyze what constitutes photorealism of images synthesized by CG, or improve the look and feel of industrial products around us. Based on these motivations, researchers from brain science, psychology, and computer vision have cooperatively started several approaches recently.

In this study, we consider the problem of recognizing several attributes associated with the surface quality of an object from its single image. Although surface quality is a fairly abstract concept, as mentioned above, we select several attributes that are considered to contribute to the formation of the surface quality, such as glossiness, roughness, transparency etc. We then call them as surface-quality attributes, or simply attributes, and regard them as the target of recognition. To be specific, choosing the attributes taking a continuous value, we consider the problem of estimating the value of each attribute from imagery.

There exist a few studies of methods for recognizing such surface-quality attributes from images. Dror et al. [4] present a method that directly learns relations between image features and several attributes and estimate the attribute values for a given image. In the field of vision science, there are a number of studies that investigates what kinds of cues biological vision uses to recognize such attributes [1, 2] or analyzes what parts of human brains are used to do this by using fMRI [3].

In the study of Dror et al. the images are synthesized by CG, where some surface reflectance model such as the Torrance-Sparrow model is chosen, and are used for training and testing a classifier. The problem is then formulated as estimating a particular parameter of the chosen reflectance model, from a given image; the parameter is regarded as a surface-quality attribute itself. This approach greatly simplifies the problem, as it is easy to synthesize images of an object for which a chosen attribute has an arbitrary value. This makes it possible to have a large number of pairs of an image and an attribute that can be used for training the classifier.

However, this approach of using CG synthesized images has several problems.

Firstly, it can only deal with simple re-

978-4-9906441-1-6 ©2012 IAPR

flectance/appearance models; thus, the resulting images are simple, as well. In this approach, one has to find a parameter in the assumed reflectance model that is translatable to a surface-quality attribute. In the case of complex models, which enables the synthesis of photorealistic images, it is unclear to relate each surface-quality attribute to which parameter of the models.

Secondly, it is difficult to deal with surface-quality attributes that are shared by different materials. Although surface quality has a close connection to materials, it is natural to think of it as a concept defined at a higher level than materials. For example, glossiness, softness, etc. are general attributes that can be used for all sorts of materials. In the case of CG synthesized images, it is hard to relate the parameters in two different models for different materials, as each material tends to need a particular reflectance model.

Thirdly, the approach lacks a direct relation between what the trained classifier recognizes and what humans recognize. The trained classifier merely recognizes a parameter of the assumed reflectance model. It could be significantly different from what humans recognize.

Based on these considerations, we consider the problem of recognizing surface qualities from natural images. In the case of natural images, we can no longer create samples by simply varying parameters of reflectance models. A solution would be giving a value to each sample by hand and creating a set of pairs of an image and an attribute value, as is done in the studies of object recognition. However, it is quite hard for us humans to answer an absolute value of a particular attribute. This can be understood when considering a scenario in which we are shown an image and asked to answer its glossiness in the range of 0 to 100. Even if we manage to do this, it will be necessary to normalize the results to compensate for drifts within a person and differences among individuals.

In this study, inspired by the work of Parikh and Grauman [5], we propose to create training data by ranking pairs of images in terms of the target attribute and then train an *attribute function* which represents the strength of the target attribute using the created samples. To be specific, each training sample gives relative information between a pair of images, that is, "image A is greater than image B in terms of glossiness." It is much easier for us to make such a relative decision than to tell an absolute value in terms of an attribute. We then train an attribute function of the target attribute using methods of learning to rank. In this paper, we consider only surface-quality attributes taking a continuous value, and exclude those taking a binary value, such as either natural or man-made.

2. Approach

2.1. Learning an Attribute Function from Relative Information

In this section, we describe how to learn an attribute function $f(\boldsymbol{x})$, which computes the strength of the target attribute of an object from its image, or strictly its representation $\boldsymbol{x}$ as a feature. We use a set of relative information in terms of the target attribute between a pair of images for learning an attribute function. The overall procedure of learning is basically the same as Ranking-SVM [6], which converts learning to rank based on relative information into a classification problem and trains a function representing the ranking by SVM.

Let $\mathcal{I} = \{\boldsymbol{I}_1, \boldsymbol{I}_2, ..., \boldsymbol{I}_n\}$ denote the set of training images, and $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n\}$ denote the set of feature vectors extracted from $\mathcal{I}$. Each feature vector $\boldsymbol{x}_i$ is a d-dimensional vector extracted from $\boldsymbol{I}_i$ by using the methods that will be described later. Let $f : R^d \to R$ denote an attribute function of the target attribute, which computes the strength of the attribute from a given feature vector $\boldsymbol{x}$. Let $\mathcal{O} = \{(s_1, t_1), (s_2, t_2), ..., (s_m, t_m)\}$ denote the set of relative information given as training data. Each sample (s_j, t_j) indicates "$\boldsymbol{I}_{s_j}$ is greater than $\boldsymbol{I}_{t_j}$ in terms of the attribute". This implies that f should satisfy the inequality

$$f(\boldsymbol{x}_{s_j}) > f(\boldsymbol{x}_{t_j}). \tag{1}$$

The goal of learning is to obtain f that satisfies as many inequalities given by $\mathcal{X}$ and $\mathcal{O}$ as possible.

In this study, we use the following linear function for the attribute function f:

$$f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}. \tag{2}$$

Then Eq.(1) is rewritten as

$$\boldsymbol{w}^\top (\boldsymbol{x}_{s_j} - \boldsymbol{x}_{t_j}) > 0. \tag{3}$$

To find $\boldsymbol{w}$ that satisfies Eq.(3), we introduce slack variables $\xi_{s,t}$'s and consider the following optimization problem:

$$\min_{\boldsymbol{w}} |\boldsymbol{w}|_1 + C \sum_{(s,t)\in\mathcal{O}} \xi_{s,t}^2 \tag{4}$$

$$\text{s.t.} \quad \forall\mathcal{O} \ \xi_{s,t} \geq 0, \ \boldsymbol{w}^\top (\boldsymbol{x}_s - \boldsymbol{x}_t) \geq 1 - \xi_{s,t}.$$

This is the same problem as the one solved by L1-regularized L2-loss SVM. Thus, learning f based on $\mathcal{X}$ and $\mathcal{O}$ reduces to learning a linear binary classifier from features $\{\boldsymbol{x}_j^+ = \boldsymbol{x}_{s_j} - \boldsymbol{x}_{t_j}, \boldsymbol{x}_j^- = \boldsymbol{x}_{t_j} - \boldsymbol{x}_{s_j}\}$. We use LIBLINEAR [7] to solve Eq.(4).

2.2. Creation of training data

We use a part of Flickr Material Database [8] (FMD) to create a data set for our experiments. FMD, which

Table 1. Recognition accuracies with various image features.

attributes	SIFT-BoVW	SIFT-BoVW (dense)	RGB-BoVW	Neural Model in [1]	SIFT-BoVW and RGB-BoVW	SIFT-BoVW(dense) and RGB-BoVW
glossiness	**0.78**	0.77	0.57	0.74	0.69	0.73
transparency	0.77	0.78	0.66	0.75	**0.81**	0.81
smoothness	0.84	**0.87**	0.72	0.75	0.83	0.83
coldness	**0.72**	0.72	0.54	0.65	0.66	0.69

is created for the study of material recognition, consists of 1000 images of ten different materials (100 for each) that are collected from Flickr. We choose the five materials: metal, glass, plastic, stone, and wood, which gives 500 images in total. The backgrounds are removed by using the segmentation masks supplied in FMD and are not used in the subsequent feature extraction stage. These 500 images are split into two sets of 250 images, and one is used for training and the other for testing.

For each set of 250 images, we manually created sets of relative information by the following procedures: (1) two images are randomly selected from the set and shown to a respondent, (2) the respondent is asked to answer which image is greater in terms of a particular surface-quality attribute. The four attributes, glossiness, transparency, smoothness, and coldness, are considered in this study. The respondent is also allowed to choose “impossible to rank”, when he or she thinks that the two images are equal or, that it is irrelevant to rank with the specified attribute. For each attribute, we created about 200 samples of such relative information for each of the training and testing sets of 250 images.

It should be noted that we do not use the samples of “impossible to rank”, as this is contrastive to the method of [5] which uses the samples of equally ranked pairs for training. Unlike the tasks considered in [5], in our task, the samples of “impossible to rank” do not always mean two images are equal in terms of the attribute, they could mean that the comparison of the two images is by itself irrelevant. This is why we ignore the samples of “impossible to rank”.

2.3. Image Features

It is expected that different types of image features are effective for recognizing different surface-quality attributes. For example, capturing highlight generated by specular reflections should be effective to recognize glossiness. Hence we test several types of image features in our experiments to investigate which image feature is better to recognize each attribute. The image features used are described as follows.

SIFT-BoVW Keypoints and their local descriptors are extracted by SIFT [9] from a luminance image. They are quantized by the standard procedures of BoVW [10] to form a histogram of the descriptors. Then it is used as the global feature of the input image. We employed this approach, since it is a baseline method that is widely used for various image recognition tasks. It is expected that it can capture the highlights and textures of surfaces, as the SIFT descriptor is based on the spatial derivatives of luminance images.

SIFT-BoVW (dense sampling) The image feature is created in the same way as above except that the local descriptors are densely sampled on a regular grid.

RGB-BoVW 6000 patches of 3×3 pixels are randomly sampled on a RGB image, and their histogram is created by BoVW in the 9-dimensional patch space. Then it is used as an image feature.

Model of Neural Mechanism for encoding a skewness Motoyoshi et al. reported that a skewness of a luminance histogram is correlated with the perception of surface glossiness, and proposed a model of the neural mechanism for approximately encoding it [1]. The model is represented as a nonlinear filter composed of linear filters, nonlinear functions and spatial pooling processes. We apply it to an image, and use the histogram of the output as an image feature.

SIFT-BoVW and RGB-BoVW An image feature is created by concatenating SIFT-BoVW and RGB-BoVW features. Concatenated feature of SIFT-BoVW (dense sampling) and RGB-BoVW is also used as an image feature.

Combined Features SIFT-BoVW and RGB-BoVW features are combined and used as an image feature. The combined feature is simply crated by concatenating the two different features.

3. Experimental Results

For the four attributes, glossiness, transparency, smoothness, and coldness, we evaluate the accuracies of estimation of their values by the learned functions with the image features described above. The estimation accuracy is measured by the error rate, which is obtained by dividing the number of samples for which

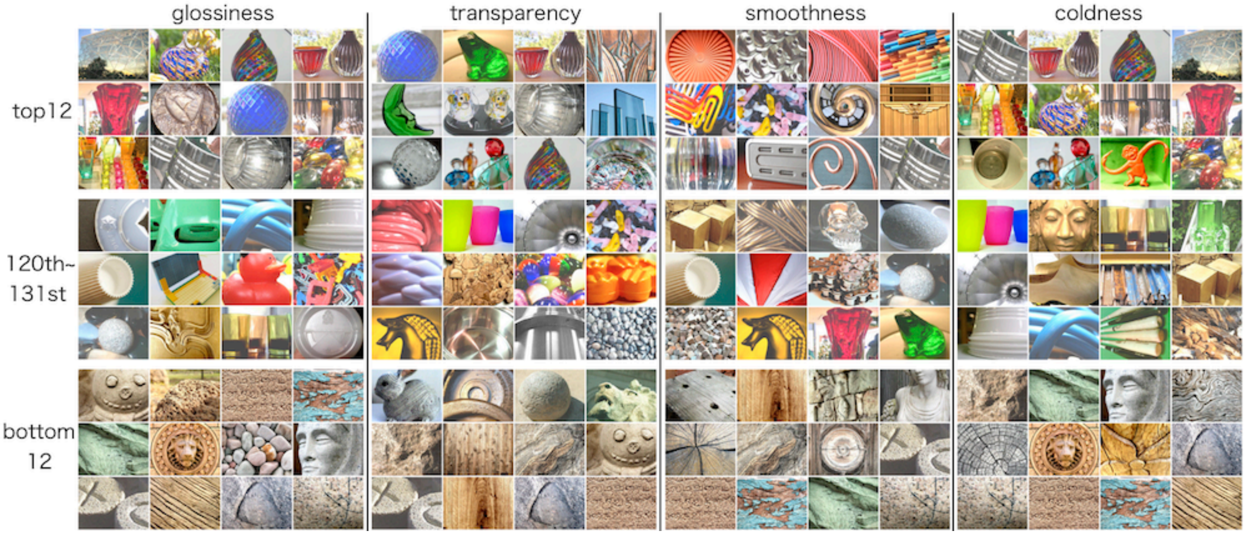

Figure 1. Testing images ordered by learned attribute functions.

the ranking is correctly reproduced by the attribute values computed by f with the number of samples in the testing data.

Tab.1 shows the result. The accuracies for the best image feature are by far above 0.5 (the chance rate) for each attribute. The results show that, as far as the features we tested are concerned, SIFT-BoVW is the best feature for the recognition of glossiness and coldness, and SIFT-BoVW (dense sampling) is the best feature for transparency, and the concatenated feature of SIFT-BoVW and RGB-BoVW is the best for transparency.

Fig.1 shows selected ordered lists of testing images sorted by the learned attribute functions of glossiness, transparency, smoothness and coldness. It can be observed that the images are ordered correctly with only a few exceptions. The erroneous estimations are, for example, the clearly opaque wooden object is ranked high (4th) and the transparent glass horse is ranked medium group in the list of transparency. These may be considered to be attributable to the fact that only simple image features are used here or that the size of training data is small.

4. Conclusion

In this paper we consider the problem of estimating surface-quality attributes of an object such as glossiness and transparency from its image. We have proposed to use a set of relative information to learn a function representing the strength of the target attribute based on learning to rank. The experimental results show that the proposed method is very promising; it can recognize glossiness, transparency, smoothness, and coldness with high accuracy, despite the fact that the size of training data is small and only simple features are used for the task. Future work includes making it possible to recognize other attributes and developing more effective image features.

Acknowledgement

This work was supported by KAKENHI 23135501.

References

[1] I. Motoyoshi, S. Nishida, L. Sharan, and E.H. Adelson. Image statistics and the perception of surface qualities. *Nature*, 447(7141):206–209, 2007.

[2] I. Motoyoshi. Highlight–shading relationship as a cue for the perception of translucent and transparent materials. *Journal of Vision*, 10(9), 2010.

[3] C. Hiramatsu, N. Goda, and H. Komatsu. Transformation from image-based to perceptual representation of materials along the human ventral visual pathway. *NeuroImage*, 57(2):482–494, 2011.

[4] R.O. Dror, E.H. Adelson, and A.S. Willsky. Recognition of surface reflectance properties from a single image under unknown real-world illumination. In *Proc. the Workshop on Identifying Objects Across Variations in Lighting at CVPR*, 2001.

[5] D. Parikh and K. Grauman. Relative Attributes. In *Proc. ICCV*, 2011.

[6] R. Herbrich, T. Graepel, and K. Obermayer. Large margin rank boundaries for ordinal regression. In *Advances in Large Margin Classifiers*, pages 115–132. MIT Press, 2000.

[7] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[8] C. Liu, L. Sharan, E. H. Adelson, and R. Rosenholtz. Exploring features in a Bayesian framework for material recognition. In *Proc. CVPR*, 2010.

[9] D.G. Lowe. Object recognition from local scale-invariant features. In *Proc. ICCV*, volume 2, pages 1150–1157. Ieee, 1999.

[10] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proc. ICCV*, pages 1470–1477, 2003.