Combining multiple features for color texture classification

Article *in* Journal of Electronic Imaging · July 2016

DOI: 10.1117/1.JEI.25.6.061410

3 authors:

Claudio Cusano
University of Pavia
77 PUBLICATIONS **839** CITATIONS
SEE PROFILE

Paolo Napoletano
Università degli Studi di Milano-Bicocca
67 PUBLICATIONS **258** CITATIONS
SEE PROFILE

Raimondo Schettini
Università degli Studi di Milano-Bicocca
305 PUBLICATIONS **3,281** CITATIONS
SEE PROFILE

Combining multiple features for color texture classification

Claudio Cusano,[a] Paolo Napoletano,[b,*] and Raimondo Schettini[b]
[a]University of Pavia, Department of Electrical, Computer, and Biomedical Engineering, Via Ferrata 1, 27100 Pavia, Italy
[b]University of Milan-Bicocca, Department of Informatics, Systems, and Communication, Viale Sarca 336, 20126 Milano, Italy

Abstract. The analysis of color and texture has a long history in image analysis and computer vision. These two properties are often considered as independent, even though they are strongly related in images of natural objects and materials. Correlation between color and texture information is especially relevant in the case of variable illumination, a condition that has a crucial impact on the effectiveness of most visual descriptors. We propose an ensemble of hand-crafted image descriptors designed to capture different aspects of color textures. We show that the use of these descriptors in a multiple classifiers framework makes it possible to achieve a very high classification accuracy in classifying texture images acquired under different lighting conditions. A powerful alternative to hand-crafted descriptors is represented by features obtained with deep learning methods. We also show how the proposed combining strategy hand-crafted and convolutional neural networks features can be used together to further improve the classification accuracy. Experimental results on a food database (raw food texture) demonstrate the effectiveness of the proposed strategy. *© 2016 SPIE and IS&T* [DOI: 10.1117/1.JEI.25.6.XXXXXX]

Keywords: color texture classification; color texture features; color texture database; ensemble of classifiers.
Paper 16249SS received Mar. 26, 2016; accepted for publication Jun. 20, 2016.

1 Introduction

The recognition of color texture is a widely studied topic in computer vision and pattern recognition. Most of the research efforts on this subject have been devoted to the definition of suitable descriptors able to capture the distinctive properties of the texture images while being invariant, or at least robust, with respect to some variations in the acquisition conditions, such as rotations and scalings of the image, changes in brightness, contrast, and light color temperature.[1]

Previous works have made an extensive evaluation of descriptors that encode both color and texture information.[1,2] Results have shown that these descriptors perform poorly when both the color and position of the light source change. A possible strategy to exploit color in texture classification consists in the extraction of image features that are invariant (or at least robust) with respect to changes in the illumination. For instance, Seifi et al., proposed to characterize color textures by analyzing the rank correlation between pixels located in the same neighborhood and by using a correlation measure that is related to the colors of the pixels, and is not sensitive to illumination changes.[3] Cusano et al.[4] proposed a descriptor that measures the local contrast: a property that is less sensitive than color itself to variations in the color of the illuminant. The same authors then enhanced their approach by introducing a novel color space where changes in illumination are even easier to deal with.[5] Other strategies for color texture recognition have been proposed by Drimbarean and Whelan,[6] who used Gabor filters and co-occcurrence matrices, by Bianconi et al.,[7] who used ranklets and the discrete Fourier Transform, and by Palm, who experimented with integrative co-occurrence matrices.[8] Comparative studies on the combination of color and texture features have been documented by Porebski et al.[9] and Iakovidis et al.[10]

Recent works suggested that, in several application domains, carefully designed features can be replaced by features automatically learned from a large amount of data with methods based on deep learning.[11] Cimpoi et al.,[12] for instance, used Fisher vectors to pool features computed by a convolutional neural network (CNN) trained for object recognition. Cusano et al.[13] have shown that none of the descriptors in the state of the art is robust enough to any possible variation of illuminations especially when its magnitude is large.

Since different visual descriptors induce different sensitivities to variations of the lighting conditions, we designed a classification strategy based on an ensemble of descriptors that capture different aspects of images such as color, texture, and color contrast, showing a high degree of robustness to specific changes in lighting (e.g., changes in intensity, color temperature, light direction, etc.). We show that a proper combination of these descriptors makes it possible to achieve a very high classification accuracy even under large variations of the lighting conditions. To demonstrate the effectiveness of our approach we conducted an extensive experimentation with the raw food texture (RawFooT) database,[13] a collection of texture images that have been carefully acquired to allow comparative studies of texture recognition under controlled variations of the light intensity, direction, and color. The results prove how our approach significantly boosts the performance of the individual descriptors, making it possible to improve the state of the art.

The focus on the descriptors and on their combination presents the advantage that the classifier is not required to

*Address all correspondence to: Paolo Napoletano, E-mail: napoletano@disco.unimib.it

learn a model of the variations in the lighting conditions. This allowed us to experiment with a very simple classifier (we used the nearest neighbor classifier), to work with a small number of training examples, and to limit the dependence of our findings on the composition of the training set. Note, however, that the proposed strategy is very general and that it can be easily adapted to more complex classification models such as, for instance, support vector machines (SVMs).

The paper is organized as follows: Sec. 2 presents in detail the ensemble of descriptors we selected and the combination rule we designed. Section 3 describes the experimental setup and reports a comparison, in terms of classification accuracy, with several other descriptors and combination rules of the state of the art. Finally, Sec. 4 summarizes the work and highlights some promising directions for future research on this topic.

2 Proposed Method

The performance of texture classification methods can be boosted by enriching the information extracted from the images and by encoding it into suitable descriptors. In particular, it has been observed that the combination of color and texture information can be very effective.[14] However, the actual effectiveness of such a combination depends on several factors such as the degree of variability in the acquisition conditions. For instance, Mäenpää and Pietikäinen[2] have shown how a simple concatenation of color and texture descriptors is beneficial when the illuminant color is stable, and is detrimental when the illuminant color is variable.

Combining strategies represents a common approach for the improvement of the performance in classification problems (see the work by Kuncheva for a detailed introduction to the topic[15]). Kittler et al.[16] proposed a theoretical framework in which combination methods are modeled as the application of suitable operators to the posterior probabilities computed by multiple classifiers (this approach is often referred to as "late fusion" to distinguish it from the "early fusion" combining that operates directly in the feature space).[17] In their framework, each element to be classified (that in the following we assume to be an image) is represented by multiple descriptors $\mathbf{x}_1,\ldots,\mathbf{x}_R$, and must be assigned to one class among a set $\omega_1,\ldots,\omega_m$ of mutually exclusive classes. Two combining operators among the others have a particularly clear interpretation: the product and the sum.[18] Under the product combination rule the probability that an image is of class ω_j is considered as proportional to the product of the posterior probabilities computed from the individual descriptors

$$P(\omega_j|\mathbf{x}_1,\ldots,\mathbf{x}_R) \propto \prod_{i=1}^{R} P(\omega_j|\mathbf{x}_i). \tag{1}$$

The rule corresponds to the assumption that the distributions of the descriptors are conditionally independent given the class labels, i.e.,

$$P(\mathbf{x}_1,\ldots,\mathbf{x}_R|\omega_j) = \prod_{i=1}^{R} P(\mathbf{x}_i|\omega_j). \tag{2}$$

The sum combination rule, instead, assumes that all the descriptors encode the same information, with some additional noise. More precisely, it assumes that the posterior probabilities given a single descriptor are just a noisy version of the true probabilities

$$P(\omega_j|\mathbf{x}_1,\ldots,\mathbf{x}_R) = P(\omega_j|\mathbf{x}_i) + e_i, \quad i = 1,\ldots,R, \tag{3}$$

where e_i is a zero-mean noise term. On the basis of these assumptions, the sum rule states that the real posterior probability is approximated by the mean of the R posteriors

$$P(\omega_j|\mathbf{x}_1,\ldots,\mathbf{x}_R) \simeq \frac{1}{R}\sum_{i=1}^{R} P(\omega_j|\mathbf{x}_i). \tag{4}$$

Summing up, the product rule is expected to be effective when the descriptors encode independent information, while the sum rule is expected to produce reliable results when the descriptors are highly correlated.

In the case of texture classification, descriptors are not always independent. In fact, there are several descriptors that, in different ways, try to capture the same kind of information. On the other hand, it is probably safe to consider some groups of descriptors as mutually independent. For instance, color and gray-level texture descriptors are supposed to encode potentially orthogonal aspects of the images. On the basis of these considerations, we propose a new combination rule, the "product of means" rule, that takes into account the possibility of both independent and nonindependent descriptors. To begin with, the set of descriptors $\mathbf{x}_1,\ldots,\mathbf{x}_R$ is partitioned into K subsets $S_1,\ldots S_K$, where each subset contains closely related descriptors. For each subset, the sum rule is used to obtain a posterior probability and then, with the product rule, the final probability estimate

$$P(\omega_j|\mathbf{x}_1,\ldots,\mathbf{x}_R) \simeq \frac{1}{Z}\prod_{k=1}^{K}\frac{1}{|S_k|}\sum_{\mathbf{x}_i\in S_k} P(\omega_j|\mathbf{x}_i), \tag{5}$$

where Z is a term that scales the estimates to have unitary sum.

In practice, we partitioned the descriptors in three subsets on the basis of the way in which they exploit color information: the S_1 subset contains descriptors extracted from the gray-level image; the S_2 subset contains descriptors that make full use of the color information; the S_3 subset contains descriptors based on color contrast. The three sets of descriptors are better described in the next section. We used these 1 descriptors in a nearest-neighbor setup: in Sec. 3.2, we detail how we computed the posterior probabilities for the individual descriptors. Figure 1 summarizes the proposed approach.

2.1 *Visual Descriptors*

The set S_1 includes the local binary patterns (LBP) and bag of sift (BoVW), both calculated on the gray-level image defined as the luma of the image, as defined by the NTSC 2 standard[19]

$$L = 0.299R + 0.587G + 0.114B. \tag{6}$$

We considered the 243-dimensional LBP feature vector with a circular neighborhood of radius 2 and 16 elements, and no-rotation invariant and uniform patterns.[2] The uniformity measure of a pattern is the number of bitwise transitions from 0 to 1 or vice versa. An LBP is called uniform if

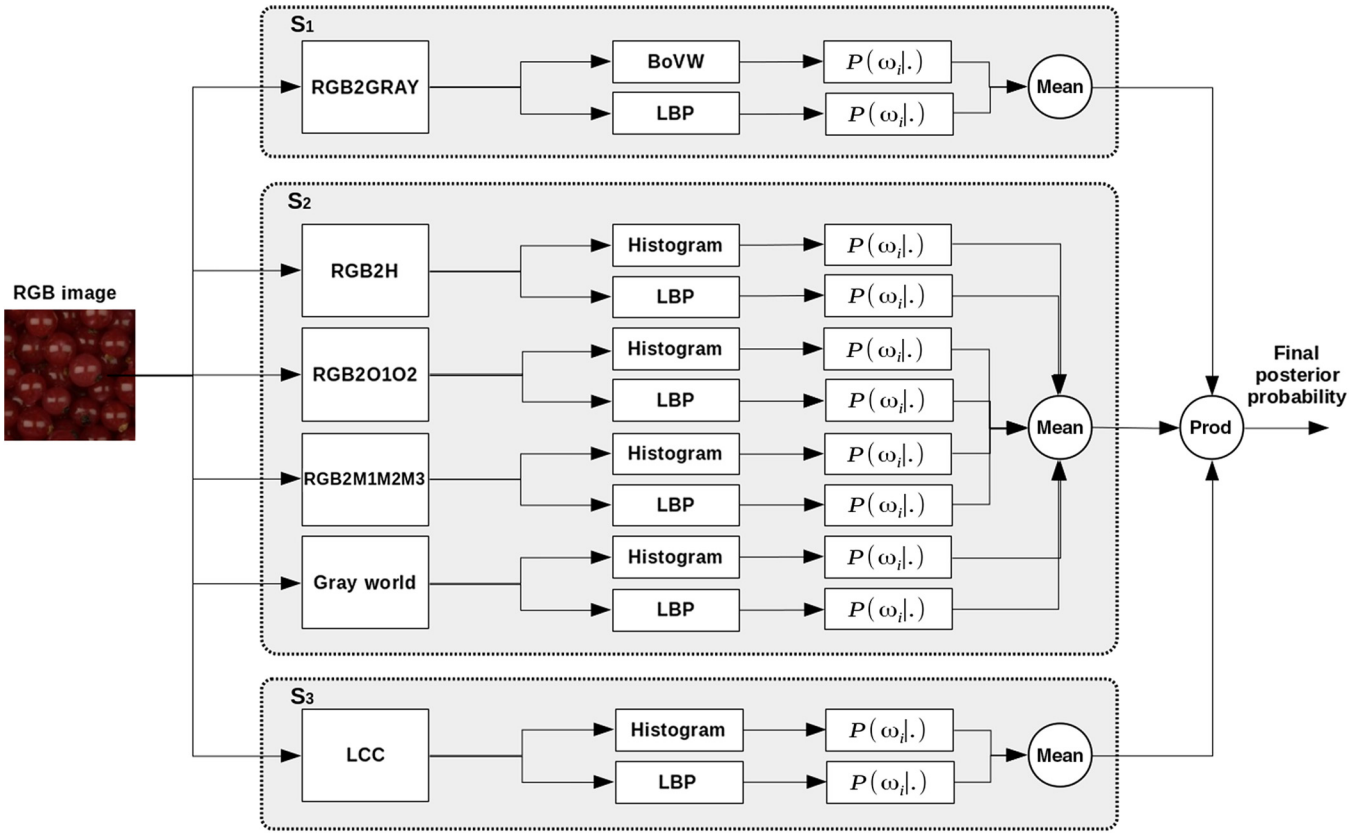

Fig. 1 Scheme of the proposed "product of means" strategy.

its uniformity measure is at most 2. As demonstrated in Ref. 2, for the neighborhood considered, the number of possible patterns is 243.

The BoVW descriptor consists in the aggregation of local descriptors according to the quantization defined by a codebook of visual words.[20] As local descriptors, we used 128-dimensional dense SIFT obtained by considering a spatial histogram of local gradient orientations. The codebook of 1024 visual words has been built on images from external sources.

The set S_2 includes the LBP and color histogram, both calculated on several color channels that are invariant to changes in the illumination conditions such as the "hue" channel, O_1 and O_2 channels, m_1, m_2, and m_3 channels and RGB channels after the "gray-world" preprocessing.

For each single channel histogram, we considered 256 bins. Multichannel histograms have been obtained concatenating single channel histograms.

The hue channel is taken from the HSV color space. It has been demonstrated that the hue channel is invariant to highlight, illumination intensity, illumination direction, and surface orientation.[21] The $O_1O_2O_3$ color space is obtained by decorrelating RGB channels:

$$O_1 = (R-G)/\sqrt{2},\ O_2 = (R+G-2B)/\sqrt{6},$$
$$O_3 = (R+G+B)/\sqrt{3}. \tag{7}$$

In particular, the chromaticity information is represented by O_1 (red-green channel) and O_2 (yellow-blue channel), whereas the intensity information is given by O_3.

It has been demonstrated that the O_1O_2 channels have invariant properties to highlight whereas the O_3 channel has no invariant property at all.[22] The color ratio model $m_1m_2m_3$ has been proposed by Gevers and Smeulders.[23] It considers the color ratio between two neighboring pixels x_1 and x_2 as follows:

$$m_1 = \frac{R(x_1)G(x_2)}{R(x_2)G(x_1)}, \quad m_2 = \frac{R(x_1)B(x_2)}{R(x_2)B(x_1)},$$
$$m_3 = \frac{G(x_1)B(x_2)}{G(x_2)B(x_1)}. \tag{8}$$

It has been demonstrated that this color representation is invariant to illumination color, intensity, illumination direction, and surface orientation. We also considered the gray-world algorithm proposed by Finlayson and Trezzi.[24] This algorithm is based on the assumption that the average value of the R, G, and B components of a given image should converge to a gray color given a sufficient amount of color variations. The gray-world preprocessing should neutralize the color of the illuminant.

The set S3 includes the LBP and color histogram, both calculated on the local color contrast (LCC) map proposed by Cusano et al.[25] The LCC map is obtained by comparing the color at a given location with the average color in a surrounding neighborhood. To make it robust against changes in the color of the illuminant, the LCC is computed in terms of the angular difference between the color vectors. The LCC map has been demonstrated to be invariant with respect to

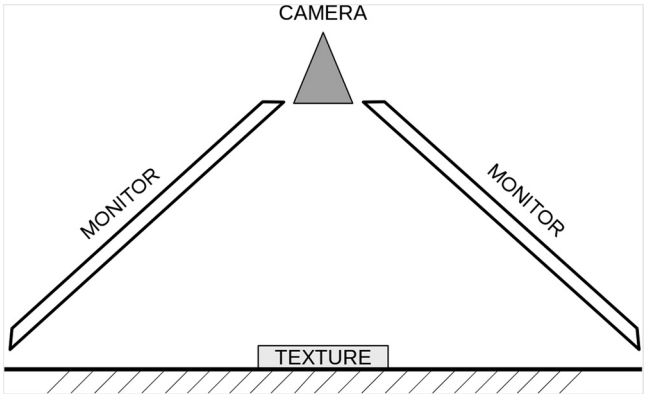

Fig. 2 Setup used to acquire the raw food texture database.

rotations and translations of the image plane and with respect to several transformations in the color space.

3 Experiments

3.1 *Data*

The RawFooT database has been specially designed to investigate the robustness of descriptors and classification methods with respect to variations in the lighting conditions, with a particular focus on variations in the color of the illuminant. The database includes 68 samples of raw food, including various kinds of meat, fish, cereals, fruit, etc. Textures were acquired in a dark room with a camera placed with the optical axis perpendicular to the surface of the sample and under 46 lighting conditions that may differ in the light direction, in the illuminant color, in its intensity, or in a combination of these factors. Figure 2 shows the setup used to acquire the dataset. The whole database includes $68 \times 46 = 3128$ images. Figure 3 shows an image of each sample.

The illuminants were simulated by controlling the light of a pair of LED monitors positioned above the sample and tilted by 45 deg. The two monitors were colorimetrically characterized using a spectral colorimeter, in order to render the desired chromaticites using the device *RGB* coordinates. The 46 shots for each texture sample are the following:

Intensity variations: Four shots were taken while illuminating the whole monitors with neutral light (D65) at different levels of intensity (100%, 75%, 50%, and 25% of the maximum achievable level).

Light direction: Nine shots were taken with the light (D65) coming from different angles. In the first eight of these shots only a band covering 40% of a single monitor has been lit. The angle between the direction of the light coming from the center of the illuminated band and the surface of the sample were 24 deg, 30 deg, 36 deg, 42 deg, 48 deg, 54 deg, 60 deg, and 66 deg.

Daylight: Twelve shots were taken while simulating natural daylight at 12 different color temperatures in the range from 4000 to 9500 K with a step of 500 K (we will refer to these as D40, D45, and D95). The whole monitors were lit during these shots.

Indoor illumination: Six shots were taken while simulating an artificial light with a color temperature of 2700, 3000, 4000, 5000, 5700, and 6500 K on the two whole monitors. We will refer to these as L27, L30, and L65.

Color and direction: Nine shots were taken by varying both the color and the direction of the illuminant. The combinations of three colors (D65, D95, and L27) and of

Fig. 3 Overview of the 68 classes included in the raw food texture database. For each class, it is shown the image taken under D65 at direction $\theta = 24$ deg.

three directions (24 deg, 60 deg, and 90 deg) were considered.

Multiple illuminants: Three shots were taken while the sample was illuminated by two illuminants with different colors (D65, D95, or L27). Bands covering the lower 40% of both the monitors were lit, using two different colors on the two monitors.

Primary colors: Three shots were taken under pure red, green, and blue illuminants.

The final texture images were obtained by cropping the central region of 800×800 pixels from each 3944×2622 original texture sample. Figure 4 shows the 46 shots taken for two different texture samples.

3.1.1 *Data split for classification*

For each of the 68 classes, we considered 16 patches obtained by dividing the original texture image, that is of size 800×800 pixels, in 16 nonoverlapping squares of size 200×200 pixels. For each class, we selected eight patches for training and eight for testing alternating them in a chessboard pattern. We form subsets of $68 \times (8 + 8) = 1088$ patches by taking the training and test patches from images taken under different lighting conditions. In this way, we defined several subsets, grouped in nine texture classification tasks:

1. **No variations (NOVAR)**: Forty-six subsets. Each subset is composed of training and test patches taken under the same lighting condition.
2. **Light intensity (INT)**: Twelve subsets obtained by combining the four intensity variations. Each subset is composed of training and test patches with different light intensity values.
3. **Light direction (DIR)**: Seventy-two subsets obtained by combining the nine different light directions. Each subset is composed of training and test patches with different light directions.
4. **Daylight temperature (DAY)**: One hundred and thirty-two subsets obtained by combining all 12 daylight temperature variations. Each subset is composed of training and test patches with different light temperatures.

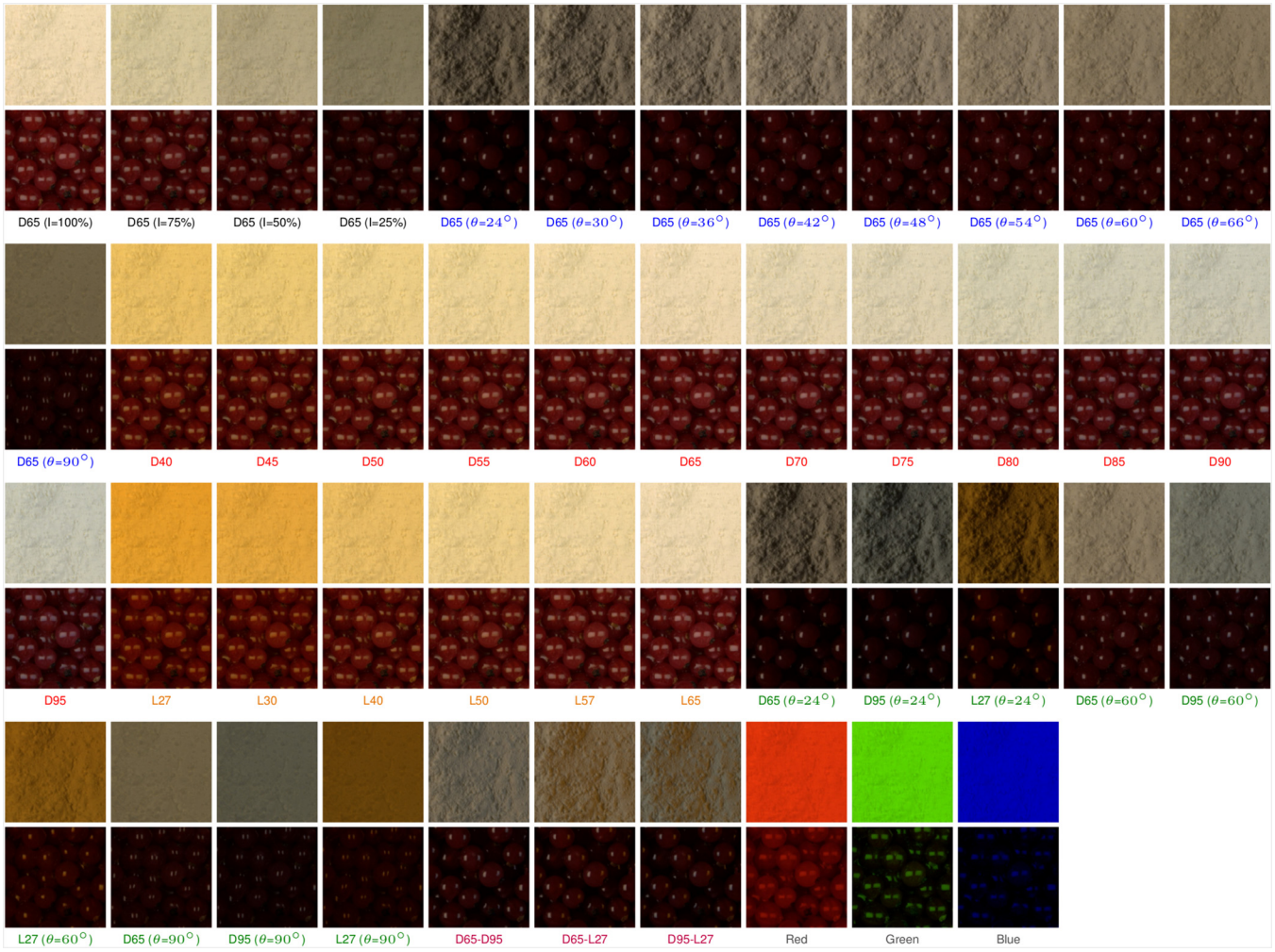

Fig. 4 Overview of the 46 lighting conditions in the raw food texture database: the top lines represent the "flour" class while bottom lines represent the "currant" class. The colors of the labels indicate sets of acquisition conditions used to define the experiments: variable light intensity, variable light direction, simulated daylight color, simulated artificial color, variable color and/or direction, multiple illuminants, and primary colors.

5. **LED temperature (LED)**: Thirty subsets obtained by combining all six LED temperature variations. Each subset is composed of training and test patches with different light temperatures.
6. **Daylight versus LED (DvL)**: Seventy-two subsets obtained by combining 12 daylight temperatures with six LED temperatures.
7. **Temperature or direction (T ∨ D)**: Seventy-two subsets obtained by combining all nine combinations of color temperatures and light directions. Each subset is composed of training and test patches where either the color or the direction (or both) change.
8. **Temperature and direction (T ∧ D)**: Thirty-six subsets obtained by combining all nine combinations of color temperatures and light directions. Each subset is composed of training and test patches where both the color and the direction change.
9. **Multiple illuminant (MULT)**: Six subsets obtained by combining the three acquisitions with multiple illuminants.

In the experiments, for each of the nine tasks, we computed the average classification rate over the subsets. Note that some tasks (notably T ∨ D and T ∧ D) are expected to be more challenging than the others because they include, on average, larger differences in the illumination conditions.

3.2 *Experimental Setup*

In all the experiments, we used the nearest neighbor classification strategy: given a patch in the test set, its distance with respect to all the training patches is computed. The prediction of the classifier is the class of the closest element in the training set. All the experiments have been conducted under the "maximum ignorance" assumption, i.e., no information about the lighting conditions of the test patches is available for the classification method and for the descriptors. Performance is reported as classification rate (i.e., the ratio between the number of correctly classified images and the number of test images). Note that more complex classification schemes (e.g., SVMs) would have been viable. We decided to adopt the simplest one, that is the nearest neighbor classifier, in order to focus the evaluation on the descriptors themselves and not on the classifier; therefore, the most of the computational time is on feature extraction.

Given a new image represented by the multiple descriptors $\mathbf{x}_1, \ldots, \mathbf{x}_R$, its assignement to one of the classes $\omega_1, \ldots, \omega_m$ is done according to the combining rules described in Sec. 2. These rules require that the posterior probabilities for each descriptor are available. Inspired by Kittler et al.,[16] we defined them on the basis of the distances between the descriptors

$$P(\omega_j|\mathbf{x}_i) = \frac{\exp[-d_j(\mathbf{x}_i)^2]P(\omega_j)}{\sum_{k=1}^{m}\exp[-d_k(\mathbf{x}_i)^2]P(\omega_k)}, \tag{9}$$

where $P(\omega_j)$ is the *a-priori* probability that the image is of class ω_j [$P(\omega_j) = 1/m$ in our experiments]. The normalized distance $d_j(\mathbf{x}_i)$ between the descriptor $\mathbf{x}_i$ and the class ω_j is computed from a training set $T_{i,j}$ of training examples of descriptors (of the same type of $\mathbf{x}_i$) computed on images of class ω_j

$$d_j(\mathbf{x}_i) = \min_{\mathbf{x}_i' \in T_{i,j}} \frac{\|\mathbf{x}_i - \mathbf{x}_i'\|_1}{\sigma_i}, \tag{10}$$

where σ_i is the standard deviation of the distances between pairs of descriptors, estimated on the whole training set $T_i = \bigcup_j T_{i,j}$. We chose to adopt the L_1 distance after some preliminary tests where we compared it against a few other common distance measures.

3.3 *Results*

Table 1 reports the average performance we obtained in the nine classification tasks considered. Together with all the results of the individual descriptors, the table reports the performance of several combining strategies including the proposed product of means. The competing combining strategies are: the simple concatenation of the descriptors that, in practice, represents an early fusion of the image descriptors; the sum, product, minimum, maximum, and median combination rules described by Kittler et al.;[16] two voting schemes: the majority vote rule (where each descriptor votes for a class and the most voted is chosen) and the Borda count rule (where each descriptor determines a ranking among the classes, and the class with the best average rank is chosen).

In the "NOVAR" task, the introduction of combining clearly boosts the performance of the individual descriptors, almost reaching 100% of classification accuracy. The improvement of combining is even more significant in the more challenging classification tasks that include variation in the illumination. Even the simple varation in the intensity of illumination (INT) causes a noticeable drop in the classification performance for almost all the descriptors, including those that are supposed to be invariant (such as LBP). This result can be explained by taking into account the fact that the relative amount of acquisition noise depends on the illumination conditions.

Our proposed product of means combination rule clearly outperforms the other combination strategies considered. In fact, it obtained the highest classification accuracy in five of the nine classification tasks. Moreover, in three of the other four tasks it obtained performance very close to that of the best method (less than 1% of difference). The only task in which our combining rule is clearly suboptimal is the classification under variable artificial lights (LED). This task is the one characterized by the largest variation in the color of the illuminant. In this case, color-based descriptors are very low performing; therefore, it is not surprising that any combining considering them are not so well performing. In this 7 case, the gray-level descriptor (BoVW) obtained the best result (88.87%). However, even in this task our combination strategy (reaching 84.11% of accuracy) outperformed by a large margin, all the other combinations considered (ranging from 23.90% of a simple concatenation to 74.41% of Borda count), by showing that our proposal is able to mitigate the influence of low performing features in the overall combination.

3.4 *CNN-Based Methods*

Following the trend in image recognition, features extracted from CNNs can be adopted for texture classification as well. CNNs allow for leveraging very large datasets of labeled

Table 1 Classification rates (%) of combining methods, applied to a selection of color descriptors. Methods have been evaluated by introducing different kinds of variations in the illumination conditions (NOVAR: no variations; INT: varying intensity; DIR: varying light direction; DAY: daylight, varying the color temperature; LED: artificial light, varying the color temperature; DvL: daylight versus artificial LED lights; T ∨ D: varying color temperature or light direction; T ∧ D: varying color temperature and light direction; and MULT: with multiple illuminants).

Method	NOVAR	INT	DIR	DAY	LED	DvL	T ∨ D	T ∧ D	MULT
Single descriptors									
LBP angle	59.71	38.07	35.68	41.20	21.26	28.56	15.29	10.04	42.34
Hist angle	74.48	23.21	26.79	51.39	24.37	35.25	9.72	6.88	42.06
Hist L	78.32	6.83	20.97	49.94	27.18	38.05	10.45	7.98	47.33
LBP L	80.37	51.15	47.09	77.76	70.77	73.15	29.54	18.66	76.99
Hist O_1O_2	92.21	18.21	45.38	24.42	24.11	22.49	11.83	9.21	28.31
LBP O_1O_2	85.00	59.34	56.75	56.16	33.15	39.17	21.09	14.97	46.51
Histogram H	85.65	48.15	50.48	31.74	26.31	24.85	15.71	9.95	34.16
LBP H	66.96	40.44	37.65	34.44	28.14	28.37	15.29	10.26	39.58
GrayWorld Hist *RGB*	98.81	47.87	36.08	51.90	22.42	35.12	13.41	8.36	27.24
GrayWorld LBP *RGB*	93.63	80.91	61.94	77.88	47.09	58.19	27.10	17.11	72.40
LBP $m_1m_2m_3$	90.72	61.29	51.50	53.82	23.87	34.86	17.97	9.38	48.84
BoVW	89.73	87.38	67.38	90.02	**88.87**	89.53	51.59	**39.34**	88.60
Combinations									
Concatenation	90.87	61.40	51.52	53.86	23.90	34.89	17.99	9.40	48.84
Sum	99.97	75.35	77.67	63.01	58.50	53.48	33.31	21.77	78.52
Product	**99.99**	73.09	82.41	66.17	59.20	55.41	34.81	24.09	78.40
Minimum	95.66	49.74	71.27	49.43	41.26	38.89	23.20	15.10	60.60
Maximum	97.76	56.97	63.34	49.05	45.61	42.02	26.43	17.93	65.59
Median	98.41	82.32	76.92	80.85	60.82	64.58	35.08	25.09	78.22
Majority	99.45	87.84	76.77	90.91	74.28	78.45	42.18	28.37	88.88
Borda count	99.71	89.17	80.57	**92.04**	74.41	78.63	44.72	31.51	91.64
Prod. of means	99.89	**93.70**	**85.77**	91.55	84.11	**83.11**	**51.69**	38.56	**92.52**

images, by learning intermediate image representations that can be used for various image classification problems. To do so, the most common way consists in taking a CNN trained for object recognition and using the activation values of its last hidden layer as an image descriptor.[26] In our experiments, we considered the descriptors computed by two widely used CNNs: the AlexNet[27] and the VGG very deep at 16 layers[28] that both produce features of 4096 components. An alternative approach consists of selecting the last convolutional layer (instead of the last fully connected layer) and in using its activations as local descriptors. These are then pooled with a Fisher vector encoding. We included this approach in our experiment in its most performing setup, as identified by Cimpoi et al.:[12] medium sized VGG net (VGG-M) producing 512-dimensional local descriptors that are pooled with a codebook of 64 elements, yielding a final very large descriptor of $4096 \times 64 \times 2 = 65{,}536$ components. Table 2 shows how, in most tasks, the use of CNN-based features allows obtaining better results than those obtained by any individual "handcrafted" descriptor or by any of their combinations. With the Fisher vector encoding we obtained worse results than with the direct use of CNNs, probably because the rather

Table 2 Classification rates (%) of methods based on CNNs. They have been evaluated by introducing different kinds of variations in the illumination conditions (NOVAR: no variations; INT: varying intensity; DIR: varying light direction; DAY: daylight, varying the color temperature; LED: artificial light, varying the color temperature; DvL: daylight versus artificial LED lights; T ∨ D: varying color temperature or light direction; T ∧ D: varying color temperature and light direction; and MULT: with multiple illuminants).

Method	NOVAR	INT	DIR	DAY	LED	DvL	T ∨ D	T ∧ D	MULT
CNN-based									
BVLC AlexNet	94.88	85.00	80.25	95.79	88.56	89.64	48.08	39.37	90.44
VGG VeryDeep 16	98.21	94.10	91.23	97.41	93.69	93.67	70.81	63.64	96.60
VGG M + FV	92.67	54.67	71.76	78.87	65.56	68.88	38.49	28.77	84.74
Combinations									
Concatenation	99.53	96.00	86.21	96.42	91.76	92.47	60.15	46.21	95.53
Sum	99.96	95.42	92.98	95.84	91.33	90.22	65.23	55.16	97.52
Product	**99.99**	89.26	91.55	85.80	82.35	78.49	55.63	43.69	93.05
Minimum	97.86	60.03	80.07	58.93	53.93	49.43	32.91	23.22	72.79
Maximum	99.30	90.43	89.40	91.76	85.53	84.47	60.18	51.89	95.99
Median	99.51	89.71	84.28	90.26	73.24	76.02	46.14	34.83	88.24
Majority	99.72	92.91	83.38	95.81	86.80	88.80	53.27	39.60	94.03
Borda count	99.79	93.40	85.00	96.33	87.93	89.72	54.79	41.23	94.85
Prod. of means	99.90	**97.67**	**94.04**	**98.74**	**96.51**	**96.20**	**73.92**	**65.22**	**98.71**

large descriptors it produces require a more sophisticated classification strategy than nearest neighbor (e.g., SVMs, as done by Cimpoi et al.[12]). Our proposed strategy can be easily extended to include CNN features as well. In fact, it is enough to introduce CNN features as a new set (S_4) to be used in Eq. (5). Our modified solution obtained the best results on all the nine tasks often with quite a large improvement with respect to the use of single CNN features.

The inclusion of CNN features in our combination framework makes it is possible to achieve very high classification rates for all the classification tasks except those corresponding to multiple sources of variabilities (i.e., color and direction of light). This fact proves that even CNN features cannot completely model textures when the training and the test acquisition conditions are very different, as in the case of the T ∨ D and T ∧ D classification tasks.

4 Summary

We presented a strategy for color texture classification that makes use of an ensemble of heterogeneous image descriptors and that combine them with a novel combination rule. By exploiting, in a suitable way, the correlation among the descriptors, our strategy allows obtaining a high classification rate even in the presence of large variations of the illumination conditions.

We assessed our strategy on the RawFooT dataset and we found that, in most cases, it outperforms several other combination strategies from the state of the art. Moreover, it also outperforms all the single descriptors included in the experimentation with the only exception being the BoVW descriptor in two of the nine classification tasks considered.

In another experiment, we evaluated the performance of image descriptors extracted with CNNs. The inclusion of CNN descriptors in our combination strategy allowed obtaining the best results, by quite a large margin, on all the classification tasks.

References

1. F. Bianconi et al., "Theoretical and experimental comparison of different approaches for color texture classification," *J. Electron. Imaging* **20**(4), 043006 (2011).
2. T. Mäenpää and M. Pietikäinen, "Classification with color and texture: jointly or separately?," *Pattern Recognit.* **37**(8), 1629–1640 (2004).
3. M. Seifi et al., "Color texture classification across illumination changes," in *Conf. on Colour in Graphics, Imaging, and Vision*, pp. 332–337 (2010).
4. C. Cusano, P. Napoletano, and R. Schettini, "Combining local binary patterns and local color contrast for texture classification under varying illumination," *J. Opt. Soc. Am. A* **31**(7), 1453–1461 (2014).
5. C. Cusano, P. Napoletano, and R. Schettini, "Local angular patterns for color texture classification," *Lect. Notes Comput. Sci.* **9281**, 111–118 (2015).
6. A. Drimbarean and P. Whelan, "Experiments in colour texture analysis," *Pattern Recognit. Lett.* **22**(10), 1161–1167 (2001).
7. F. Bianconi et al., "Robust color texture features based on ranklets and discrete Fourier transform," *J. Electron. Imaging* **18**(4), 043012 (2009).
8. C. Palm, "Color texture classification by integrative co-occurrence matrices," *Pattern Recognit.* **37**(5), 965–976 (2004).
9. A. Porebski, N. Vandenbroucke, and L. Macaire, "Comparison of feature selection schemes for color texture classification," in *2nd Int. Conf. on Image Processing Theory Tools and Applications (IPTA '10)*, pp. 32–37 (2010).
10. D. Iakovidis, D. Maroulis, and S. Karkanis, "A comparative study of color-texture image features," in *Proc. of IEEE Int. Workshop on Systems, Signal and Image Processing*, pp. 205–209 (2005).

11. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature* **521**(7553), 436–444 (2015).
12. M. Cimpoi, S. Maji, and A. Vedaldi, "Deep filter banks for texture recognition and segmentation," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 3828–3836 (2015).
13. C. Cusano, P. Napoletano, and R. Schettini, "Evaluating color texture descriptors under large variations of controlled lighting conditions," *J. Opt. Soc. Am. A* **33**, 17–30 (2016).
14. F. S. Khan et al., "Evaluating the impact of color on texture recognition," in *Computer Analysis of Images and Patterns*, pp. 154–162, Springer (2013).
15. L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*, John Wiley & Sons (2004).
16. J. Kittler et al., "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.* **20**(3), 226–239 (1998).
17. S. Bianco et al., "Local detectors and compact descriptors for visual search: a quantitative comparison," *Digital Signal Process.* **44**, 1–13 (2015).
18. D. M. Tax et al., "Combining multiple classifiers by averaging or by multiplying?" *Pattern Recognit.* **33**(9), 1475–1485 (2000).
19. "Bt 601: studio encoding parameters of digital television for standard 4:3 and wide-screen 16:9 aspect ratios," (1995).
20. J. Sivic and A. Zisserman, "Video Google: a text retrieval approach to object matching in videos," in *Proc. Ninth IEEE Int. Conf. on Computer Vision*, pp. 1470–1477, IEEE (2003).
21. D. Lee and K. N. Plataniotis, "A taxonomy of color constancy and invariance algorithm," in *Advances in Low-Level Color Image Processing*, pp. 55–94, Springer, Netherlands (2014).
22. T. Gevers and H. Stokman, "Classifying color edges in video into shadow-geometry, highlight, or material transitions," *IEEE Trans. Multimedia* **5**(2), 237–243 (2003).
23. T. Gevers and A. W. Smeulders, "Color-based object recognition," *Pattern Recognit.* **32**(3), 453–464 (1999).
24. G. Finlayson and E. Trezzi, "Shades of gray and colour constancy," in *Color and Imaging Conf.*, pp. 37–41 (2004).
25. C. Cusano, P. Napoletano, and R. Schettini, "Local angular patterns for color texture classification," *Lect. Notes Comput. Sci.* **9281**, 111–118 (2015).
26. A. S. Razavian et al., "CNN features off-the-shelf: an astounding baseline for recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW '14)*, pp. 512–519 (2014).
27. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NIPS '12)*, pp. 1097–1105 (2012).
28. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR* (2014).

Claudio Cusano is an associate professor at the University of Pavia. He took the PhD in computer science at the University of Milano-Bicocca. Since April 2001, he has been a fellow of the ITC Institute of the Italian National Research Council. The main topics of his current research concern 2-D and 3-D imaging, with a particular focus on image analysis and classification, and on face recognition.

Paolo Napoletano is currently a post-doc researcher at the Department of Informatics, Systems, and Communication of the University of Milano-Bicocca. From 2007 to 2012, he worked as a post-doc researcher at the Department of Electronic and Computer Engineering of the University of Salerno, focusing on information retrieval and data mining. In 2007, he received a doctor of philosophy degree (PhD) in information engineering from the University of Salerno with a thesis focused on computational vision and pattern recognition, advisor professor Giuseppe Boccignone. In 2003, he received a master's degree in telecommunications engineering from the University of Naples Federico II, with a thesis focused on transmission of electromagnetic fields. His current research interests focus on image and video analysis and information retrieval.

Raimondo Schettini is a professor at the University of Milano Bicocca (Italy). He is a vice director of the Department of Informatics, Systems, and Communication, and head of the Imaging and Vision Lab. He has been associated with the Italian National Research Council since 1987, where he led the color imaging lab from 1990 to 2002. He has been a team leader in several research projects and published more than 300 refereed papers and six patents about color reproduction, and image processing, analysis, and classification. He is a fellow of the International Association of Pattern Recognition for his contributions to pattern recognition research and color image analysis.