ISASE 2019

Image Retrieval System Using Japanese Sound Symbolic Words for Surface Texture

Koichi YAMAGATA*, Tatsuki KAGITANI, and Maki SAKAMOTO**

The University of Electro-Communications Department of Informatics
1-5-1,Chofugaoka,Chofu,Tokyo182-8585,Japan
** koichi.yamagata@uec.ac.jp*
*** maki.sakamoto@uec.ac.jp*

Abstract: We propose a system to retrieve appropriate images from a database for a given Japanese sound symbolic word (SSW) expressing texture. We use some features of images calculated from gray-level co-occurrence matrices (GLCM). Using GLCMs and SSWs contribute reduction of calculation cost and easy operation of the system.
Keywords: *onomatopoeia, sound symbolism, texture impression*

1. INTRODUCTION

In recent research, it was found that Japanese sound symbolic words (SSWs) have a strong association with sensations and visual perception of surface texture [1]. For example, “nuru-nuru” indicates sliminess, while “zara-zara” indicates roughness. In the current study, we developed a system to retrieve images from a database by utilizing SSWs expressing surface textures.

There exist many texture feature analysis techniques. Among them, gray-level co-occurrence matrix (GLCM) is the most remarkable method for the extraction of textural features. GLCM is a second-order statistics methods, which is based on local information about gray levels in pair of pixels [2, 3]. The matrix is defined over the image with distribution of co-occurring values for a given inter pixel distance d as follows. Suppose an image I to be analyzed has n columns and m rows. Suppose that the gray level appearing at each pixel is quantized to g levels. The image I can be represented as a function that assigns some gray level in $\{1, \dots, g\}$ to each pair of coordinates. A co-occurrence matrix P_d defined over an $n \times m$ image I with respect to a given inter pixel distance d is

$$P_d(i,j) = \sum_{(\Delta x,\Delta y)\in\{-d,0,d\}^2\setminus(0,0)} P_{(\Delta x,\Delta y)}(i,j),$$

where

$$P_{(\Delta x,\Delta y)}(i,j) = \sum_{x=1}^{n}\sum_{y=1}^{m}\begin{cases}1, & if\ (I(x,y), I(x+\Delta x, y+\Delta y)) = (i,j)\\ 0, & otherwise.\end{cases}$$

Several statistic feature measures calculated from the co-occurrence matrix were introduced with the intent to describe the texture of the images [4,5]. The following equations define some of these features. Let $p(i,j)$ be the (i,j)th entry in a normalized GLCM. The mean and standard deviations for the rows and columns of the matrix are

$$\mu_x = \sum_{ij} i \cdot p(i,j), \qquad \mu_y = \sum_{ij} j \cdot p(i,j),$$
$$\sigma_x = \sum_{ij} (i-\mu_x)^2 \cdot p(i,j),$$
$$\sigma_y = \sum_{ij} (j-\mu_y)^2 \cdot p(i,j).$$

The features are as follows.

Energy (angular second moment):

$$X_1 = \sum_{ij} p(i,j)^2. \tag{1}$$

Entropy:

$$X_2 = -\sum_{ij} p(i,j)\log p(i,j). \tag{2}$$

Dissimilarity:

$$X_3 = \sum_{ij} |i-j|\, p(i,j). \tag{3}$$

Cluster Shade:

$$X_4 = \sum_{ij} (i+j-\mu_x-\mu_y)^3\, p(i,j). \tag{4}$$

Cluster Prominence:

$$X_5 = \sum_{ij} (i+j-\mu_x-\mu_y)^4\, p(i,j). \tag{5}$$

Maximum Probability:

$$X_6 = \max_{ij} p(i,j). \tag{6}$$

2. Materials and Methods

2.1 Task Design

We performed a psychophysical experiment using images from the Flickr Material Database (FMD) as visual stimuli [6].

In the experiment, cropped sections of FMD images were presented, and participants answered spontaneously and freely using sound symbolic words to describe the surface textures shown in the images. We analyzed participants' responses by evaluating sound symbolic words as quantitative adjectives [7].

2.2 Participants

The number of participants were 100 (25 women and 75 men, mean age = 20.6), and they were divided into 10 groups. They have no known impediment in speech or in vision. They were not informed of the purpose of the experiment.

2.3 Apparatus and Stimuli

The experimental stimuli used in this study were obtained from the FMD (http://people.csail.mit.edu/celiu/CVPR2010/FMD/) (Sharan et al. 2014 [51]), which is the image database developed for studying human material categorization. This database was constructed with the specific purpose of capturing a range of real-world appearances of common materials (e.g. glass, plastic etc.). Each image in this database (100 images per category, 10 categories) was selected manually from Flickr.com (under Creative Commons license) to ensure a variety of illumination conditions, compositions, colors, texture and material sub-types.

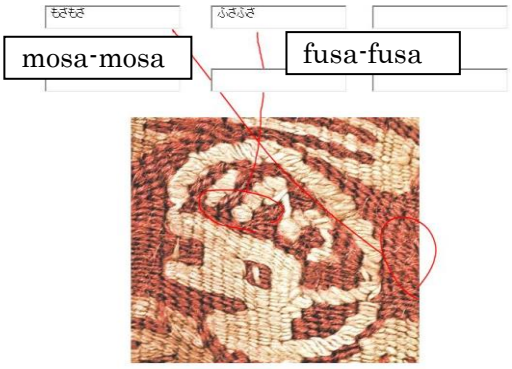

Fig. 1: An example of a questionnaire to crop sections of FMD images.

Because multiple objects and textures are included in FMD images, we conducted an experiment to create a new image dataset suitable for texture. We marked the part of the visual stimulus participants focused on when describing the surface texture. Ten participants participated in this experiment, we cropped each image section that three or more participants marked (Fig.1). Since the average size of the image sections marked by participants was approximately 100 pixels, we cropped square images of 150 × 150 pixels. Consequently, we obtained a total of 1,946 image samples, and classified them into 10 groups. Each group of visual stimuli was presented to each participant group.

2.4. Procedure

We conducted a psychophysical experiment to associate image textures with feelings. We used the cropped image stimuli in this experiment. During the test, participants were instructed to answer spontaneously and freely with 1–6 sound symbolic words expressing the texture of each material. An example answer is shown in Fig. 2. The sound symbolic word in the left cell is 'gowa-gowa,' which refers to a coarse and stiff texture. The sound symbolic word in the middle cell is 'zara-zara', which refers to a dry and rough texture.

In this experiment, each trial was conducted in an isolated test room under controlled lighting conditions. Participants were kept at a viewing distance of approximately 50 cm from a touch panel display showing the visual stimuli. The visual stimuli were presented vertically at eye-height, in a random order.

Fig. 2: An example of a questionnaire to associate image textures with feelings using cropped FMD images.

2.5. Data analysis

In this study, we develop a system to associate image features with SSWs expressing textures. This system enables us to retrieve images from database by SSWs. The overview of our system is shown in Fig. 3. Our system consists of three modules. The first one converts SSWs into quantitative adjectives. The second one converts features of images into quantitative adjectives. The third one compares SSWs and images by calculating cos similarities of quantitative adjectives given by above two modules. Let us explain in details below.

In the first module, a given SSW inputted in the text field is converted into a 43-dimensional vector expressing a texture impression. This module is based on a system proposed by Doizaki et al. (2017), which estimates the fine impression of sound symbolic words [8]. In the system, when a word that intuitively expresses a texture is input into the text field, information close to evaluations

against the 26 pairs of touch adjectives is obtained based on an analysis of the sounds of the word. This method for quantifying qualitative data uses quantification theory class I (a type of multiple regression analysis). Further, Kwon et al. expanded the paired adjectives to 43 pairs of adjectives (see Table 1) which is more suitable to describe the visual perception of the texture of objects [1].

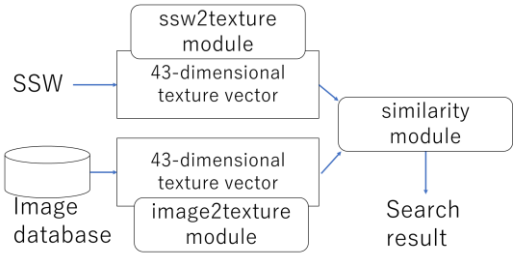

Fig. 3: The overview of surface texture retrieval system which retrieves images by input of retrieval SSWs.

Table 1: List of 43 adjective pairs

Y1	bright - dark	Y23	simple- complex
Y2	warm - cool	Y24	like - dislike
Y3	thick - thin	Y25	slippery - sticky
Y4	relieved - uneasy	Y26	sharp - dull
Y5	good - bad	Y27	static - dynamic
Y6	impressive- unimpressive	Y28	fashionable- unfashionable
Y7	happy - sad	Y29	pleasant - unpleasant
Y8	stable - unstable	Y30	masculine - feminine
Y9	comfortable- uncomfortable	Y31	elastic - nonelastic
Y10	hard - soft	Y32	glossy - nonglossy
Y11	regular - irregular	Y33	strong - weak
Y12	clean - dirty	Y34	bumpy - flat
Y13	modern- old_fashioned	Y35	smooth - rough
Y14	eccentric - ordinary	Y36	stretchy - nonstretchy
Y15	fresh - annoying	Y37	intense - calm
Y16	natural - artificial	Y38	loud - plain
Y17	familiar - unfamiliar	Y39	positive - negative
Y18	wet - dry	Y40	western_style- Japanese_style
Y19	sharp - mild	Y41	young - old
Y20	heavy - light	Y42	luxurious - austere
Y21	elegant - vulgar	Y43	repulsive - nonrepulsive
Y22	firm - fragile		

The second module converts 6-dimensional feature vectors of images into 43-dimensional adjective vectors by using a linear regression model

$$Y_i = a_i^1 X_1 + \cdots + a_i^6 X_6 + b_i, \qquad (1)$$

where $Y = (Y_1, \ldots, Y_{43})$ are rating values of adjectives introduced in Table 1, $X = (X_1, \ldots, X_6)$ are features of a given image calculated from GLCM P_d (see equations (1) $-$(10)), with coefficients a_i^j and intercepts b_i $(1 \le i \le 43, 1 \le j \le 6)$.

We performed this linear regression on the dataset obtained by the experiment of 1,946 cropped FMD images associated with some SSWs. The 42-dimensional adjective vectors for this regression is derived by translating SSWs into adjective vectors by using above SSW-adjective module and taking averages of them for each image. We set the inter pixel distance to 1 for calculation of GLCM, because the correlation coefficient between X_j and Y_i was relatively high in this case for most i and j. The standardized partial regression coefficients are shown in Table 2. Zeros in this table imply that each corresponding coefficient was not significant at 0.1% level of significance. We can see that X6 (Inverse Difference Moment) has large influence on most texture adjectives, particularly “unstable” and “dislike”. On the other hand, X5 (Homogeneity) has large influence on “stable” and “like”.

Table 2: Standardized partial regression coefficients

	X1	X2	X3	X4	X5	X6
Y1	-0.072	-0.294	0.59	-0.341	0.449	-0.941
Y2	-0.063	0.169	-0.065	-0.338	0.574	-1.127
Y3	0.048	0.527	-0.837	0.111	-0.327	0.466
Y4	-0.079	-0.082	0.223	-0.143	0.738	-1.149
Y5	-0.086	-0.149	0.316	-0.149	0.763	-1.163
Y6	0.069	0.405	-0.72	0.242	-0.614	0.983
Y7	-0.078	-0.127	0.329	-0.237	0.648	-1.119
Y8	-0.085	-0.113	0.232	-0.099	0.899	-1.289
Y9	-0.081	-0.151	0.295	-0.075	0.755	-1.097
Y10	0.08	0.066	-0.296	0.376	-0.635	1.162
Y11	-0.019	-0.456	0.742	-0.135	0.264	-0.545
Y12	-0.077	-0.288	0.51	-0.152	0.597	-0.961
Y13	-0.053	-0.304	0.617	-0.362	0.498	-1.022
Y14	0	0	0.124	0	0	-0.086
Y15	-0.068	-0.267	0.459	-0.101	0.687	-1.017
Y16	0	0	0	0	0	0
Y17	-0.064	-0.19	0.425	-0.24	0.622	-1.079
Y18	0	0	0.06	0	-0.393	0.222
Y19	0.054	-0.155	0.072	0.279	-0.524	0.977
Y20	0.056	0.485	-0.842	0.295	-0.348	0.697
Y21	-0.072	-0.218	0.36	-0.036	0.718	-0.998
Y22	0	0	0	0	0	0
Y23	-0.039	-0.347	0.591	-0.242	0.523	-0.979
Y24	-0.081	-0.104	0.236	-0.097	0.772	-1.146
Y25	-0.043	-0.341	0.503	0.057	0.513	-0.623
Y26	0	0	0	0	0	0
Y27	-0.064	-0.165	0.312	-0.18	0.706	-1.035
Y28	-0.049	-0.373	0.668	-0.221	0.49	-0.901
Y29	-0.065	-0.192	0.427	-0.311	0.537	-1.04
Y30	0.063	0.351	-0.702	0.405	-0.428	0.953
Y31	0	0	0.017	0	0.231	-0.425
Y32	-0.048	-0.407	0.788	-0.469	0.32	-0.878
Y33	0.074	0.292	-0.538	0.198	-0.669	1.005
Y34	0.04	0.519	-0.855	0.143	-0.401	0.663
Y35	-0.054	-0.321	0.662	-0.395	0.512	-1.07
Y36	-0.04	-0.072	0.292	-0.351	0.527	-1.056
Y37	0.075	0.251	-0.525	0.324	-0.665	1.191
Y38	0	0	0	0	0	0
Y39	-0.067	-0.279	0.548	-0.29	0.419	-0.876
Y40	-0.05	-0.271	0.555	-0.372	0.387	-0.869
Y41	-0.054	-0.354	0.683	-0.361	0.37	-0.877
Y42	-0.05	-0.198	0.397	-0.104	0.674	-1.07
Y43	0.068	0.327	-0.573	0.105	-0.631	0.891

The third module calculate cosine similarities of 43-dimensional adjective vectors between a given SSW and all images in database. A cosine similarity between

adjective vectors $v, w \in \mathbb{R}^{43}$ is

$$s(v, w) = \frac{v \cdot w}{|v||w|}.$$

All images in database are sorted in descending order by the cosine similarities. Fig.4 shows an example output of our image retrieval system when input SSW is pikapika. The shown images in this figure are ordered by the similarities. The upper left image has the highest similarity to pikapika.

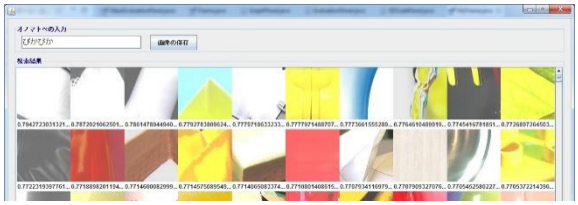

Fig.4: Example output of our image retrieval system when input SSW is pikapika. Shown images are ordered by the cosine similarities. The upper left image has the highest similarity to pikapika.

3. System Evaluation

To evaluate the accuracy of our system, we conducted a psychological experiment to 8 Japanese undergraduates aged 21 to 26 (6 males and 2 females). Participants were asked to evaluate the top 10 output images of our image retrieval system when input SSWs were “sukkiri”, “mossari”, “sikkari”, “sara”, and “turun” on a scale from 1 to five. These 5 SSWs were not used to construct SSW-adjective module. Each output image is regarded as correct if its average score is better than 3.5. The accuracy rates for “sukkiri”, “mossari”, “sikkari”, “sara”, and “turun” were 1.0, 0.1, 0.1, 0.9, and 0.7 respectively. The low accuracies of “mossari” and “sikkari” were caused by weak relations of these SSWs to texture impression. On the other hand. other SSWs had better accuracies than 0.7. This result shows that our system can retrieve appropriate image for SSWs expressing texture.

4. Conclusion

This study proposed a method to retrieve appropriate images from database for a given SSW expressing texture. The output images are ordered by cosine similarities of 43-dimmensional adjective vectors between images and the given SSW. The adjective vectors of images are obtained by a linear regression model with 6 predictor values calculated from GLCM: angular second moment, contrast, dissimilarity, entropy, homogeneity, and inverse difference moment. The accuracy rates for input SSWs expressing texture were better than 0.7. The calculation of GLCM of an image is not so heavy, and inputting SSWs is easy. Therefore, our system is expected to contribute to quick and easy image retrievals.

ACKNOWLEDGMENTS

This work was supported by a Grant-in-Aid for Scientific Research on Innovative Areas Shitsukan (No. 23135510 and 25135713) from MEXT Japan and JSPS KAKENHI Grant Number 15H05922 (Grant-in-Aid for Scientific Research on Innovative Areas Innovative SHITSUKSAN Science and Technology) from MEXT, Japan.

REFFERENCES

[1] Jinhwan Kwon, Tatsuki Kagitani and Maki Sakamoto; Holistic Processing Affects Surface Texture Perception: Approach from Japanese Sound Symbolic Words, Journal of Cognitive Science; 18(3), pp.321-340, 2017.

[2] Meng Liang; 3D co-occurrence matrix based texture analysis applied to cervical cancer screening, Department of Information Technology; UPPSALA UNIVERSITET, 2012.

[3] rafel C. Gonzalez and Richard E. Woods; Digital Image Processing, Pearson Education; Inc. Upper Saddle River, new Jersey 07458; Third Edition, pp.830-836, 2008.

[4] Robert M. Haralick; Statistical and structural approaches to texture, Proc. IEEE; vol. 67, no. 5, pp. 786 - 804, 1979.

[5] D. O. Aborisade, J. A. Ojo, A. O. Amole and A. O. Durodola; Comparative Analysis of Textural Features Derived from GLCM for Ultrasound Liver Image Classification, International Journal of Computer Trends and Technology; 11(6), pp.239-244, 2014.

[6] L. Sharan, R. Rosenholtz and E. H. Adelson; Material Perception: What Can You See in a Brief Glance?, Journal of Vision; 14(9), 12, 2014.

[7] M. Sakamoto, J. Yoshino, R. Doizaki and M. Haginoya; Metal-like Texture Design Evaluation Using Sound Symbolic Words, International Journal of Design Creativity and Innovation; 4(3-4), pp.181-194, 2015.

[8] R. Doizaki, J. Watanabe and M. Sakamoto; Automatic Estimation of Multidimensional Ratings from a Single Sound-symbolic Word and Wordbased Visualization of Tactile Perceptual Space, IEEE Transactions on Haptics; vol. 10, no. 2, pp. 173–182, 2017.