九州大学大学院総合理工学府報告
第 43 巻 第 2 号 42-48 頁 令和 4 年 2 月

Engineering Sciences Reports, Kyushu University
Vol. 43, No. 2, pp. 42-48, FEB, 2022

Discrimination of Complex Odors with Gas Chromatography-Mass Spectrometry Data by Texture Image Analysis and Machine Learning

Chaiyanut JIRAYUPAT[*1,2] Kazuki NAGASHIMA[*2,3†] Takuro HOSOMI[*2,3]
Tsunaki TAKAHASHI[*2,3] Wataru TANAKA[*2]
Masaki KANAI[*4] and Takeshi YANAGIDA[*2,4†]
[†]E-mail of corresponding author: kazu-n@g.ecc.u-tokyo.ac.jp, yanagida@cm.kyushu-u.ac.jp

(Received January 26, 2022, accepted February 1, 2022)

Conventional odor discrimination is generally performed by gas chromatography–mass spectrometry (GC–MS) that identifies specific marker molecules. Such marker identification process is, however, labor-intensive, and the limited number of identified marker molecules is often insufficient to discriminate complex odors. In this study, we have demonstrated a facile method for discriminating complex odors with GC–MS data by combining texture image analysis (TIA) and machine learning (ML). We extracted various texture features (*i.e.*, contrast, energy, homogeneity, correlation, dissimilarity and angular second moment) of two-dimensional (2D) MS maps by TIA, and used them as datasets for ML. Each texture feature contains a lot of molecular information appeared in 2D MS maps, and thus serves as an effective parameter for discriminating complex odors. Based on this method, we successfully performed the discrimination of breath samples collected from the persons of different blood glucose levels with higher performances and reliability than the conventional approach.

Key words: *Odor discrimination, texture image analysis, machine learning, 2D MS map, GLCM*

1. Introduction

Odor analysis is a promising technique for non-invasively characterizing a subject based on the species and the concentrations of contained volatile chemical compounds. This type of analysis has recently attracted much attention in various scientific and industrial fields such as pathology,[1-3] pharmacology,[4,5] healthcare,[6-12] food industry,[13-15] fragrance and perfume industry,[16-18] environmental conservation,[19-21] agriculture[22-24] and so on.[25-27] The odor analysis is performed by two-step process consisting of i) a marker molecules identification and ii) a discrimination or classification of odors based on the identified specific markers. In odor analysis, GC–MS is conventionally employed for identifying marker molecules.[28-33] However, the marker identification process is labor-intensive, and limited number of identified marker molecules is often insufficient to discriminate complex odors.

The texture image analysis (TIA) is a useful way to effectively collect large amount of information in an image.[34,35] TIA has recently been applied to medical image analyses and successfully demonstrated its performances on the tumor identification and the radiotherapy beyond the sense of human eyes.[36,37] Despite the advantage of TIA, it has rarely been applied to odor discrimination. These backgrounds motivated us to investigate the applicability of TIA to the discrimination of complex odors.

In this study, we demonstrated the discrimination of human breath samples with GC–MS data by combining texture image analysis and machine learning (TIA–ML). In this method, various texture features were extracted from two-dimensional (2D) MS maps. Each texture feature contains a lot of

*1 Department of Molecular and Material Sciences, Interdisciplinary Graduate School of Engineering Sciences, Kyushu University, Graduate Student
*2 Department of Applied Chemistry, Graduate School of Engineering, The University of Tokyo
*3 Japan Science and Technology Agency, PRESTO
*4 Division of Integrated Materials, Institute for Materials Chemistry and Engineering, Kyushu University

molecular information appeared in 2D MS maps, and thus serves as an effective parameter for discriminating complex odors. Based on this method, we successfully performed the discrimination of breath samples collected from persons with different blood glucose levels. The performance and reliability of the TIA–ML method were discussed in comparison with those of a conventional marker identification approach.

2. Experimental

2.1 Collections of Breath Samples and Blood Glucose Data

The human breath samples were collected from healthy volunteers under fasting condition (8–10 h). To control the blood glucose levels, the volunteers took a 150 mL aqueous solution of 50g glucose (TRELAN-G50, AY Pharmaceuticals). The blood glucose level of volunteers was measured by a glucometer with conventional fingerstick method and a flash glucose monitoring system (FreeStyle Libre, Abbott). Each 50 breath samples were collected from the persons with high blood glucose level (HBG, ≥ 125 mg/dL) and low blood glucose level (LBG, < 120 mg/dL). The exhaled breath was collected using a 10 L gas sample bag (Smart bag PA, GL Sciences). The 500 mL of collected breath was then transferred to an adsorbent-filled tube (Packed Liner with Tenax GR, mesh 80/100 #2414-1021, GL Science Inc.) using an air pump at the pumping rate of 50 mL/min. The sample tubes were sealed and stored in a refrigerator at –18 °C until they were used for the GC–MS measurements.

2.2 Breath Component Analysis by GC–MS

Total ion current (TIC) chromatograms and MS chromatograms of the breath samples were obtained by GC–MS (Shimadzu, GCMS-QP2020) equipped with an inlet temperature control unit (OPTIC). A InertCap 5MS/NP capillary column (60 m length, 0.25 mm inner diameter, 1 µm thickness, GL Sciences) was used, and the temperature profile of GC oven was set as follows: (i) held at 40 °C for 5 min, (ii) elevating to 280 °C at a rate of 5 °C/min, and (iii) held at 280 °C for 5 min. The inlet temperature was set at 300 °C with split-less mode. The temperatures of the ion source and the GC-to-MS junction were both set at 200 °C. The vacuum pressure in the ionization chamber was 9.9×10^{-5} Pa. He (99.9999% pure) was used as a carrier gas in column and a purge gas, and the flow rates were set at 1 mL/min and 5 mL/min, respectively. The MS measurements were carried out with a single quadrupole MS analyzer in a mode of electron ionization with positive ion analysis and the full scan data acquisition. A mass to charge ratio (m/z) was characterized in the range of 35–300. The obtained data were analyzed by GCMS Solution ver. 4.45 SP1.

2.3 Texture Image Analysis and Machine Learning of 2D GC–MS Data

GC–MS data was analyzed by the TIA–ML method and the conventional marker identification approach. The workflows of the TIA–ML method and the conventional marker identification approach are shown in Figs.1(a) and (b). For the TIA–ML method, firstly, all MS chromatograms, *i.e.*, the series of retention time–signal abundance data, were combined and converted into a 2D MS map as the functions of m/z (x-axis) and retention time (y-axis). The range of m/z and retention time used for analysis were 35–300 and 3–58 min, respectively. The image resolution of 2D MS map was set to be 1300 × 3700 pixels. The intensity of 2D MS map was scaled by a power law (γ = 0.5), displayed by 256 colors, and normalized via the highest peak using Matplotlib ver. 3.5.1. To investigate the robustness of the TIA-ML method, the influences of position alignment and noise reduction in 2D MS map were examined. The details of such image processing for 2D MS map can be seen in our previous study.[38)]

To extract texture features of the 2D MS map, TIA was performed with gray-level co-occurrence matrix (GLCM)[39)] using Scikit-image ver. 0.19.1. The GLCM functions characterize the textures of an image by calculating the number of pairs of pixels with specific values in a specified spatial distance, creating GLCM maps, and then extracting statistical texture feature values from the matrix of GLCM map. In this study, the distance of 1 pixel and the angle of 45° were used. GLCM maps of contrast, energy, homogeneity, correlation, dissimilarity, and angular second moment (ASM) were created from 2D MS map using the formulas shown in Table 1. Then the texture feature was obtained by a summation of the feature values of all pixels in a GLCM map. The extracted texture features were normalized and used as datasets

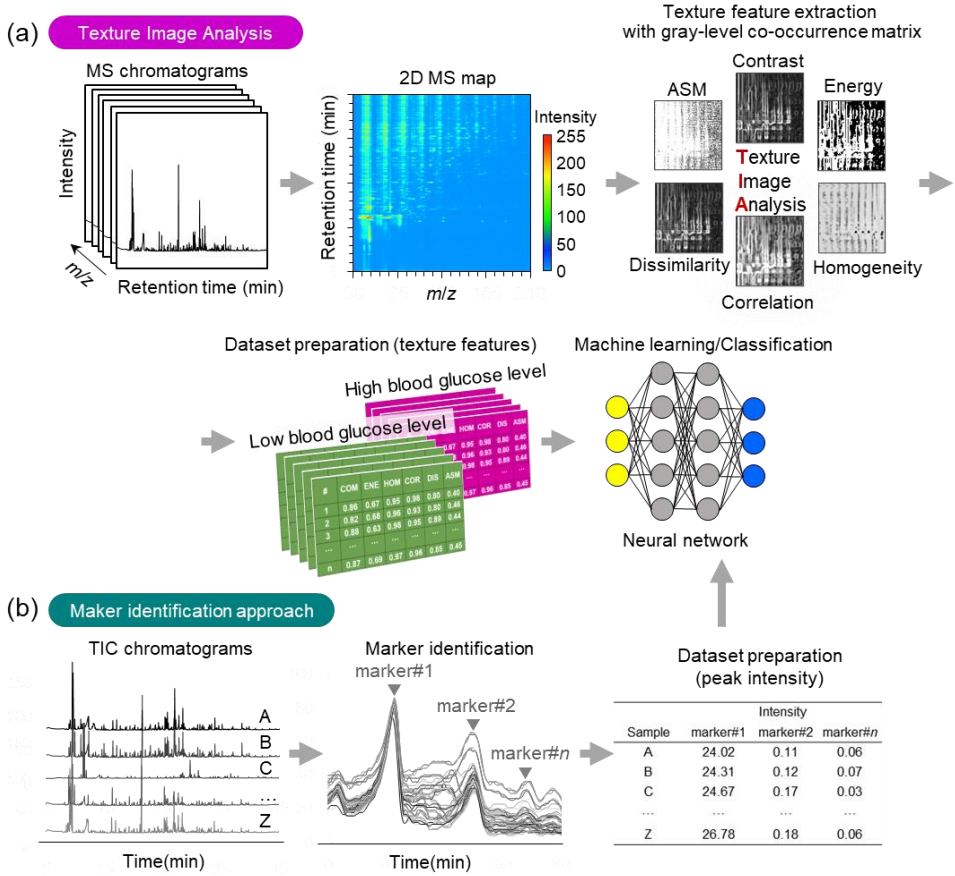

Fig.1 Graphical workflows of (a) texture image analysis (TIA) and (b) conventional marker identification approach for discriminating complex odors in GC–MS data.

for ML.

For conventional marker identification approach, peaks were detected in TIC chromatograms and their intensities were used as datasets for ML. The peak detection was performed by *CentWave* method with the parameters of ppm=10, peak width minimum=1, peak width maximum=2, snthresh=100, mzdiff=6, prefilter scan number=0.01, prefilter scan abundance=3, and bw=100, which were optimized by using the method reported by Manier *et al.*[40] In this study, for the simplicity, all detected peaks were used as the marker molecules for the discrimination of human breath samples, while the marker molecules are identified by carefully screening the detected peaks in the conventional odor discrimination study.

ML was conducted by a neural network algorithm. For ML, the datasets were divided into training data and testing data with a ratio of 70% and 30%, respectively. For enriching the training datasets while preventing overfitting, the data augmentation technique[38] was employed. In this technique, the intensity of 2D MS maps was randomly modulated in the range of 0.0–10.0%. Consequently, the number of data increased by 100 datasets. The two-levels classification of breath samples (*i.e.*, HBG and LBG) was performed with a multilayer perceptron, which is a class of feedforward artificial neural network, using Scikit-learn ver. 1.0.2. The classifiers were optimized by the hyper-parameters and operated with the parameters of hidden_layer_sizes = (128, 128), activation = 'relu', solver = 'adam', alpha = 1, max_iter = 1000 for the TIA-ML method and hidden_layer_sizes = (256, 512), activation = 'relu', solver = 'adam', alpha = 1, max_iter = 3000 for the conventional marker identification approach.

The odor discrimination performances in the TIA–ML method and the conventional marker identification approach were evaluated by calculating and comparing their classification accuracy, sensitivity, and specificity. The averaged area under the curve of receiver operating characteristic curve (AUC–ROC) was utilized to evaluate the reliability of

Table 1 Texture features and formulas

Texture feature	Formula
Contrast	$\sum_{i,j=0}^{N-1} P_{i,j}(i-j)^2$
Energy	$\sqrt{\sum_{i,j=0}^{N-1} (P_{ij})^2}$
Homogeneity	$\sum_{i,j=0}^{N-1} \frac{P_{ij}}{1+(1-j)^2}$
Correlation	$\sum_{i,j=0}^{N-1} P_{i,j}\left[\frac{(i-u)(j-u)}{\sqrt{(\sigma_i)^2(\sigma_j)^2}}\right]$
Dissimilarity	$\sum_{i,j=0}^{N-1} P_{i,j}\lvert i-j\rvert$
ASM (Angular second moment)	$\sum_{i,j=0}^{N-1} (P_{ij})^2$

classifier. The significance of each feature for the discrimination of human breath samples was evaluated with the *p*-value obtained in *t*-test.

3. Results & Discussion

3.1 2D MS Map and Texture Features of Human Breath Samples

Figs.2(a) and (b) show the representative TIC chromatograms and 2D MS maps of breath samples collected from the persons with two different blood glucose levels (*i.e.*, HBG, ≥ 125 mg/dL and LBG, < 120 mg/dL). TIC chromatogram is a primary form of GC–MS data used to identify marker molecules in the conventional approach. Each peak in the TIC chromatogram corresponds to a component molecule species in the tested breath, and each bright spot in the 2D MS map corresponds to a fragment peak of a component molecule species. Contrary to the TIC chromatograms, where the peak intensity can be quantitatively compared, the 2D MS maps were hardly distinguishable to the human eyes.

Next, we extracted the features of GC–MS data. The texture features of the 2D MS maps were extracted by TIA with GLCM. Fig.3 shows the GLCM maps for the breath samples collected from the persons of HBG and LBG. Each texture feature was then obtained by a summation of feature values of all pixels in a GLCM map. The extracted texture features

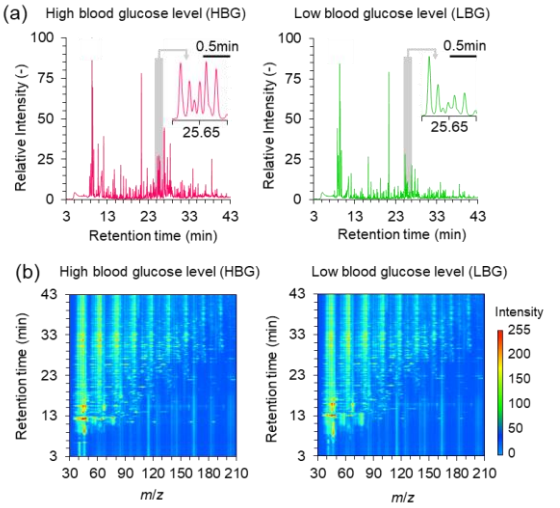

Fig.2 (a) TIC chromatograms and (b) 2D MS maps of breath samples collected from the persons with high blood glucose level (HBG, ≥ 125 mg/dL) and low blood glucose level (LBG, < 120 mg/dL), respectively. For visibility, the 2D maps are shown in the restricted range (*m*/*z*: 30–210, retention time: 3–43 min).

were normalized and used as datasets for ML. We created a classifier by ML with a neural network algorithm. As a comparison, we also created a classifier by the conventional marker identification approach. For this purpose, we identified the peaks of marker molecules in the TIC chromatograms, and the peak intensities were used as datasets for ML. By using the classifiers, we calculated the classification accuracies for the test breath samples.

3.2 Classification Performance of TIA-ML Method for Human Breath Samples

Fig.4(a) shows the classification accuracy of breath samples of HBG and LBG, plotted as a function of the number of features employed for creating the classifier. The employed features were arranged in ascending order of the *p*-values. In the conventional marker identification approach, the classification accuracy was as low as 20.0% when employing a single feature. It tended to increase with increasing the number of employed features and reached to 100% with 50 features. On the other hand, in the TIA–ML method, the classification accuracy was 83.3% with a single feature, and reached to 100% with two features. These results clearly indicated that that the TIA–ML method provided a higher classification accuracy with fewer features than the conventional marker identification approach. Note that both of the specificity and sensitivity of the TIA–ML method reached to

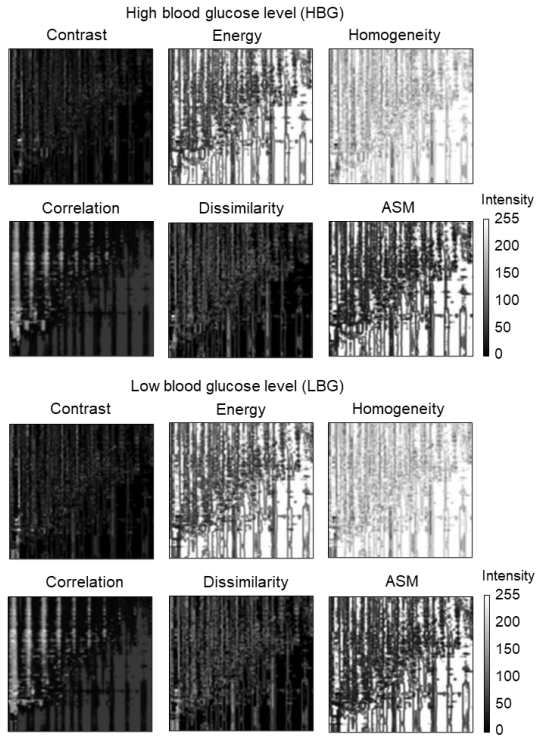

Fig.3 GLCM maps for texture features of contrast, energy, homogeneity, correlation, dissimilarity, and ASM.

100% with two features. Such excellent discrimination performance of the TIA–ML method can be interpreted by the fact that each texture feature contains a lot of molecular information.

3.3 Reliability of TIA-ML Method

To confirm the validity of above-mentioned classification performance, we evaluated the reliability of classifiers. Fig.4(b) shows the averaged AUC–ROC for the TIA–ML method and the conventional marker identification approach, plotted as a function of the number of features employed for creating the classifier. As well as the trends of classification accuracy in Fig.4(a), the AUC–ROC tended to increase by accompanying with the increase of the number of employed features. The AUC–ROC in the TIA–ML method reached to 1.00 with two features while that in the conventional marker identification approach was as low as 0.47. These results highlighted that the TIA–ML method showed better performances in both the accuracy and reliability for the discrimination of the human breath samples.

3.4 Advantage of Texture Feature

Here we discuss the contribution of each feature on the classification results in Fig.4(a).

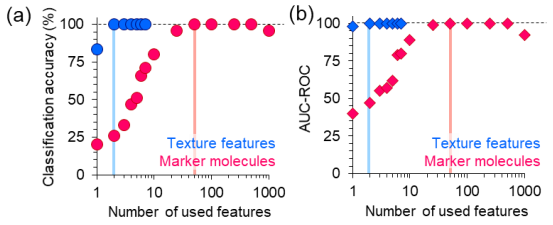

Fig.4 (a) The classification accuracy of the breath samples and (b) the averaged AUC-ROC for the TIA–ML method and the conventional marker identification approach, plotted as a function of the number of features employed for creating the classifier.

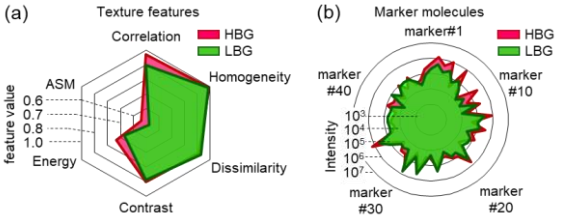

Fig.5 Radar charts for the feature values used in (a) the TIA and (b) the conventional marker identification approach, respectively. In these charts, the mean feature values of breath samples collected from the persons of HBG and LBG are plotted.

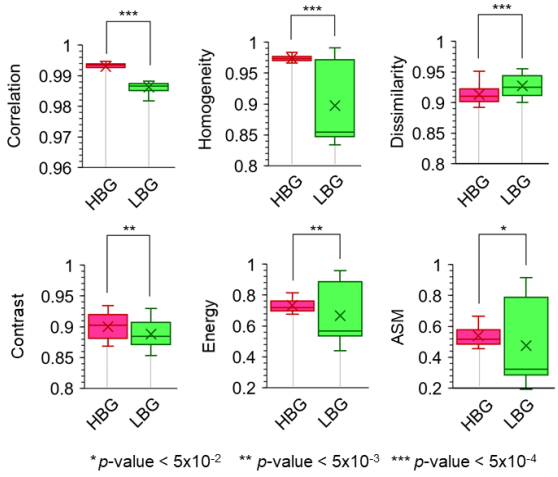

Fig.6 Box-and-whisker plots for the texture features of the breath samples collected from the persons of HBG and LBG.

In the conventional marker identification approach, the classification accuracy decreased with increasing number of features due to a so-called overlearning effect, in which the performance of classifier deteriorates by learning disturbing features. Interestingly, such an overlearning effect did not occur at all in the TIA–ML method. This indicates that all

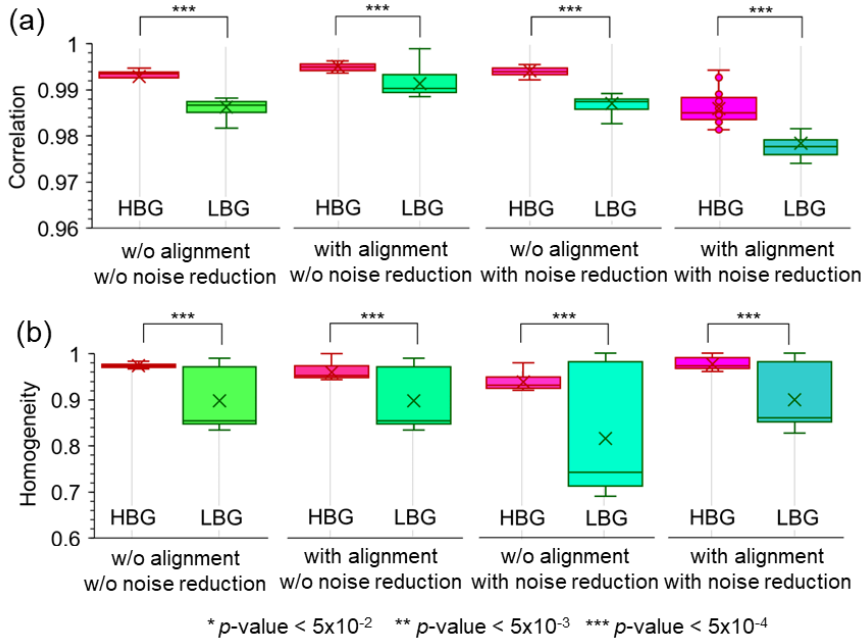

Fig.7 Box-and-whisker plots for the representative texture features (correlation and homogeneity) of the breath samples collected from the persons of HBG and LBG when performing the image processing (position alignment and noise reduction) with various combinations.

texture features positively contributed to the classification.

To gain an in-depth understanding as to the role of extracted texture features, we quantitatively compared their feature values on the human breath samples of HBG and LBG. Figs.5(a) and (b) show the radar charts of the feature values for the TIA–ML method and the conventional marker identification approach, respectively. In these charts, the mean feature values are plotted. Note that the features are arranged in ascending order of the *p*-values and displayed clockwise on the charts. We found that there was no clear relationship between the substantial difference in the feature values and the arrangement order of the features in the TIA–ML method. A similar trend was also found in the chart of the conventional marker identification approach. Fig.6 shows the box-and-whisker plots for the texture features in the human breath samples of HBG and LBG. The results showed that the distributions of feature values more clearly separated for the features of earlier order. On the other hand, the distributions of feature values overlapped in some texture features such as dissimilarity, contrast, energy and ASM. Considering the fact that no overlapping occurred in either the classification accuracy or the AUC–ROC, the classification error in each texture feature could be negligible in the assembled features (*i.e.*, combination of texture features). Such a beneficial role of texture feature might be available only when more maker molecules than contaminant molecules occupy the analyte odors.

3.5 Robustness of TIA–ML Method

Finally, we investigated the robustness of the TIA–ML method by examining the influences of position shift and noise in 2D MS maps. The position shift of spots occurs when a liquid phase of GC column deteriorates over time. The noise is caused by the deterioration of GC column and/or the contaminant molecules (*i.e.*, non-marker molecules) in analytes. Fig.7 shows the box-and-whisker plots for representative texture features of the human breath samples of HBG and LBG when performing the image processing. As examples, the effects of position alignment and noise reduction in correlation and homogeneity are shown. The essential importance of position alignment and noise reduction was demonstrated in our previous study for the identification of marker molecules by image analysis.[38] We found that the significance of each texture feature (*i.e.*, *p*-value) was in the almost same order, regardless of the position alignment and/or the noise reduction. These results are reasonably interpreted by the principle of TIA, where the spatial relationship

of pixels is emphasized in the texture features. It should be worth describing that the robustness to position shift and noise in TIC chromatograms is not available in the conventional marker identification approach. As such, the data analysis process in the TIA–ML method can be simpler than that in the conventional marker identification approach and thus the high-throughput discrimination of complex odors would be expected by the TIA–ML method.

4. Summary and Conclusion

We demonstrated a facile method for discriminating complex odors with GC–MS data by combining texture image analysis and machine learning (*i.e.*, TIA–ML method). In the proposed method, various texture features (*i.e.*, contrast, energy, homogeneity, correlation, dissimilarity and ASM) of 2D MS maps were extracted by TIA with GLCM and used as datasets for ML. Contrary to the conventional marker identification approach, which relies on the limited number of marker molecules, each texture feature contains a lot of molecular information appeared in the 2D MS map, and thus served as an effective parameter for discriminating complex odors. By the TIA–ML method, we successfully performed the discrimination of breath samples collected from the persons of different blood glucose levels with higher performances and reliability than the conventional marker identification approach. While this study was limited to a two-levels classification, the TIA–ML method is essentially applicable to a multilevel classification. Thus, we believe that the TIA–ML method paves a novel avenue in complex odor analysis.

Acknowledgments

This work was supported by KAKENHI (No. JP18H05243) and PRESTO Program (No. JPMJPR19J7) of Japan Science and Technology Corporation (JST). K.N. acknowledges JACI. This work was partly supported by the Cooperative Research Programs of “Dynamic Alliance for Open Innovation Bridging Human, Environment and Materials in Network Joint Research Center for Materials and Devices”, “Network Joint Research Center for Materials and Devices”. This study was performed with the approval of the Research Ethics Committee. We acknowledge Prof. Takao Yasui and Prof. Yoshinobu Baba in Nagoya University for their technical advices on human experiments.

References

1) H. Haick *et al.*, Chem. Soc. Rev., 43, 1423 (2014).
2) M. Hakim *et al.*, Chem. Rev., 112, 5949 (2012).
3) M. M. Ioana *et al.*, Medicina (Kaunas), 56, 118 (2020).
4) L. Schreiner *et al.*, J. Nat. Prod., 83, 834 (2020).
5) H. Wakayama *et al.*, Ind. Eng. Chem. Res., 58, 15036 (2019).
6) M. K. Nakhleh *et al.*, ACS Nano, 11, 112 (2017).
7) F. Decrue *et al.*, Anal. Chem., 93, 15579 (2021).
8) A. Z. Berna *et al.*, ACS Infect. Dis., 7, 2596 (2021).
9) T. Güntner *et al.*, ACS Sens., 4, 268 (2019).
10) G. Giovannini *et al.*, ACS Sens., 4, 1408 (2021).
11) Z. Li *et al.*, Anal. Chem., 93, 9158 (2021).
12) D. Maier *et al.*, ACS Sens., 4, 2945 (2019).
13) E. Guichard *et al.*, J. Agric. Food Chem., 68, 10318 (2020).
14) M. Bösl *et al.*, J. Agric. Food Chem., 69, 1405 (2021).
15) J. E. Grimm *et al.*, J. Agric. Food Chem., 67, 5838 (2019).
16) R. Toniolo *et al.*, Anal. Chem., 85, 7241 (2013).
17) M. A. Teixeira *et al.*, Ind. Eng. Chem. Res., 52, 963 (2013).
18) V. B. Xavier *et al.*, Ind. Eng. Chem. Res., 59, 2145 (2020).
19) J. M. Estrada *et al.*, Environ. Sci. Technol., 45, 1100 (2011).
20) J. Quintana *et al.*, Environ. Sci. Technol., 50, 62 (2016).
21) G. Ašmonaitė *et al.*, Environ. Sci. Technol., 52, 14381 (2018).
22) S. Giglio *et al.*, Environ. Sci. Technol., 42, 8027 (2008).
23) Y. Wang *et al.*, J. Agric. Food Chem., 53, 3563 (2005).
24) M. Bengtsson *et al.*, J. Agric. Food Chem., 49, 3736 (2001).
25) Q. Wang *et al.*, Sci. Rep., 10, 14856 (2020).
26) C. Wongchoosuk *et al.*, Sensors, 9, 7234 (2009).
27) S. K. Jha, Rev. Anal. Chem., 36, 20160028 (2017).
28) M. P. Styczynski *et al.*, Anal. Chem., 79, 966 (2007).
29) S. Zhang *et al.*, ACS Omega, 5, 26402 (2020).
30) B. Demarcq *et al.*, J. Agric. Food Chem., 69, 3175 (2021).
31) L. Sun *et al.*, J. Agric. Food Chem., 69, 9350 (2021).
32) Z. Jia *et al.*, ACS Omega, 3, 5131 (2018).
33) D. K. Trivedi *et al.*, ACS Cent. Sci., 5, 599 (2019).
34) E. R. Thaler *et al.*, Expert. Rev. Med. Devices., 2, 559 (2005).
35) N. Feizi *et al.*, Trends. Analyt. Chem., 138, 116239 (2021).
36) N. B. Bahadure *et al.*, Int. J. Biomed., 2017, 1 (2017).
37) T. S. Kumar *et al.*, Biomed. Signal Process. Control., 73, 103440 (2022).
38) J. Chaiyanut *et al.*, Anal. Chem., 93, 14708 (2021).
39) A. Kassner and R.E. Thornhill, Am. J. Neuroradiol., 31, 809 (2010).
40) S. K. Manier *et al.*, Drug Test. Anal., 11, 752 (2019).