Computers and Electronics in Agriculture 110 (2015) 233–240

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Strawberry cultivar identification and quality evaluation on the basis of multiple fruit appearance features

Kyosuke Yamamoto [a], Seishi Ninomiya [a,*], Yoshitsugu Kimura [b], Atsushi Hashimoto [b], Yosuke Yoshioka [c], Takaharu Kameoka [b]

[a] Graduate School of Agricultural and Life Sciences, The University of Tokyo, 1-1-1 Midori-cho, Nishi-Tokyo, Tokyo 188-0002, Japan
[b] Graduate School of Bioresources, Mie University, 1577 Kurima-machiya, Tsu, Mie 514-8507, Japan
[c] Graduate School of Life and Environmental Sciences, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan

ARTICLE INFO

Article history:
Received 23 December 2013
Received in revised form 12 October 2014
Accepted 19 November 2014

Keywords:
Cultivar identification
Image analysis
Quality evaluation
Strawberry

ABSTRACT

The appearances of agricultural products are important indices for evaluating the quality of commodities and the characteristics of different varieties. In general, the appearances are evaluated by experts based on visual observations. However, the concern regarding this method is that it lacks objectivity, and it is not quantifiable because it depends greatly on an empirical knowledge. In addition, agricultural products have multiple appearance features; therefore, several of them need to be analyzed simultaneously for correct evaluation of the appearance. In this study, we developed a new image analysis system that can simultaneously evaluate multiple appearance characteristics such as the color, shape and size, of agricultural products in detail. To evaluate the effectiveness of this system, we conducted quality evaluations and cultivar identification on the basis of cluster analysis, multidimensional scaling and discriminant analysis of the appearance characteristics. The results of the cluster analysis revealed that strawberries could be classified on the basis of their appearance characteristics. Furthermore, we were able to visualize the small differences in the appearance of the fruit based on multiple characteristics on a two-dimensional surface by performing multidimensional scaling. The results demonstrate that our system is effective for qualitative evaluations of the appearance of strawberries. The results of the discriminant analysis revealed that the accuracy of strawberry cultivar classification using 14 cultivars was <42%, when only single feature was used. However, the rate increased to 68% after combining the three features. These results indicate that our system exploits the advantage of analyzing multiple appearance characteristics.

1. Introduction

The appearances of agricultural products such as the color, texture, shape and size are important indices for evaluating the quality of commodities and the characteristics of varieties, and they affect the purchasing decisions of consumers (Brosnan and Sun, 2004). Therefore, appearance characteristics are important in breeding programs. In general, appearances are evaluated by experts based on visual observation, which is time-consuming and labor intensive (Clement et al., 2012). In addition, this method lacks objectivity and it is not quantifiable because it depends greatly on empirical knowledge (Brosnan and Sun, 2004).

To overcome these problems, several studies have been conducted to evaluate the external appearance of agricultural products using image analysis, particularly during quality assessments. In recent studies, automatic citrus grading was used to classify fruits on the basis of their appearance (López-García et al., 2010). Defects on the fruit surface were detected on the basis of statistical analyses of color information. In addition, k-nearest neighbor classification was performed using color information to evaluate the quality of hazelnut peeling (Pallottino et al., 2009). A cucumber grading system (Clement et al., 2013) applied an active contour method to analyze the shape of cucumbers, where the length and curvature

Abbreviations: NSDH, normalized spatial distribution histogram; CDE, color distribution entropy; I-CDE, improved CDE; EMD, earth movers distance; HCA, hierarchical cluster analysis; MDS, multidimensional scaling; PCA, principal component analysis; ANOVA, analysis of variance; LDA, liner discriminant analysis; H, Hue; S, saturation; V, value.

* Corresponding author at: Institute for Sustainable Agro-ecosystem Services, The University of Tokyo, 1-1-1 Midori-cho, Nishi-Tokyo, Tokyo 188-0002, Japan. Tel.: +81 42 463 1613; fax: +81 42 464 4391.
E-mail address: snino@isas.a.u-tokyo.ac.jp (S. Ninomiya).

http://dx.doi.org/10.1016/j.compag.2014.11.018
0168-1699/

were calculated by the active contour method, which were used as grading indices. Broken rice grains were detected (Lin et al., 2010) by evaluating the shapes of rice grains by the velocity representation method. New indices for melon fruit color evaluation were developed by methods based on content-based image retrieval (Yoshioka and Fukino, 2009). Compared with the traditional indices used for fruit color evaluation, these new indices could detect more color differences among cultivars.

Thus, several studies have been conducted to evaluate the external appearance of agricultural products. However, most of these studies focused on a single characteristic such as the color or shape. In general, a single characteristic is not sufficient to describe the appearance of agricultural products because they have multiple and complex features. For example, the color and/or size of a fruit with an ideal shape may be undesirable because of its immaturity. In this case, the shape, color and size should be analyzed simultaneously to evaluate the quality appropriately. Some studies evaluated three or more characteristics such as the color, shape and size of fruit (Singh et al., 1993), e.g., a tomato grading system (Sarkar et al., 1984; Jahns et al., 2001; Clement et al., 2012) and an automated strawberry grading system (Liming and Yanchao, 2010). However, they only evaluated simple features such as the dominant color of the fruit surface, fruit diameter ratios and the minimum curvature of the boundary, which are not sufficient to describe the complex appearance of agricultural products. Therefore, important information such as the color distributions on a fruit surface and subtle changes in fruit shape is ignored.

This study aimed to develop a new image analysis system to simultaneously evaluate multiple appearance characteristics of agricultural products in detail. This image analysis system had two components: image acquisition and image analysis. The image acquisition component utilized the previously developed chromatic image acquisition system (Yamamoto et al., 2011). We developed image analysis methods for evaluating images obtained using this image acquisition system. In addition, we conducted two analyses to evaluate the effectiveness of the system: quality evaluation and cultivar identification on the basis of fruit appearance. We used strawberries as the target crop.

A quality evaluation was conducted to visualize the relationships between the image analysis results and the actual appearance of fruit. On the basis of the visualization analysis results, we would discuss the possibility of performing strawberry fruit quality evaluations using our system.

Table 1
Strawberry cultivars used in this study. All of the cultivars were used for quality evaluation, and the cultivars marked with asterisks (*) were additionally used for cultivar identification. Some of cultivar ID is missing (e.g. 03), because some cultivars we grew did not produce any fruits to be harvested. "No." represents the number of strawberry samples of each cultivar.

Cultivar ID	Cultivar	No.	Cultivar ID	Cultivar	No.
01	Kurume 16*	28	22	Himiko	9
02	REGINA*	20	23	BELRUB	1
04	Kikyo 5*	12	25	Kurume 32	9
05	TROUBADOUR	15	26	SELVA	1
06	APPLEVER	17	27	GALA	2
07	FAIRFAX	3	28	Morioka 16	11
08	Tsukushi*	14	29	Koki	3
09	VOLA	3	30	Harunoka	14
10	Natsusaki*	16	31	Hokowase*	20
11	Yamagata 1	2	33	Red Pearl	6
12	DONNER	5	34	Kurume 48	17
13	Akashi*	9	35	Sagahonoka*	23
14	MARLATE*	18	36	Benihoppe*	20
15	TORO*	21	37	Amao*	16
18	Aiberry	4	39	Akihime*	13
20	CHANDLER*	37	40	Toyonoka	7
21	BHDL 17	5	41	Tochiotome	16

It is well known that each strawberry cultivar has distinctive appearance characteristics. However, these differences are quite minute; therefore, it is very difficult for non-professionals to recognize them. Thus, if our image analysis system could identify strawberry cultivars based on the appearance of the fruit, this would demonstrate that our system is capable of evaluating small differences that only professionals can recognize normally. Hence, we used cultivar identification to evaluate our system.

Although several studies have been conducted to sort and grade agricultural products by evaluating their appearances using image analysis, almost none of them discussed about utilization of the developed technologies for breeding purposes. Generally, features relevant to sorting and grading are related to the features measured in breeding programs, and hence such technologies can be used for the purpose of phenotyping. Therefore, we will also discuss about capability of our system for the utilization for breeding purposes based on the result of this study.

2. Materials and methods

2.1. Crop materials and image acquisition

This study used 34 strawberry cultivars, which were grown at the National Agriculture and Food Research Organization Institute of Vegetable and Tea Science (Mie, Japan). The cultivars and the numbers of strawberry samples are presented in Table 1. The strawberries were harvested during January and February in 2010. Images of the fruit surfaces were acquired within 72 h of harvesting using the chromatic image acquisition system developed in a previous study (Yamamoto et al., 2011).

To remove the background from acquired images, the images were converted from the RGB color space into the HSV color space ($0 \leq H \leq 179, 0 \leq S \leq 255, 0 \leq V \leq 255$) using OpenCV (Itseez, 2014). This confirmed that the saturation values were very different for strawberries and the background. Therefore, the images were binarized using a saturation value threshold of 45 to extract the strawberry fruit. The images backgrounds of which were removed, were used for the further analyses.

2.2. Image analysis methods

2.2.1. Color analysis

In this study, color histograms (Swain and Ballard, 1991) and the color distribution entropy (CDE) (Sun et al., 2006) were used to analyze the color of strawberries. The color information in an image can be analyzed quantitatively by combining color histogram and CDE, which describes the appearance frequency of colors and the color spatial information, respectively.

CDE is based on the normalized spatial distribution histogram (NSDH) and the definition of information entropy (Shannon, 1948). NSDH is based on the annular color histogram (Rao et al., 1999). Suppose that A_i is the set of pixels, where i is the color bin, $|A_i|$ is the number of pixels in A_i, and C_i is the center of gravity coordinate of pixels in A_i. N concentric circles referred to as annular circles are drawn, where C_i is the center. Suppose that $|A_{ij}|$ is the number of pixels inside the jth circle ($1 \leqslant j \leqslant N$). The NSDH of color bin i, P_i, is defined as follows:

$$P_i = \{P_{i1}, P_{i2}, \cdots, P_{iN}\} \tag{1}$$

$$P_{ij} = \frac{|A_{ij}|}{|A_i|} \tag{2}$$

The CDE of color bin i, E_i, is defined as follows:

$$E_i(P_i) = -\sum_{j=1}^{N} P_{ij} \log_2 (P_{ij}) \tag{3}$$

However, the CDE obtained using Eq. (3) has certain problems such as the symmetrical property of the entropy, when it is used to describe the distribution of pixels in an image. Thus, the improved CDE, I-CDE, was introduced (Sun et al., 2006). The I-CDE of color bin i, E'_i, is defined as follows:

$$f(j) = 1 + \frac{j}{N} \tag{4}$$

$$g(P_i) = 1 + \frac{\sum_{j=1}^{N}(P_{ij} \times j)}{N} \tag{5}$$

$$E'_i(P_i) = -g(P_i)\sum_{j=1}^{N} f(j)P_{ij}\log_2(P_{ij}) \tag{6}$$

Eq. (6) measures the distribution of pixels with color bin i. Since I-CDE is calculated based on the definition of information entropy, the pixels with color bin i are more dispersed as E'_i increases.

In this study, we modified I-CDE to make it suitable for analyzing an image of only one object. The original I-CDE was developed to describe an image that contains various objects; thus, the radius of the Nth annular circle was determined based on the distributions of colors in an image (Rao et al., 1999). In the modified I-CDE, the radius was determined based on the highest fruit diameter and the color-spatial information was normalized against the fruit size. The modified I-CDE value increases when the pixels are more dispersed relative to the fruit size.

A color distance between samples A and B is calculated as follows:

$$D_{\text{color}}(\text{A},\text{B}) = \sum_{i=1}^{n} \min\left(h_i^{\text{A}}, h_i^{\text{B}}\right) \times \frac{\min\left(E_i''^{\text{A}}, E_i''^{\text{B}}\right)}{\max\left(E_i''^{\text{A}}, E_i''^{\text{B}}\right)} \tag{7}$$

where, h_i^{A} and h_i^{B} are the color histograms of A and B for color bin i, respectively, and $E_i''^{\text{A}}$ and $E_i''^{\text{B}}$ are the modified I-CDE values of A and B of for color bin i, respectively. The variables used in color analysis are enumerated in Appendix A.

Before conducting the color image analysis, an image was converted into HSV color space. The H, S and V values were then rescaled to 0–15, 0–3, and 0–3, respectively, to reduce the sampling error on color caused by the fluctuation of the lighting condition and the computational cost for the very high dimension data. Thus, the number of colors was reduced to 256 ($0 \leqslant i \leqslant 255$). The number of color was determined as the sufficient number to describe color feature of strawberry fruits of different cultivar and quality in our preliminary experiment. Therefore, each fruit had 512 dimensions of data, i.e., a 256-dimensional color histogram and a 256-dimensional modified I-CDE.

2.2.2. Shape analysis

The shape of a strawberry was analyzed using a shape analysis tool based on the $r\theta\phi$ polar coordinates and tangent lines (Kondou et al., 1998, 2002). In the tool, the origin is set at the center of gravity of the target shape. Let θ be the angle between the x-axis and a vector from the origin to a point on the target shape, and let ϕ be the angle between the vector and tangent line at the point. Thus, the target shape can be described using parameters θ and ϕ, which do not depend on the size of the target. Therefore, similar shapes have closer θ and ϕ values. The relationship between the xy coordinate system and the $r\theta\phi$ coordinate system (Fig. 1) is defined as follows.

$$x = r \cdot \cos\theta \tag{8}$$

$$y = r \cdot \sin\theta \tag{9}$$

$$\frac{dy}{dx} = -\tan(\phi - \theta) \tag{10}$$

$$\frac{dr}{d\theta}\tan\phi + r = 0 \tag{11}$$

If we suppose that $\phi = \alpha\theta + \beta$, then Eq. (12) can be obtained.

$$r^{\alpha}\sin(\alpha\theta + \beta) = C \tag{12}$$

In this study, θ was measured every 0.1 radians. Thus, 63-dimensional shape data were obtained for each fruit.

The shape distance between strawberries was measured using the earth mover's distance (EMD) (Rubner et al., 2000). Thus, primarily EMD is a measure used to calculate the distance between two signatures, which is computed based on a solution to the Hitchcock transportation problem (Hitchcock, 1941). Using the shape analysis tool, strawberry shapes were described as shape signatures, which comprised the feature vector θ and its weight ϕ. We calculated the EMD between the shape signatures of two fruit and used this as the shape distance between the fruit.

2.2.3. Size analysis

In a similar study, the maximum horizontal diameter of a fruit was used as a size index for strawberries (Liming and Yanchao, 2010). However, different sized fruits sometimes have the same horizontal diameters because their shapes are highly variable, particularly in strawberries. Thus, we used the area of the projected sample (the number of pixels in each sample) as the fruit size index. The size distance between samples A and B is defined as follows:

$$D_{\text{size}}(\text{A},\text{B}) = \frac{S_{\text{A}} - |S_{\text{A}} - S_{\text{B}}|}{S_{\text{A}}} \tag{13}$$

where S_{A} and S_{B} are the number of pixels in samples A and B, respectively.

2.3. Statistical analyses of the appearance characteristics for strawberry quality evaluation and cultivar identification

2.3.1. Quality evaluation analyses based on the distance matrix among strawberry fruits

When specialists evaluate the qualitative appearance of agricultural products, they evaluate each characteristic individually before producing a comprehensive evaluation. Using the image analysis methods, three types of appearance distance among fruits were obtained. We used the sum of the distances as the comprehensive distance of the appearance characteristics among fruits. Thus, the relationships among fruits were represented as a

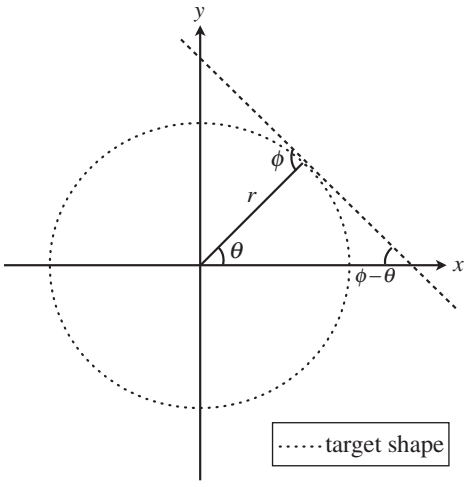

Fig. 1. The relationship between the xy coordinate system and the $r\theta\phi$ coordinate system.

distance matrix. We performed the following statistical analyses of the distance matrix: (i) hierarchical cluster analysis (HCA) to classify the strawberries on the basis of their appearance characteristics and (ii) multidimensional scaling (MDS) to visualize the appearance characteristics of strawberries on a two-dimensional surface. In addition, we calculated the following appearance characteristics of each strawberry fruit: the average fruit color values (L^*, a^*, b^*, H, S and V) calculated using OpenCV (Itseez, 2014), the shape distance generated from a circle model based on EMD and the fruit size. We calculated the average values of the appearance characteristics for each cluster obtained by HCA to compare the clusters. In addition, we performed a multiple linear regression analysis using the two-dimensional scores derived from MDS and the appearance characteristics to determine the meaning of each dimension. All of the statistical analyses were performed using R (R Developmet Core Team, 2011).

2.3.2. Cultivar identification using principal components analysis and liner discriminant analysis

Using the color and shape analysis methods, 512-dimensional color data (color histogram and CDE with 256 colors) and 63-dimensional shape data (ϕ measured every 0.1 radians) were obtained for each fruit. The data were subjected to principal components analysis (PCA) to summarize the color and shape data. Analysis of variance (ANOVA), Tukey's multiple comparisons test and linear discriminant analysis (LDA) of the principal component scores were then performed to investigate the possibility of cultivar identification using our image analysis system. We performed ANOVA and LDA using the standardized size data (the number of pixels). Instead of using separate datasets for training and testing, we performed a leave-one-out cross-validation to assess the performance of the discriminant functions obtained from the LDAs of the principal components and the size data. The accuracy of cultivar identification was evaluated as follows:

$$\text{Accuracy} = \frac{\text{number of fruit classified correctly}}{\text{total number of fruit}} \quad (14)$$

In these analyses, we used the 14 strawberry cultivars with larger samples among the cultivars we harvested (Table 1). All of the statistical analyses were performed using R (R Developmet Core Team, 2011).

3. Results and discussion

3.1. Quality evaluation of strawberry fruit on the basis of distances of the appearance characteristics among fruits

The results obtained with eight clusters are presented in Fig. 2, where strawberries with similar appearance tend to be classified into the same clusters.

Table 2 presents the average values for the: (i) fruit color (H, S and V), (ii) shape distances generated from the circle model on the basis of EMD and (iii) fruit size of each cluster. The results reveal that clusters 5 and 7 have low V values and the surface color of the strawberries in these clusters is dark red. In contrast, the V values of clusters 4 and 6 are high and they are light red. For the shape distances generated from the circle model, the strawberries in clusters 1 and 8 are close to the circle, whereas the strawberries in clusters 4, 6 and 7 are distant from the circle. Thus, the former clusters include rounded fruit and the latter include cone-shaped fruit. On the basis of average fruit size, clusters 4 and 8 contain small fruit, whereas clusters 1 and 5 contain large fruit. As described above, the strawberries were classified based on their appearance using HCA. These results suggest that strawberries could be sorted based on their quality using our approach because the appearance characteristics of strawberry fruit are important indices of their commodity value. Furthermore, strawberry fruits which looked similar to human eyes were classified into the same cluster. Therefore, we can expect that the appearance distances on color, shape and size we defined in this study is very close to a human sense and feasible. However, we have not compared these results with human sensory evaluation of multiple persons and need a further study to confirm it. In Japan, the standard appearance of strawberry fruits is defined for each grade and variety, and farmers evaluate each fruit appearance by comparing it with the standard appearance. Such evaluation is unstable because each person may give different evaluation on the same fruit, and even the same person sometimes gives different evaluation to the same fruit. In the future study, we need to conduct the comparison between the human sensory and our method's evaluations to show how stable our method is, and to assess the general applicability of our method for the quality evaluation of strawberry fruits.

The HCA results were highly dependent on the number of clusters, which was determined by the user. In this study, we selected eight clusters based on the results of our preliminary experiment, although this may change when different samples are used. In future research, the number of clusters should be determined objectively using statistical approaches (Jolion et al., 1991; Pelleg and Moore, 2000).

Fig. 3 shows a two-dimensional MDS map for the strawberries, where the appearance of strawberries changes in the first (x-axis) and second (y-axis) dimensions. The results of the multiple linear regression analysis showed that there was a significant correlation ($r = 0.950$, $p < 0.001$) with the first dimensional score, surface L^* value and fruit size. In addition, the second dimensional score was positively correlated with the a^* and b^* values for the fruit surface and fruit size ($r = 0.813$, $p < 0.05$). These results indicate that the first dimension is related to the brightness and size of fruit, whereas the second is related to the color and size of fruit. Thus, we visualized small differences in fruit appearance based on multiple characteristics.

Previously, the appearance of agricultural products has been evaluated qualitatively and subjectively by the naked eye. Thus, it has been difficult to visualize detailed differences in appearance, as presented in Fig. 3. Our approach is useful for several purposes. For example, we could use the MDS results to extract the best strawberry fruit from large volumes of harvested fruit. Fruit with similarities in appearance are closer in the MDS map, whereas the distance of fruit from the ideal appearance in the map can be used as a strawberry fruit appearance quality index. Desirable fruit can be extracted by setting a distance threshold. In addition, we determined the meaning of each dimension based on the results of the multiple linear regression analysis, so the dimensional scores of any other fruit can be calculated. Thus, fruits that were not used in the MDS analysis can be plotted in the map.

During breeding, the appearance distance between the fruit of existing cultivars and newly developed cultivars can be measured using our approach. Each strawberry cultivar has distinctive appearance characteristics and the desired features vary greatly depending on their usage. In addition, the consumer preferences for strawberry fruit appearance have specific trends. Therefore, the appearance distances from existing cultivars, including various popular and unpopular cultivars, can be used to determine the use and commodity value of newly developed cultivars. Furthermore, genetic information related to the appearance quality could be estimated by linking genetic information to the MDS map. The strawberry cultivars (*Fragaria* × *ananassa* Duch.) used in this study are octoploid, and those used in molecular genetic studies are still rarely conducted (Bianco et al., 2009). Our image analysis system provides a new method for phenotyping fruit appearance, which may contribute to genetic research and fruit breeding.

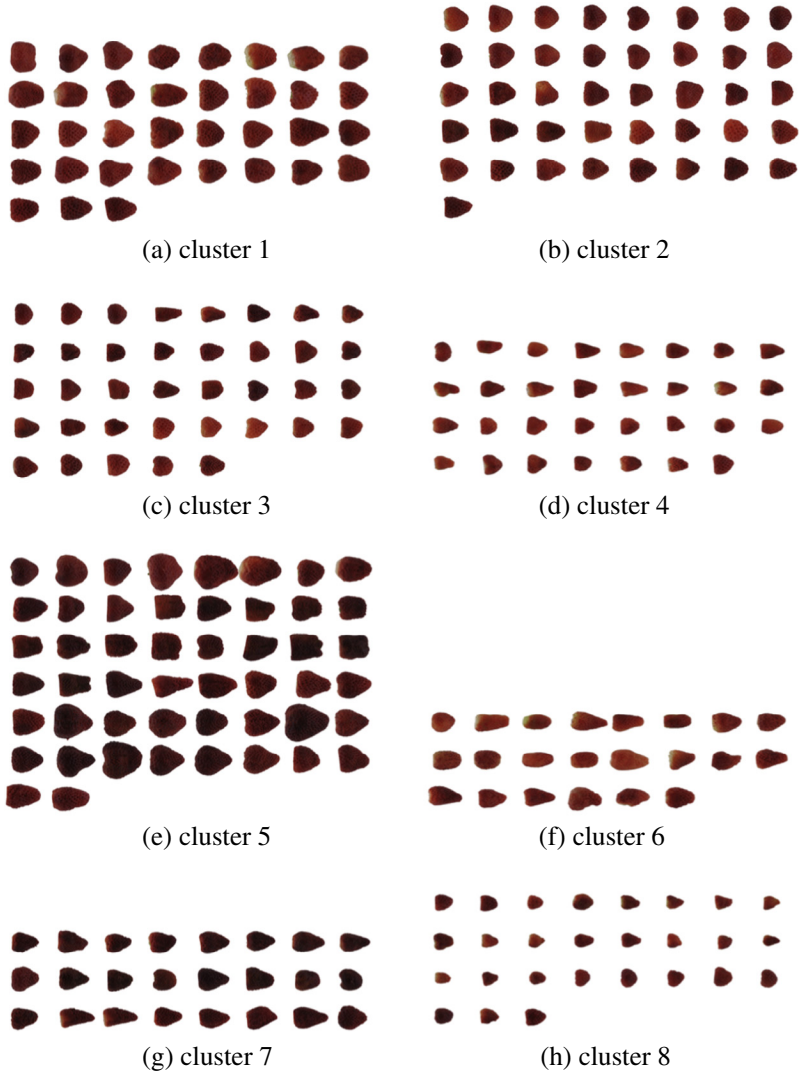

Fig. 2. Results of the cluster analysis using eight clusters. Fruit with similar appearance were classified into the same clusters.

Table 2
Average values of the surface color (H, S and V), shape distances generated from the circle model and fruit size for each cluster.

Cluster	H	S	V	Shape distance	Size ($\times 10^5$)
1	13.62	185.96	129.31	0.075	6.65
2	21.61	182.79	126.89	0.081	4.73
3	31.13	181.09	118.30	0.092	3.48
4	9.83	187.03	142.52	0.103	2.77
5	60.84	168.95	91.98	0.082	7.40
6	8.52	186.78	141.57	0.132	4.65
7	68.94	161.99	81.53	0.097	4.73
8	32.82	179.52	123.32	0.077	2.08

3.2. Cultivar identification based on the fruit appearance characteristics

The results of the PCA of the color and shape data for the 14 cultivars with larger amounts of samples are presented in Table 3. Three principal components provided a good summary of the color and shape data, which explained 64.0% and 64.7% of the total variance, respectively.

Figs. 4 and 5 present the variation in the first to third principal component scores for the color and shape data, whereas Fig. 6 presents the variation in the fruit size among the 14 cultivars. There were significant differences among the 14 cultivars with respect to each principal component and the fruit size (Table 4).

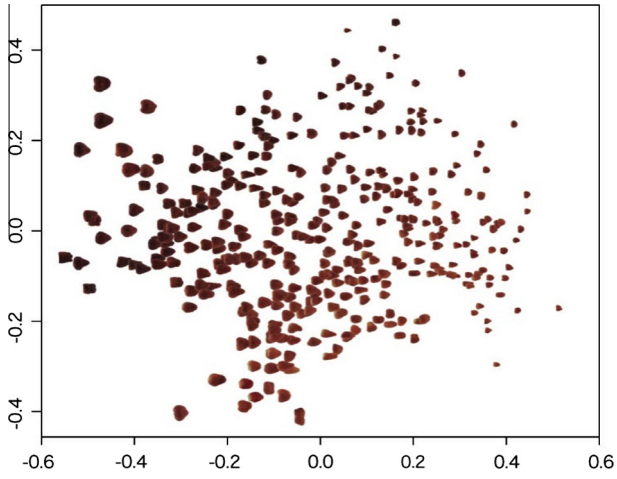

Fig. 3. Result of MDS mapping. The first dimension (x-axis) was significantly correlated with the surface lightness and fruit size and the second dimension (y-axis) was significantly correlated with the surface color and fruit size.

Table 3
Eigenvalues and contributions of the principal components for the color and shape data.

Component	Color			Shape		
	Eigenvalue	Proportion (%)	Cumulative (%)	Eigenvalue	Proportion (%)	Cumulative (%)
1	6.32	40.99	40.99	1.16	46.53	46.53
2	3.47	12.35	53.34	0.54	10.28	56.81
3	3.22	10.65	63.99	0.48	7.92	64.73
4	2.09	4.48	68.47	0.45	7.18	71.91
5	1.92	3.76	72.23	0.33	3.72	75.63
6	1.85	3.52	75.75	0.30	3.12	78.75

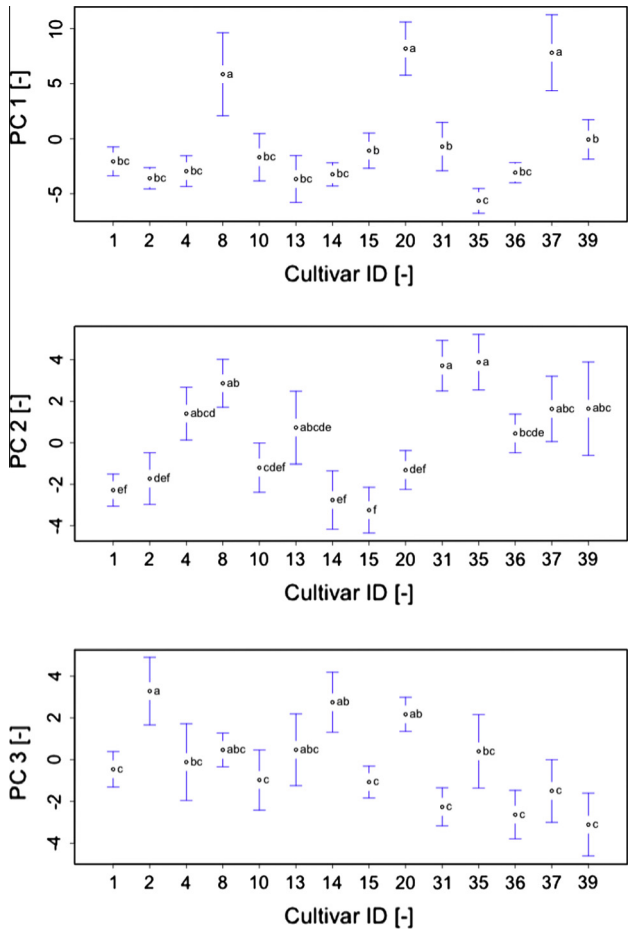

Fig. 4. Variations in the first to third principal component scores based on the results of the color image analysis of 14 strawberry cultivars. The mean values of the cultivars with different letters in each plot were significantly different ($p < 0.05$). Table 1 provides a key to the cultivars numbered on the horizontal axis.

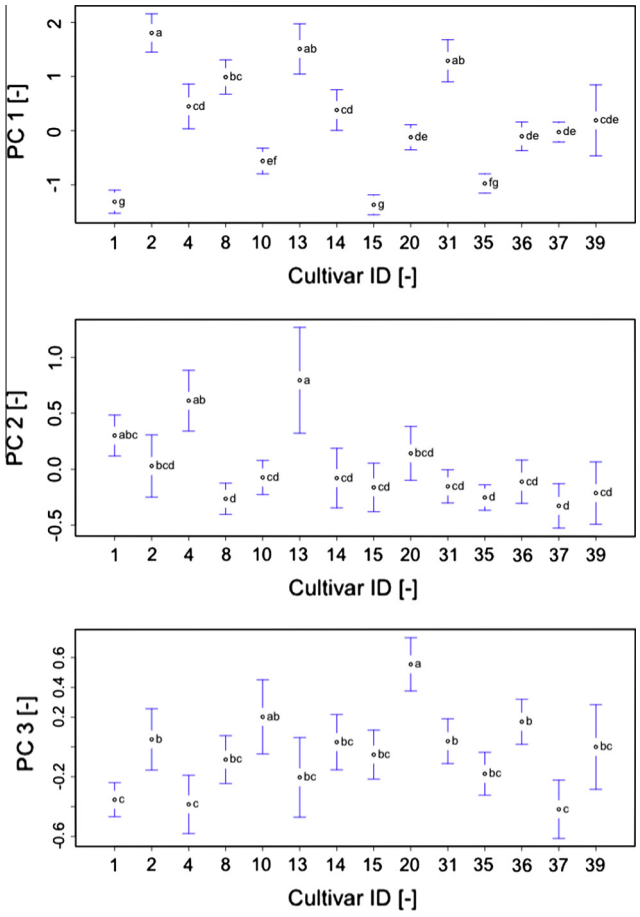

Fig. 5. Variations in the first to third principal component scores based on the results of the shape analysis of 14 strawberry cultivars.

The results of Tukey's multiple comparisons tests are shown in Figs. 4–6. In each graph, the cultivars with different letters were significantly different from each other. On the basis of these results, we characterized some of the cultivars. For example, as presented in Fig. 4, "Tsukushi" (cultivar ID: 8), "CHANDLER" (cultivar ID: 20), and "Amao" (cultivar ID: 37) were distinguished from the other 11 cultivars using the first principal component score for color. This result indicates that these three cultivars had distinctive characteristics in terms of their color. Furthermore, the fruit of "Amao" was relatively larger than that of the other cultivars; therefore, we could distinguish it from the other 13 cultivars using the first principal score component for color and the fruit size (presented in Fig. 6). "Kurume 16" (cultivar ID: 1) and "TORO" (cultivar ID: 15) had distinctive shapes; therefore, they were distinguished using the first principal component score for shape (Fig. 5). However, as shown in Figs. 4–6, most of the cultivars could not be distinguished based on a single feature.

Table 5 presents the accuracy of cultivar classification using the first to *n*th principal components from the color and shape data. No significant differences were found among most of the cultivars on the basis of a single feature and the accuracy rates of cultivar classification using color, shape and size data were less than 41.6%, 39.7% and 25.1%, respectively. In contrast, the rate increased to 68.0%, when using all three features of the first three principal components from the color and shape data (cumulative contribution ratios are more than 60%) and the size data. Most of the cultivars differed with respect to at least one appearance characteristic; therefore, combining the three features increased the accuracy. These results indicate that our system was advantageous for analyzing multiple appearance characteristics.

The appearance characteristics of agricultural products have been previously evaluated based on visual inspections, and

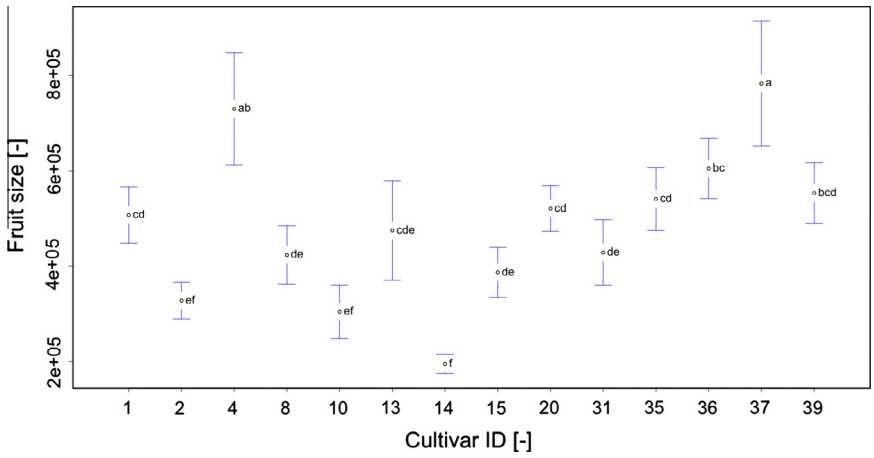

Fig. 6. Variations in fruit size among the 14 strawberry cultivars.

Table 4
Results of the ANOVA of the first three principal components for the color and shape data and fruit size.

Source	df	PC1		PC2		PC3	
		SS	F	SS	F	SS	F
(a) Color							
Cultivar	13	5730.3	22.7[a]	1498.1	17.1[a]	980.2	10.7[a]
Error	253	4906.4		1705.6		1784.2	
(b) Shape							
Cultivar	13	253.8	48.0[a]	19.7	6.5[a]	22.1	11.1[a]
Error	253	103.0		59.2		38.7	
(c) Size							
Cultivar	13	5.5	21.6[a]				
Error	253	4.9					

[a] $p < 0.001$.

Table 5
Discriminant analysis classification results using the first to *n*th principal components from the color and shape data.

PCs[a]	Accuracy rate (%)	
	Color	Shape
1	23.2	27.0
1–2	28.1	30.7
1–3	34.8	39.0
1–4	35.6	37.8
1–5	37.5	39.7
1–6	41.6	39.7

[a] 1 – *n*: first to *n*th principal components used for discriminant analysis.

recorded as categorical data. For example, the shapes of strawberry fruit have been recorded as text terms such as "sphere," "oblate" and "circular cone." However, the appearance characteristics vary continuously; therefore, these categorical data are not adequate for describing the small differences among strawberry fruits. In contrast, our method can describe the appearance characteristics based on quantitative data, and it can also distinguish cultivars. Thus, our method can be used as a new index of the appearance characteristics instead of using categorical data.

We used a number of cultivars, and it was quite difficult even for specialists to identify all of the cultivars; however, the accuracy rate we obtained was still fairly low. In fact, one-third of the fruit were misclassified as incorrect cultivars. Table 6 presents the classification accuracy for each cultivar. Although, as discussed above, "Amao" (cultivar ID: 37) was discriminable from the other cultivars based on the first principal component of color and size, the accuracy of cultivar identification of the cultivar was comparatively low. The fruits of "Amao" were mainly misclassified into "Benihoppe" (cultivar ID: 36) and "Akihime" (cultivar ID: 39). In fact, there were no significant differences in the features other than the first principal component of color and size among the cultivars (Figs. 4–6). Thus, such similarity in the most of the features may have caused the accuracy of misclassification. This might be a drawback of using multiple features, and we need to carefully investigate it in the future study.

In this study, we only analyzed the fruit surface appearance characteristics, but several important characteristics used for strawberry cultivar identification are revealed in cross-section

Table 6
Classification results for each cultivar using linear discriminant analysis.

From cultivars (ID)	Number of fruits classified into cultivars (ID)														Accuracy (%)
	1	2	4	8	10	13	14	15	20	31	35	36	37	39	
01	22	0	0	0	1	0	0	3	0	0	2	0	0	0	78.6
02	0	17	0	0	0	0	2	0	0	1	0	0	0	0	85.0
04	0	0	9	0	0	0	0	0	0	0	2	1	0	0	75.0
08	0	2	0	9	0	0	0	0	1	2	0	0	0	0	64.3
10	1	0	0	0	9	0	1	1	2	1	1	0	0	0	56.3
13	0	1	1	0	0	5	0	0	0	2	0	0	0	0	55.6
14	0	3	0	0	2	1	12	0	0	0	0	0	0	0	66.7
15	2	0	0	0	3	0	0	14	1	0	1	0	0	0	66.7
20	1	1	0	1	3	0	3	4	23	0	0	0	1	0	62.2
31	0	0	0	3	2	0	0	0	0	14	0	0	0	1	70.0
35	1	0	0	0	0	0	0	1	0	0	19	2	0	0	82.6
36	1	0	1	0	0	0	0	0	0	0	1	15	0	2	75.0
37	0	0	1	0	0	0	0	1	0	0	0	2	10	2	62.5
39	1	0	0	0	0	1	0	3	0	5	0	0	0	3	23.1

such as the ratio of white to red flesh. Thus, the utilization of this type of information would increase the classification accuracy.

The variation within the samples of each cultivar in this study is not as wide as they would be in practical cases, because all the samples we used were grown under a single environment and each cultivar was harvested at the same timing. In this study, we focused on the methods to evaluate appearances quantitatively by utilizing image analyses and we concluded that, as the first step of the study, the overall variation among the samples from all cultivars was diverse enough to evaluate the methods. In the future study, we need more samples grown and harvested under various conditions for the practical use of the system.

4. Conclusions

In this study, we developed a new image analysis system that can simultaneously evaluate multiple appearance characteristics of agricultural products. To evaluate the effectiveness of this system, we conducted quality evaluations and cultivar identification on the basis of statistical analyses of the appearance characteristics.

The result of the cluster analysis revealed that strawberries could be classified on the basis of their appearance characteristics. This result indicated that the appearance distances we defined in this study was able to represent the distance close to a human sense. By performing MDS, we were able to visualize the small differences in the appearance of the fruit based on multiple characteristics on a two-dimensional surface. Since we were able to quantitatively evaluate the distance on multiple fruit appearances, we plan to utilize our system for breeding purposes in the future study.

The results of the discriminant analysis revealed that the accuracy of strawberry cultivar classification using 14 cultivars was <42%, when only single feature such as color, shape and size was used. However, the rate increased to 68% after combining the three features. These results indicate that our system exploits the advantage of analyzing multiple appearance characteristics. There are several important characteristics used for strawberry cultivar identification also in cross-section of fruits. Thus, the utilization of this type of information would increase the classification accuracy in the future study.

Acknowledgements

This study was supported by a Grant-in-Aid for "Accelerating Utilization of University IP Program" from the Japan Science and Technology Agency (JST). This research was also supported by a Grant from the Ministry of Agriculture, Forestry and Fisheries (Genomics-based Technology for Agricultural Improvement, NGB-3002).

Appendix A. List of variables used in the color analysis

i	color bin
A_i	set of pixels color bin of which is i
$\|A_i\|$	the number of pixels in A_i
C_i	the center of gravity coordinate of pixels in A_i
N	the number of annular circles
$\|A_{ij}\|$	the number of pixels inside the jth circle $(1 \leq j \leq N)$
P_i	normalized spatial distribution histogram of color bin i
E_i	CDE of color bin i
E'_i	I-CDE of color bin i
E''_i	modified I-CDE of color bin i

References

Bianco, L., Lopez, L., Scalone, A.G., Di Carli, M., Desiderio, A., Benvenuto, E., Perrotta, G., 2009. Strawberry proteome characterization and its regulation during fruit ripening and in different genotypes. J. Proteomics 72, 586–607.

Brosnan, T., Sun, D.-W., 2004. Improving quality inspection of food products by computer vision––a review. J. Food Eng. 61, 3–16.

Clement, J., Novas, N., Gazquez, J.-A., Manzano-Agugliaro, F., 2012. High speed intelligent classifier of tomatoes by colour, size and weight. Spanish J. Agric. Res. 10, 314–325.

Clement, J., Novas, N., Gazquez, J.-A., Manzano-Agugliaro, F., 2013. An active contour computer algorithm for the classification of cucumbers. Comput. Electron. Agric. 92, 75–81.

Hitchcock, F.L., 1941. The distribution of a product from several sources to numerous localities. J. Math. Phys. 20, 224–230.

Itseez, 2014. OpenCV. URL <http://code.opencv.org/projects/opencv> (accessed 10.02.14).

Jahns, G., Møller Nielsen, H., Paul, W., 2001. Measuring image analysis attributes and modelling fuzzy consumer aspects for tomato quality grading. Comput. Electron. Agric. 31, 17–29.

Jolion, J., Meer, P., Bataouche, S., 1991. Robust clustering with applications in computer vision. IEEE Trans. Pattern Anal. Mach. Intell. 13, 791–802.

Kondou, H., Itou, H., Ishikawa, H., Motonaga, Y., Hashimoto, A., Kameoka, T., 1998. In: Color Chart for Fruits of Grape 'Aki Queen' by Digital Image Processing. Agricultural Information Technology in Asia and Oceania, Wakayama, Japan, pp. 197–202.

Kondou, H., Kitamura, H., Nishikawa, Y., Motonaga, Y., Hashimoto, A., Naknishi, K., Kameoka, T., 2002. Shape evaluation by digital camera for grape leaf. In: Third Asian Conference for Information Technology in Agriculture. Beijing, China, pp. 586–590.

Liming, X., Yanchao, Z., 2010. Automated strawberry grading system based on image processing. Comput. Electron. Agric. 71, S32–S39.

Lin, P., Chen, Y., He, Y., 2010. Identification of broken rice kernels using image analysis techniques combined with velocity representation method. Food Bioprocess Technol. 5, 796–802.

López-García, F., Andreu-García, G., Blasco, J., Aleixos, N., Valiente, J.-M., 2010. Automatic detection of skin defects in citrus fruits using a multivariate image analysis approach. Comput. Electron. Agric. 71, 189–197.

Pallottino, F., Menesatti, P., Costa, C., Paglia, G., Salvador, F.R., Lolletti, D., 2009. Image analysis techniques for automated hazelnut peeling determination. Food Bioprocess Technol. 3, 155–159.

Pelleg, D., Moore, A.W., 2000. X-means: extending *K*-means with efficient estimation of the number of clusters. In: Proceedings of the Seventeenth International Conference on Machine Learning, ICML'00. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 727–734.

R Developmet Core Team, 2011. R: A Language and Environment for Statistical Computing. R Found. Stat. Comput., R Foundation for Statistical Computing, Vienna, Austria, <http://www.r-project.org>.

Rao, A., Srihari, R.K., Zhang, Z., 1999. Spatial color histograms for content-based image retrieval. Proc. Eleventh Int. Conf. Tools with Artif. Intell., 183–186.

Rubner, Y., Tomasi, C., Guibas, L.J., 2000. The earth mover's distance as a metric for image retrieval. Int. J. Comput. Vis. 40, 99–121.

Sarkar, N., Little, A., Wolfe, R., 1984. Computer vision based system for quality separation of fresh market tomatoes. Am. Soc. Agric. Eng., 1714–1718.

Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27, 379–423.

Singh, N., Delwiche, M.J., Johnson, R.S., 1993. Image analysis methods for real-time color grading of stonefruit. Comput. Electron. Agric. 9, 71–84.

Sun, J., Zhang, X., Cui, J., Zhou, L., 2006. Image retrieval based on color distribution entropy. Pattern Recognit. Lett. 27, 1122–1126.

Swain, M.J., Ballard, D.H., 1991. Color indexing. Int. J. Comput. Vis. 7, 11–32.

Yamamoto, K., Yoshitsugu, K., Togami, T., Yoshioka, Y., Hashimoto, A., Kameoka, T., 2011. A chromatic image analysis system using content-based image retrieval. Agric. Inf. Res. 20, 139–147.

Yoshioka, Y., Fukino, N., 2009. Image-based phenotyping: use of colour signature in evaluation of melon fruit colour. Euphytica 171, 409–416.