Robust Grape Detector Based on SVMs and HOG Features

Pavel Škrabánek[1]*, Petr Doležel [1]

Abstract
Detection of grapes in real-life images is a serious task solved by researchers dealing with precision viticulture. In the case of white wine varieties, grape detectors based on SVMs classifiers, in combination with a HOG descriptor, have proven to be very efficient. Simplified versions of the detectors seem to be the best solution for practical applications. They offer the best known performance vs. time-complexity ratio. As our research showed, a conversion of RGB images to grayscale format, which is implemented at an image pre-processing level, is ideal means for further improvement of performance of the detectors. In order to enhance the ratio, we explored relevance of the conversion in a context of a detector potential sensitivity to a rotation of berries. For this purpose, we proposed a modification of the conversion, and we designed an appropriate method for a tuning of such modified detectors. To evaluate the effect of the new parameter space on their performance, we developed a specialized visualization method. In order to provide accurate results, we formed new datasets both for tuning and evaluation of the detectors. Our effort resulted in a robust grape detector which is less sensitive to image distortion.

Keywords
computer vision — image recognition — conversion to grayscale — HOG features — SVM classifier — agricultural machinery

[1]*Department of Process Control, University of Pardubice, Pardubice, Czech Republic*
***Corresponding author**: pavel.skrabanek@upce.cz*

Contents

1. Introduction

Detection of grapes in real-life images is a serious task solved by many researchers dealing with precision viticulture [1]. Grape detectors are employed in various applications, e.g. in autonomous vineyard sprayers, harvesters or in the process of yield estimation [2, 3, 4, 5]. Various types of image processing, feature extraction and classification algorithms can be employed when detecting berries or bunches of grapes in RGB images.

A bunch detector designed by Reis at el. [6] employs colour mapping, morphological dilation, and stem detection. It showed correct white wine bunch classification at 90.53 % and red wine at 97.14 %, demonstrating an increase in complexity of detection depending on grape colour. Thus, let us focus only on solutions aimed at white variety detection.

A detector introduced by Berenstein et al. [2] was based on the decision tree algorithm and was applicable both for the bunch and the berry detection. Its detection rate of bunches was 90.45 %, and the detection rate of single grapes was 90.10 %. An exceptional single grape detector was developed by Nuske at el. [3]. It showed overall precision at 98.00 % but recall was lower at 63.70 %. Their detector utilized radial symmetry transformation, Gabor filters, and a *k*-nearest neighbours classifier.

A comparable method [7], aimed at detection of white wine varieties, considered support vector machines (SVMs) classifiers in combination with histograms of oriented gradients (HOG). Specifically, its average accuracy by a 10-fold cross-validation (CV) was 98.23 % for a linear kernel and 98.96 % for a radial basis function (RBF) kernel. Similar results were achieved in their evaluation on real-life images [8].

While excellent, these detectors are computationally intensive to an extent that makes them impractical in viticulture applications. Modification of an image pre-processing (IP), where an input RGB image is first converted to a grayscale format and then adjusted with a linear contrast normalization, can be used to reduce the time complexity of the original solutions [9]. The impact of the normalization step was shown to be negligible in the grape detectors [9, 8], but appropriateness of the grayscale conversion remains to be tested.

The evaluation of the detectors requires sets of labelled object images. These sets might significantly influence evaluation results. Škrabánek et al. [7, 9] discovered inconsistency in the performance of the detectors that depended on the evaluation set. Although the detectors have a high score in the 10-fold CV and an evaluation on real-life images, they do not achieve such results on test sets [7]. The test sets contained distorted positive samples, where 75 % of the positive samples in each test set were artificial samples created by a rotation of original samples. The detectors were tested on two categories of test sets. The poorer grape detection results were present for both categories. For example, the average accuracy of the original detector with the linear kernel was as low as 87.85 % in a test set [7].

Here, we address both the relevance of the grayscale conversion and a potential sensitivity of the detectors to the image rotation. We test a modification of the conversion and propose a parameter tuning method, designed for such modified grape detectors. To evaluate the effect of the new parameter space on their performance, we developed a specialized method for a visualization of the evaluation results. We applied the modified conversion and the tuning method into grape detectors developed by Škrabánek et al. [9], and we evaluated performance of these modified versions. In order to provide accurate results, we formed new datasets both for tuning and evaluation of the detectors. Our improvements resulted in a robust grape detector which is less sensitive to image distortion.

2. Materials and Methods

2.1 Original Work on Grape Detectors
The presented study evaluates and modifies previous research published in [7, 9]. Herein, we provide a brief summary of the published grapes detectors and the test sets.

2.1.1 Objective of the Detectors
The grape detectors were aimed at the recognition of grapes of white varieties in object images. RGB object images I of dimensions 40×40 px (pixel) were considered in [7, 9, 8].

The object images were square viewports of source images of a resolution 1936×1288 px, 24 bit.

The detectors distinguished between two classes y: 'berry' and 'not berry'. The class 'berry' was called 'positive' and the class 'not berry' was called 'negative'. Object images belonging to the class 'positive' contained berries of the circle shape of a diameter ranging between 30 and 40 px. Moreover, the middle of the berries were required to be placed in the middle of the object images with a tolerance ± 1 px. Object images, which did not satisfy this condition, belonged to the class 'negative'.

2.1.2 Structure of the Detectors
In computer vision, detection of objects in images usually consists of four successive steps. The first step is acquiring an object image I from a large real-life image; the second step is the IP resulting in a modified image Y; the third one is extraction of features; and the final step is classification of the object image using a feature vector $\mathbf{x}$. However, the grape detectors introduced in [7, 9] consist specifically of three parts only; from the IP, a features descriptor and a classifier. Although the detectors differ in structure of the IP, used feature descriptors, and settings of classifiers, they have the same arrangement of a vision pipeline (Fig. 1). Parts of the detectors are described in the context of our previous works in further details.
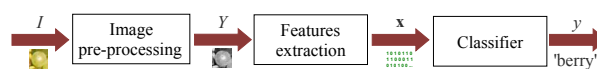

Figure 1. Vision pipeline of the grape detectors.

Image Pre-processing   IPs of the original detectors, labelled as O [7], consist of two steps; the conversion of input RGB images to grayscale format followed by a linear contrast normalization on the range $[0, 1]$. Simplified versions, which were introduced in [9], skip either the contrast normalization (simplified version 1 or S_1) or both operations (simplified version 2 or S_2). In O and S_2, the conversion is carried out according to ITU-R recommendation BT.601 [10]. This conversion belongs to techniques based on weighted means of all three colour channels [11]. It means that the conversion of the input RGB image I from the RGB model to the grayscale format is realized by eliminating the hue and saturation information, while retaining the luminance.

Features Extraction   Two types of features, a vector of normalized pixel intensities and HOG, were considered in [7]; however, only the HOG features have proven to be convenient for the detection of grapes of white varieties [7, 9, 8]. Thus, only the detectors based on the HOG features will be considered further. It means that the feature vectors $\mathbf{x}$ are extracted from Y using the HOG descriptor [12]. A standard setting of the descriptor has demonstrated to be sufficient. Specifically, a linear gradient voting into 9 bins in $0° - 180°$; cells of size

6×6 px; blocks of 2×2 cells; and 1 overlapping cell between adjacent blocks in both directions were used in [7, 9, 8].

Classifier The aim of classifiers in the grape detectors is a judging the classes y of the object images I using feature vectors **x**. SVM classifiers [13] with the linear and the RBF kernel functions were used in [7, 9]. Regardless of the used kernel function, the performance of a SVM classifier is influenced by a regularization constant C. Performance of a classifier with the RBF kernel is further influenced by a kernel width σ. A grid search algorithm [14] combined with the 10-fold CV was used to find settings of these parameters giving the maximal recognition accuracy [7].

2.1.3 Training and Evaluation of the Detectors

Five training sets were introduced in [7]. The i-th training set was denoted as T-i, where $i \in X$, and $X = \{1, 2, \ldots, 5\}$. Each training set consisted of 288 unique 'positive' and 288 unique 'negative' samples. The training set T-3 was used for the training of the detectors [7, 9, 8].

Three kinds of evaluation methods were considered by the evaluation of the original detectors O [7, 8]: the 10-fold CV, an evaluation on test sets, and an evaluation on cut-outs of one vineyard photo. The simplified detectors S_1 and S_2 were evaluated using the test sets [9] and the cut-outs [8]. In all cases, three performance measures m were used for the evaluation:

$$\text{accuracy} = \frac{|TP| + |TN|}{|TP| + |FN| + |TN| + |FP|} \cdot 100, \tag{1a}$$

$$\text{precision} = \frac{|TP|}{|TP| + |FP|}, \tag{1b}$$

$$\text{recall} = \frac{|TP|}{|TP| + |FN|}, \tag{1c}$$

where $|TP|$ is the number of correctly classified 'positive' samples, $|FN|$ is the number of misclassified 'positive' samples, $|FP|$ is the number of misclassified 'negative' samples, and $|TN|$ is the number of correctly classified 'negative' samples [15]. Naturally, for the cut-outs, unbiased variants of the measures have been used [8].

While just the training sets were needed for the 10-fold CV, appropriate datasets must be created for the remaining two evaluation methods. Creation of these datasets was sufficiently described in [7]; however, datasets used for the evaluation on test sets should be detailed here. For this purpose, two types of datasets were created: an *environment type* E, and a *grape type* G. Five sets of each type were formed [7]. The i-th test set of the type E was denoted as E-i and the i-th test set of the type G as G-i, where $i \in X$.

Each test set consisted of 200 'positive' and 200 'negative' samples. The sets were based on one vineyard row photo which was not used for creation of the training sets. To form a single test set, 50 unique 'positive' and 200 unique 'negative' samples were used. Each test set was further extended by artificial 'positive' samples [16]. The 'positive' samples were created by turning of the images through an angle φ, where $\varphi \in \{0, \pi/2, \pi, 3\pi/2\}$.

The difference between these two types of test sets consisted in the selection of the 'negative' samples. The 'negative' samples in G were composed solely of incomplete grape berries of a diameter ranging between 30 and 40 px, while the 'negative' samples in E were based on the environment only, and they did not capture even the smallest piece of the targeted berry. Examples of 'positive' samples as well as of both types of 'negative' samples are shown in Fig. 2.

Figure 2. Examples of object images of the class: a) 'positive' , b) 'negative' - grape type, c) 'negative' - environment type.

2.2 Robust Grape Detector

While the original, as well as the simplified, grape detectors showed excellent performance by the 10-fold CV [7] and by the evaluation on the cut-outs [8], considerably worse results were obtained by their evaluation on the test sets [7, 9]. For example, the original version of the detector with the linear kernel reached the following average score by the 10-fold CV: precision = 0.980, accuracy = 98.23 %, and recall = 0.987. The following average results were obtained for the same detector on the test sets of the type E: precision = 0.977, accuracy = 87.85 %, and recall = 0.776. Similar results were obtained on both types of test sets for all versions of detectors.

The drop in accuracy and recall might be caused by different reasons. Generally, the most common source of a discrepancy in evaluation results is an inadequacy of datasets used by the evaluation. In the case of the grape detectors, a distortion of positive samples caused by the rotation might be considered as well. Indeed, the HOG features are not rotation invariant [12]. Since a potential sensitivity of the detectors on the distortion caused by the rotation of images is undesirable, we started a research aimed on enhancement of the detector's robustness.

Two principally different means might help to improve the detector's performance in the presence of a distortion: either a more appropriate training set could be used for the training, or a modification in the vision pipeline could be done. In our research, we focused on the modification of the pipeline (subsection 2.2.1). Unfortunately, three additional tenable parameters were introduced by the presented modification, which considerably complicated the search for the optimal setting of the modified detector. In order to face this

disadvantage, we developed a specialized tuning methodology (subsubsection 2.2.3). The methodology was based on a visualization method which we designed for this purpose (subsubsection 2.2.2).

2.2.1 Modification of the Detector

Although the selection of a good feature vector is widely recognized to be fundamental when designing image recognition systems, the IP may also significantly influence the performance of a final solution. Thus, an intelligent use of the IP can provide benefits and solve problems that ultimately lead to better local and global feature detection [17].

IPs of the original grape detectors consisted of two operations: the conversion of RGB images to the grayscale format, and the linear contrast normalization. Our later research has positively shown that the skipping of the contrast normalization does not influence the performance of the detectors at all [9]. Nonetheless, the relevance of the conversion in the detectors was not disproved.

Importance of the conversion of RGB images to the grayscale format, in connection with the image recognition systems, was studied for example by Kanan and Cottrell [11]. They have shown that a method used by the conversion of colour images to the grayscale format may significantly influence performance of the image recognition system, even when using robust descriptors. Further, they have pointed out the fact, that different tasks involve different conversion methods. For object recognition tasks, they recommend techniques based on weighted means of the red, green, and blue image channels.

Considering their outcomes, we proposed to use a general formula for the conversion based on the weighted means. The main motivation for its use is the fact, that this formula allows better control of the conversion, which consequently may allow us to improve the performance of the detectors. The general formula can be written as

$$I^* = w_R I_R + w_G I_G + w_B I_B, \tag{2}$$

where I_R, I_G and I_B are intensity images of the red, green and blue components of the RGB image I; and w_R, w_G, and w_B are weights of the colour components in the resulting grayscale image I^*. It holds that $w_R, w_G, w_B \in [0,1]$, and $w_R + w_G + w_B = 1$. Since the conversion (2) is the single operation performed within IPs of the modified detectors, it holds that $Y = I^*$.

We modified the simplified version S_1 of the grape detector using the general formula (2), i.e. the standard conversion was replaced by this formula. The rest of the vision pipeline was not changed, i.e. the new detector consisted of the HOG descriptor and the SVM classifier. We named the new detector as a *robust grape detector*, or simply R.

Considering the essence of the proposed modification, it is apparent that the modification does not increase the computational complexity of the detector. Inherently, it may not cause a worsening of its performance. Indeed, the setting of the weighting coefficients according to the ITU-R recommendation BT.601 is just one of an infinite number of possible settings.

2.2.2 Performance Visualization for Better Understanding of Tuning Parameter Relevance

While searching for the optimal setting of an image recognition system, an appropriate variant of the grid search algorithm is typically used. For just one or two tuneable parameters, meaningfulness of evaluation results can be assessed on the basis of the raw data directly by a specialist. However, the assessment is becoming challenging when increasing the number of tuneable parameters.

Considering the parameters of the HOG descriptor to be non-tuneable, the original and the simplified versions of the detectors have only one or two tuneable parameters (depending on the used kernel function). The fixation of the HOG descriptor setting has its origin in practical reasons. The setting of the HOG descriptor, which was summarized in subsubsection 2.1.2, is appropriate in terms of exactitude vs. computational complexity. Thus, this fixed setting is used for the robust detectors as well.

The original tuneable parameters, C and σ, were extended by three additional parameters, w_R, w_G, and w_B, in the case of the robust grape detectors. These parameters were introduced by the formula (2). Thus, the tuning of a robust grape detector setting using a grid search algorithm, results in a large amount of data. An analysis of such quantities of data in the raw numeric form is really not convenient for humans. In order to facilitate the analysis, we suggested a ternary diagram for a visualization of the data. Specifically, we used the diagram to show a prospective influence of the weighting coefficients, w_R, w_G, and w_B, on the performance of an overall solution for a fixed setting of the remaining tuneable parameters.

The ternary diagram is a graph which consists of an equilateral triangle in which a given plotted point represents the relative proportions (a, b, c) of three end-members (A, B and C), usually expressed as percentages (do not confuse the end-member C with the regularization constant C). Moreover, the sum of the relative proportions is equal to a given value, e.g. for percentages, it holds that $a + b + c = 100\%$. The axis related to the member A is the left arm of the triangle. The relative proportion of A, a, is plotted on the axis where a increases downwards. The same principle is used for the remaining two axes where the bottom axis is related to the component B and the right one to the component C. The relative proportion of B increases in the right direction, and the relative proportion of C upwards. A dependent variable can be represented in different ways, e.g. using contour plot or shaded surface. For more information, see [18].

In our case, the three end-members, A, B and C, are the intensity images of the colours, I_R, I_G and I_B, respectively. The relative proportions, a, b and c, are the weights, w_R, w_G and w_B, where $w_R, w_G, w_B \in [0,1]$. Moreover, the weights are bounded by the condition $w_R + w_G + w_B = 1$. The ternary diagram is aimed to be used as a supporting tool by the evaluation

of the grape detectors. Thus, a performance measure m is the dependent variable to be displayed. Considering our previous experience, we suggested to use the shaded surface for the visualization of the results. This type of diagram gives the better idea about the influence of the weights on the performance measure.

Let us denote a setting of the weighting coefficients **w** as an ordered triple of the weights w, i.e. $\mathbf{w} = (w_R, w_G, w_B)$. Further, let us form a finite set W of settings **w** for the purpose of the graph construction, and let us call the set *grid*. The grid should uniformly cover the surface bounded by the triangle. It means that a step Δw, of a fixed size, has to be used by forming the grid. Let us express the step as $\Delta w = \frac{1}{p}$, where $p > 0, p \in \mathbb{N}$. Thus, the weighting coefficients w can take any value from $\{0, \Delta w, 2\Delta w, \ldots, 1\}$; however, a combination of the coefficients in (w_R, w_G, w_B) is bounded by the condition $w_R + w_G + w_B = 1$. It means that the grid W is a set of all admissible settings **w**.

In order to construct the ternary diagram, evaluation of a detector using a measure m has to be achieved for $\forall \mathbf{w} \in W$. Naturally, the classifier has to be trained for each of these settings separately. Settings of the classifier (C for linear kernel and C, σ for RBF kernel) must not be changed within the training-evaluation process. Once the training-evaluation process is performed for all the admissible settings **w**, construction of one diagram can be executed, where the diagram shows dependence of the used performance measure m on the weights w_R, w_G, and w_B for one particular setting of C or C, σ. An example of the diagram is shown in Fig. 3 where the recall of a grape detector is displayed.
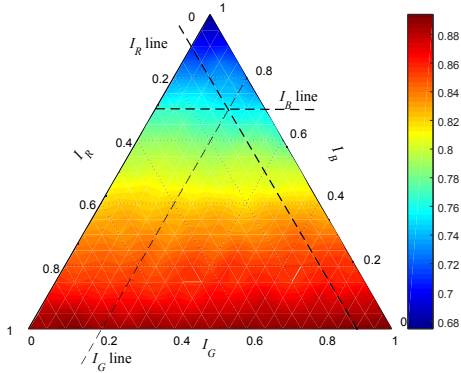
Figure 3. Demonstration of working with the ternary diagram.

Fig. 3 can be used also for explanation of how to work with the diagram. Let us suppose that the recall for $w_R = 0.1$, $w_G = 0.2$ and $w_B = 0.7$ is required to be determined. The reading of the recall value can be done using auxiliary lines which are plotted using dashed lines in Fig. 3. Each line is parallel with one of the sides of the triangle. The lines are named with respect to referred components, i.e. 'I_R line' is related to the intensity images of the red colour I_R, 'I_G line' to the intensity images of the green colour I_G, and 'I_B line' to

the intensity images of the blue colour I_B.

Positions of the lines are given by the weights w. In this example, 'I_R line' passes the left axis at the point 0.1, 'I_G line' passes the bottom axis at the point 0.2, and 'I_B line' passes the right axis at the point 0.7. The intersection of the lines positively determines the recall. The numeric value can be estimated using the colour bar. In this example, the recall is approximately 0.78. It might be noted here that only two auxiliary lines are necessary for reading of the dependent variable. Indeed, the proportion of a third component is always positively determined by the constraint of their sum, in this case by $w_R + w_G + w_B = 1$.

2.2.3 Tuning Methodology

The goal of our research was the development of a grape detector which would be invariant to the distortion caused by the rotation. One of the key steps within the development process is finding the setting of all tuneable parameters giving the best performance according to a criterion. The search for the optimal setting is usually executed using a variant of the grid search algorithm [14].

The grid search algorithms ensure a systematic evaluation of an image recognition system performance. Usually, the grid search is combined with the CV. In such a case, the evaluated system is trained on a part of a dataset and evaluated on the rest of the dataset. The training is carried out for various settings of all tuneable parameters. The settings might be assigned according to a rule or directly defined by an expert. The setting giving the best score is considered to be optimal. In order to obtain more accurate results, the search can be performed repeatedly. The search is then carried out with a finer resolution in a scaled down area. An area promising best results is selected for the finer search.

In order to develop a methodology aimed at tuning of the robust detectors, we adopted the basic principles of the grid search algorithms. Specifically, the training-evaluation process is carried out for various settings of all tuneable parameters within the search process. However, our methodology requires involvement of a computer vision specialist. Further, the evaluation of the image recognition system should be performed on a dataset affected by the target distortion. For that reason, the commonly used CV was replaced by the evaluation of the system on a specialized dataset.

In order to ensure flexibility of the method, we prepared the methodology for a multiple criteria usage. The proposed methodology allows combination of several performance measures when searching for the optimal setting. In the case of the multiple criteria, a priority of the performance measures has to be determined in advance. For the detector with the RBF kernel, the methodology consists of following steps:

1. Select performance measures and give them priorities. In such a way, a finite set M of measures m is created where each measure is paired with a priority.

2. Define admissible settings of all tuneable parameters X_C, X_σ, W, where X_C is a finite set of all admissible

settings of C, and X_σ is a finite set of all admissible settings of σ. Such a way, a parameter space Θ of features θ is formed, where $\theta \in \Theta$, and $\theta = (C, \sigma, \mathbf{w})$.

3. Perform the training-evaluation process $\forall \theta \in \Theta$ using $\forall m \in M$.

4. Display the obtained results using the diagram. Such a way, $|M| \times |X_C| \times |X_\sigma|$ graphs are obtained, where $| \bullet |$ denotes a cardinality of a set.

5. Manually evaluate the obtained results using the graphs. Identify combinations of C and σ leading to senseless results. Eliminate all settings θ containing the offending combinations of C and σ from the further processing, i.e. a new parameter space $\hat{\Theta}$ is formed where $\hat{\Theta} \subset \Theta$.

6. For each performance measure $m \in M$, find the setting θ_m^* giving the best score according to

$$\theta_m^* = \arg\max_{\forall \theta \in \hat{\Theta}} m(\theta). \tag{3}$$

7. Determine a globally optimal setting θ^* on the basis of all θ_m^*. Within this step, the priority of the measures must be taken into account; however, the functional dependences shown in the appropriate graphs must be considered as well.

This methodology can be also applied on the detector with the linear kernel. Naturally, the variable σ should be ignored in this case.

2.3 Design of Evaluation Experiments
In the experimental part, we evaluated relevance of the grayscale conversion and the potential sensitivity of the detectors to the image rotation. For this purpose, new statistically relevant datasets were created.

2.3.1 Assessment of Conversion Importance
The assessment of the conversion relevance was one of the main goals of the presented work. Two versions of the conversion were considered in the grape detectors. While the grape detectors S_1 employ the standard conversion according to the ITU-R recommendation BT.601, the robust detectors R are based on the generalized conversion (2). The detectors S_2 do not perform any conversion within the IP.

Performances of all three variants of the detectors have to be confronted concerning the assessment of the conversion relevance. In order to keep comparability of the results, the detectors should be tuned the same way. Thus, when tuning S_1 and S_2, the methodology presented in subsubsection 2.2.3 should be used in appropriately modified form. It means that steps 4 and 5 have to be left out when searching for their settings.

2.3.2 Dataset for Tuning
The proposed tuning methodology requires one training and one evaluation set. In order to keep continuity with our previous research, the training set T-3 was used in the training phase. Within the evaluation phase, a specialized (tuning) dataset should be used. The tuning set should be large enough and it should be affected by the target distortion of the 'positive' samples. Further, both types of the 'negative' samples, E and G (subsubsection 2.1.3), should be equally represented in the set. Searching for the optimal setting on such datasets may guarantee finding a setting which would be appropriate to given requirements on the robust detector.

For this purpose, a tuning set on ten photos was created. These photos were captured under the condition specified in [7]. They were captured at six different locations. None of these photos were used while forming the training set. The tuning set consisted of labelled RGB object images of size 40×40 px. The labelled object images were created from the photos using an editor [19]. To create the dataset, 1000 unique 'positive' and 4000 unique 'negative' samples were gathered. The set was extended using artificial 'positive' samples created by turning the images through the angle φ, i.e. it consists of 4000 'positive' and 4000 'negative' samples.

2.3.3 Datasets for Evaluation
In subsubsection 2.1.3, the evaluation test sets, E and G, were mentioned. Each of these sets consisted of only 400 samples, which seems to be insufficient for a credible assessment of the meaning of the conversion. In order to get meaningful results, we formed new *expanded test sets*. These sets were created in the same spirit as the original test sets. Continuity of the marking was also maintained; i.e. expanded test sets of the environment type were labelled as EX, and expanded test sets of the grape type as GX. Two expanded test sets of each type were formed. In addition, a new type of test set was introduced. This set was labelled as *expanded standard test set* or SX. The set was not affected by the distortion.

All the expanded test sets comprised of labelled RGB object images of the size 40×40 px. The labelled object images were created from vineyard row photos using the editor [19]. The sets EX and GX were based on a collection of ten unique vineyard row photos. Ten different photos were used for SX creation. The photos did not match with the photos used when creating the training and the tuning set. The photos were captured at six different locations under the conditions specified in [7].

An expanded test set, either EX or GX, consisted of 500 unique 'positive' and 2000 unique 'negative' samples. The sets were extended using artificial 'positive' samples created by turning the images through the angle φ, i.e. they consisted of 2000 'positive' and 2000 'negative' samples. The selection of the 'positive' and of the 'negative' samples followed the criteria stated in subsubsection 2.1.3. Sets EX-*i* and GX-*i* with the same index *i* shared the same collection of 'positive' samples. The standard set SX consisted of 2000 unique 'positive' and 2000 unique 'negative' samples.

2.3.4 Inquire into the Sensitivity to the Rotation

The potential sensitivity of the detectors to the image rotation was the second issue to be investigated. Suspicion on the sensitivity came from the disproportion between the results acquired by the 10-fold CV or by the evaluation on the cut-outs, and the results obtained by the evaluation on the test sets. The new expanded dataset, EX, GX, and SX, allowed us a detailed exploration of this issue. A comparison of evaluation results obtained on the expanded dataset with evaluation results obtained on the original datasets is the key. In order to get comparable results, the detectors S_1, S_2, and R tuned according to the proposed methodology, must be evaluated also on the sets E and G.

A high sensitivity of an assessed detector to the rotation would be noticeable form a comparison of evaluation results obtained on EX (GX) with evaluation results obtained on E (G). Similarity of these results would indicate the high sensitivity. Worse evaluation results obtained on EX (GX), rather than on E (G), would also confirm the high sensitivity. A low sensitivity of the detector would be visible from the comparison of evaluation results obtained on EX and GX with evaluation results obtained on SX. A worse performance on EX or GX, rather than on SX, would confirm the low sensitivity. All other results would signify its insensitivity to the rotation. In order to eliminate a potential influence of the proposed tuning methodology on the evaluation results, the new results obtained for E and G were compared with the original results.

3. Results and Discussion

3.1 Optimal Settings of Detectors

Optimal settings of the detectors were determined using the presented methodology. In order to keep the continuity of our work, the performance measures (1) were used for the evaluation of the detector performance. It means that $M = \{accuracy, precision, recall\}$. The accuracy was used for the tuning of all former versions of the detectors. For this reason, the accuracy was chosen as the primary measure with the highest priority. Detection of all grapes in a photo is essential for applications such as yield estimation. Thus, the recall was taken as the secondary measure; and consequently, the precision was considered to be the tertiary one.

Depending on the version, the detectors have up to five tuning parameters. While the weighting coefficients w are bounded by the conditions $w_R, w_G, w_B \in [0, 1]$ and $w_R + w_G + w_B = 1$, the regularization constant C, likewise the kernel width σ, must be positive. Based on our previous experience, the following settings of the parameters were used within the search process: $X_C = \{1, 10, 100, 1000\}$, $X_\sigma = \{1, 10, 20, 30, 40, 100\}$, and $p = 20$. In such a way, sets Θ of all admissible settings θ were formed.

While the search for the optimal setting of S_1 and S_2 according to the proposed methodology does not require any intervention of human, the search for the optimal setting of R cannot be performed without the computer vision expert.

3.1.1 Optimal Setting of Robust Detector with RBF Kernel

On the basis of data obtained within the training-evaluation process, 72 graphs were created. On the basis of their mutual comparison, we discovered that σ has much higher influence on the performance measured by accuracy and recall than C. Further, we found that both, C and σ, influence the performance only slightly from the perspective of precision. The main trends captured by the graphs can be demonstrated on graphs obtained for an arbitrary chosen setting of C and $\sigma \in \{1, 10, 20, 30, 40\}$. For these purpose, we chose results obtained for these values of σ and $C = 10$. The results are shown for all three measures in Figs. 4-8.

The analysis of the 72 diagrams pointed out an abnormality in the results obtained for $\sigma = 1$ and $\forall C \in X_C$. The abnormality is clearly visible in Fig. 4, where the performance of the robust detector with the RBF kernel is shown for $C = 10$ and $\sigma = 1$. For example, accuracy (Fig. 4 a)) is 0.5 for the majority of the settings $\mathbf{w} \in W$; nevertheless, excellent scores were achieved for some of them. It is apparent that a very small change in the setting of the weighting coefficients would lead to a drastic drop in accuracy. The diagrams for the other measures (Fig. 4 b) and c)) show similar discrepancies. Very similar graphs were obtained for $\forall C \in X_C$ in combination with $\sigma = 1$.

The seriousness of the discrepancies is even more apparent when comparing graphs obtained for $\sigma = 1$ with graphs obtained for higher values of σ. Graphs obtained for $C = 10$ and $\sigma > 1$ (Figs. 5-8) show consistent trends for all performance measures. Such trends were obtained for $\forall C \in X_C$, and $\forall \sigma \in \{\sigma \in X_\sigma | \sigma \neq 1\}$. It is clear that $\sigma = 1$ cannot guarantee robustness of the final solution. Thus, all evaluation results, obtained for $\sigma = 1$, were eliminated from the further processing, i.e. $\hat{\Theta} = \{(C, \sigma, \mathbf{w}) \in \Theta | \sigma \neq 1\}$. On the basis of the formula (3), results summarized in Table 1 were obtained.

Table 1. Optimal settings of the robust detector with the RBF kernel according to different measures. The results are ordered according to the priority of the measures.

	C	σ	$\mathbf{w}$
$\theta^*_{accuracy}$	10	30	(0.95, 0.05, 0.00)
θ^*_{recall}	10	40	(0.95, 0.05, 0.00)
$\theta^*_{precision}$	1	10	(0.00, 0.85, 0.15)

Diverse results were obtained for the measures $m \in M$. Although the measures have predefined priorities, meaningfulness of the results should be always considered before the globally optimal setting θ^* is determined. It is apparent from the graphs (Figs. 5-8) that the parameters $\sigma, \mathbf{w}$ had negligible influence on precision. When comparing all obtained graphs, we found that also C hardly influenced precision. Thus, precision was abandoned within the final assessment.

According to the remaining two measures, the setting providing the best performance was $\mathbf{w} = (0.95, 0.05, 0.00)$ and $C = 10$. However, the obtained results did not allow the direct determination of the optimal value of σ. Since

accuracy had the higher priority, the optimal setting of σ was determined according to this measure. As is apparent in Fig. 7 and Fig. 8, there was only insignificant difference in the performance measured using recall for $\sigma = 30$ and $\sigma = 40$. Thus, we selected $\theta^* = (10, 30, (0.95, 0.05, 0.00))$ to be the globally optimal setting for the robust detector with the RBF kernel.

3.1.2 Optimal Setting of Robust Detector with Linear Kernel

The analysis of the data obtained within the training-evaluation process was much simpler in the case of the robust detector with the linear kernel. At first, only 12 graphs were created on the basis of the obtained data. Secondly, the regularization constant C did not influence the performance of the detector, i.e. identical graphs were obtained for $\forall C \in X_C$. Thus, for the next explanation, we selected results obtained for $C = 1$ (Fig. 9).

It is apparent in the graphs (Fig. 9) that there was no anomaly in the evaluation results, and the detector reached high scores for all used measures. It means that $\hat{\Theta} = \Theta$. According to the formula (3), settings summarized in Table 2 were determined to be optimal for $\forall m \in M$.

Table 2. Optimal settings of the robust detector with the linear kernel according to different measures. The results are ordered according to the priority of the measures.

	C	$\mathbf{w}$
$\theta^*_{\text{accuracy}}$	1	(0.95,0.05,0.00)
θ^*_{recall}	1	(0.95,0.05,0.00)
$\theta^*_{\text{precision}}$	1	(0.00,0.90,0.10)

As is apparent in Fig. 9, the weighting coefficients $\mathbf{w}$ had almost no influence on precision. Thus, just accuracy and recall were taken into account when the final decision was made. Since identical optimal settings were obtained for both measures, we can positively recommend the setting $\theta^* = (1, (0.95, 0.05, 0.00))$ as the globally optimal setting for the robust detector with the linear kernel. However, we should point out here that any $C \in X_C$ might be used as well.

3.1.3 Summary of Optimal Settings

Herein, we provide a summary of optimal settings for various versions of the detectors. The summary can be found in Table 3.

3.2 Evaluation of Detectors

In order to investigate the opened issues, evaluations of the detectors on the expanded, as well as on the original datasets were done. The results obtained on the expanded sets, EX, GX and SX, are summarized in Table 4. The results obtained on the original sets, E and G, are stated in Table 4 and Table 5, respectively.

3.3 Discussion of Evaluation Results

Two main issues were opened in this article. Our findings are presented in following text.

Table 3. Optimal settings for various versions of the detectors where: S_2 - detector without any conversion, S_1 - detector with the conversion according to ITU-R recommendation BT.601, R - detector with optimally set-up generalized conversion (2).

version	**Linear kernel**		**RBF kernel**		
	C	$\mathbf{w}$	C	σ	$\mathbf{w}$
S_2	1	-	1	40	-
S_1	1	-	10	30	-
R	1	(0.95,0.05,0.00)	10	30	(0.95,0.05,0.00)

3.3.1 Assessment of Conversion Importance

The importance of the conversion in the grape detectors was assessed on the basis of results obtained on EX, GX and SX (Table 4). Let us consider the results obtained for the detectors with the linear kernel at first. From comparison of S_1 and R with S_2, it follows that the standard conversion according to the ITU-R recommendation, as well as the generalized conversion according to (2), enhanced scores obtained using accuracy and recall on EX and GX. The improvement is more evident for recall. Precision was slightly better for S_1 for these datasets; however, a small downturn in precision was registered for R. Considering the significant improvement in recall, the conversions seemed to be valuable parts of the detectors. However, the results obtained on SX did not confirm these outcomes. Thus, using of a conversion for detectors with the linear kernel cannot be positively recommended.

In the case of the detectors with the RBF kernel, the results spoke definitely on behalf of the conversions. The detectors S_1 and R out performed the detector S_2 in accuracy, and especially in recall, on all the expanded sets, i.e. on EX, GX and SX. Within all the experiments, the changes in precision were marginal. Let us focus now only on the detectors S_1 and R. From the results, it is apparent that R always out performed S_2 in accuracy and recall. We considered the downturn of precision of R to be marginal. Indeed, it never fell below 0.9800. Thus, from the perspective of the performance, the robust detector with the RBF kernel can be positively recommended as the most reliable solution. With accuracy $\geq 94.37\,\%$, recall ≥ 0.9010 and precision ≥ 0.9816, the robust grape detector with the RBF kernel is fully comparable with the state-of-the-art solutions aimed at the detection of single grapes of white varieties.

3.3.2 Inquire into the Sensitivity to the Rotation

As the first step, we compared the results obtained for the detectors S_1 and S_2 (Tables 5-6), tuned according to the proposed methodology, with the original results [9]. We found that the new results are almost identical to the original ones. We came to the conclusion that the tuning methodology did not influence the results significantly.

In the second step, we compared the results achieved by the detectors on type matching sets, i.e. results obtained

Table 4. Evaluation of the detectors trained on T-3, setup according to Table 3, on new datasets of environment type (EX), grape type (GX), and standard type (SX), where $|P| = 2000$ and $|N| = 2000$. Following versions of the detectors where evaluated using the measures (1): S_2 - detector without any conversion, S_1 - detector with the conversion according to ITU-R recommendation BT.601, R - detector with the generalized conversion (2).

kernel	version	EX-1			EX-2			GX-1			GX-2			SX		
		accuracy	precision	recall	accuracy	precision	recall	accuracy	precision	recall	accuracy	precision	recall	accuracy	precision	recall
linear	S_2	91.73%	0.9755	0.8560	90.93%	0.9712	0.8435	91.87%	0.9788	0.8560	91.10%	0.9751	0.8435	95.23%	0.9804	0.9230
	S_1	94.15%	0.9768	0.9045	92.07%	0.9714	0.8670	94.63%	0.9869	0.9045	92.40%	0.9786	0.8670	94.73%	0.9833	0.9100
	R	94.20%	0.9580	0.9245	93.07%	0.9600	0.8990	95.07%	0.9757	0.9245	93.83%	0.9756	0.8990	94.70%	0.9671	0.9255
RBF	S_2	93.27%	0.9909	0.8735	92.20%	0.9879	0.8545	93.25%	0.9904	0.8735	92.20%	0.9879	0.8545	95.83%	0.9909	0.9250
	S_1	94.87%	0.9854	0.9110	93.73%	0.9888	0.8845	95.33%	0.9940	0.9120	93.63%	0.9866	0.8845	95.85%	0.9914	0.9250
	R	95.85%	0.9816	0.9345	94.37%	0.9852	0.9010	96.35%	0.9920	0.9345	94.50%	0.9879	0.9010	95.87%	0.9873	0.9295

Table 5. Evaluation of the detectors trained on T-3, setup according to Table 3, on original datasets of environment type (E), where $|P| = 200$ and $|N| = 200$. Following versions of the detectors where evaluated using the measures (1): S_2 - detector without any conversion, S_1 - detector with the conversion according to ITU-R recommendation BT.601, R - detector with the generalized conversion (2).

kernel	version	E-1			E-2			E-3			E-4			E-5		
		accuracy	precision	recall	accuracy	precision	recall	accuracy	precision	recall	accuracy	precision	recall	accuracy	precision	recall
linear	S_2	86.50%	1.0000	0.7300	87.25%	0.9869	0.7550	85.50%	0.9671	0.7350	86.25%	0.9801	0.7400	84.50%	0.9859	0.7000
	S_1	88.50%	0.9873	0.7800	88.50%	0.9753	0.7900	87.50%	0.9573	0.7850	87.75%	0.9748	0.7750	87.00%	0.9868	0.7500
	R	88.25%	0.9693	0.7900	90.25%	0.9763	0.8250	88.75%	0.9586	0.8100	88.25%	0.9693	0.7900	87.75%	0.9809	0.7700
RBF	S_2	88.00%	0.9935	0.7650	89.50%	0.9938	0.7950	88.00%	0.9872	0.7700	87.75%	0.9935	0.7600	86.25%	1.0000	0.7250
	S_1	89.50%	1.0000	0.7900	90.25%	0.9879	0.8150	89.00%	0.9815	0.7950	89.25%	0.9937	0.7900	87.50%	1.0000	0.7500
	R	88.50%	0.9936	0.7750	89.50%	0.9877	0.8000	88.50%	0.9812	0.7850	88.00%	0.9634	0.7900	86.75%	0.9933	0.7400

on E (G) were compared with results obtained on EX (GX). We found that all detectors tuned according to the proposed methodology achieved significantly better results on the expanded sets EX and GX (Table 4), rather than on the original sets E (Table 5) and G (Table 6). It is evident that the poor performance of the detectors, which was reported in [7, 9], was mainly due to the inappropriateness of the sets E and G. Thus, we recommended using exclusively the new expanded datasets for the evaluation.

In the third step, we compared the results achieved by the detectors on the expanded sets (Table 4). For both types of the kernels, we observed a connection between the performance of the detectors and the grayscale conversion. While the detectors S_2 (without the conversion) had considerably lower recall on EX and GX than on SX, the robust detectors reached almost identical results on all these sets. The detectors S_1 showed results on the borderline between S_2 and R. Thus, we came to the conclusion that the detectors are sensitive to the image rotation; however, the sensitivity can be suppressed by the grayscale conversion. We further found that the robust detector with the RBF kernel is almost resistant to this distortion.

4. Conclusion

The grape detectors based on SVMs classifiers and HOG features were appropriate solutions for detection of single grapes of white varieties, supported by excellent results by the 10-fold CV and the evaluation on the real-life images. However, results obtained by the evaluation on the test sets prompted a thorough examination of the detector performance for a confirmation of its expected merits. Our results showed that the grayscale conversion should be excluded when the SVM classifier with the linear kernel is used in combination with the HOG features. Using the linear kernel and skipping the entire image pre-processing, ensures a low time complexity of the final solution, while keeping excellent performance on standard datasets. Such solution might be used in applications where worse performance under unfavourable conditions is not critical, e.g. in autonomous vineyard sprayers.

The robust grape detector with the RBF kernel is fully comparable with the state-of-the-art solutions aimed at detection of single grapes of white varieties. The detector has greater time complexity than a detector without the grayscale conversion, but this disadvantage is counterbalanced by its excellent performance. Since the robust grape detector with the RBF kernel provides excellent results under standard conditions, as well as under unfavourable conditions, we recommended its usage in applications where the high accuracy and recall are essential, e.g. for the yield estimation.

In the presented application, the modification of the grayscale conversion proved to be valuable. We believe that its usage is not limited to the grape detection. The modification, together with the tool set for parameter optimization, as well as the novel visualization method introduced in this contribution, might be used in various application areas.

Acknowledgments

Table 6. Evaluation of the detectors trained on T-3, setup according to Table 3, on original datasets of grape type (G), where $|P| = 200$ and $|N| = 200$. Following versions of the detectors where evaluated using the measures (1): S_2 - detector without any conversion, S_1 - detector with the conversion according to ITU-R recommendation BT.601, R - detector with the generalized conversion (2).

| kernel	version	G-1 accuracy	precision	recall	G-2 accuracy	precision	recall	G-3 accuracy	precision	recall	G-4 accuracy	precision	recall	G-5 accuracy	precision	recall
linear	S_2	86.25%	0.9560	0.7600	81.00%	0.9697	0.6400	87.00%	0.9933	0.7450	85.75%	0.9673	0.7400	84.75%	0.9728	0.7150
	S_1	86.25%	0.9618	0.7550	79.25%	0.9606	0.6100	86.00%	0.9865	0.7300	87.50%	0.9870	0.7600	84.75%	0.9793	0.7100
	R	87.75%	0.9576	0.7900	80.25%	0.9618	0.6300	86.50%	0.9803	0.7450	86.75%	0.9565	0.7700	86.00%	0.9500	0.7600
RBF	S_2	88.75%	0.9874	0.7850	81.00%	0.9844	0.6300	89.00%	1.0000	0.7800	87.25%	0.9934	0.7500	88.00%	0.9935	0.7650
	S_1	89.50%	0.9877	0.8000	81.00%	0.9844	0.6300	88.25%	1.0000	0.7650	87.25%	0.9869	0.7550	87.75%	0.9935	0.7600
	R	89.00%	0.9699	0.8050	82.25%	0.9924	0.6500	87.00%	0.9933	0.7450	87.25%	0.9806	0.7600	87.75%	0.9935	0.7600

References

[1] Jaume. Arnó Satorra, José Antonio Martínez Casasnovas, Manuel Ribes Dasi, and Joan Ramon Rosell Polo. Review. precision viticulture. research topics, challenges and opportunities in site-specific vineyard management. *Spanish Journal of Agricultural Research*, 7(4):779 – 790, 2009.

[2] Ron Berenstein, OhadBen Shahar, Amir Shapiro, and Yael Edan. Grape clusters and foliage detection algorithms for autonomous selective vineyard sprayer. *Intelligent Service Robotics*, 3(4):233 – 243, 2010.

[3] Stephen Nuske, Supreeth Achar, Terry Bates, Srinivasa Narasimhan, and Sanjiv Singh. Yield estimation in vineyards by visual grape detection. In *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*, pages 2352 – 2358. IEEE, 2011.

[4] Maria-Paz Diago, Christian Correa, Borja Millán, Pilar Barreiro, Constantino Valero, and Javier Tardaguila. Grapevine yield and leaf area estimation using supervised classification methodology on RGB images taken under field conditions. *Sensors*, 12(12):16988 – 17006, 2012.

[5] Scarlett Liu and Mark Whitty. Automatic grape bunch detection in vineyards with an SVM classifier. *Journal of Applied Logic*, 2015.

[6] M.J.C.S. Reis, R. Morais, E. Peres, C. Pereira, O. Contente, S. Soares, A. Valente, J. Baptista, P.J.S.G. Ferreira, and J. Bulas Cruz. Automatic detection of bunches of grapes in natural environment from color images. *Journal of Applied Logic*, 10(4):285 – 290, 2012.

[7] Pavel Škrabánek and Thomas Philip Runarsson. Detection of grapes in natural environment using support vector machine classifier. In *Proceedings of the 21st International Conference on Soft Computing MENDEL 2015*, pages 143 – 150, Brno, Czech Republic, 23–25 June 2015. Brno University of Technology.

[8] Pavel Škrabánek and Filip Majerík. Evaluation of performance of grape berry detectors on real-life images. In *Proceedings of the 22nd International Conference on Soft Computing MENDEL 2016*, pages 217–224, Brno, Czech Republic, 8–10 June 2016. Brno University of Technology.

[9] Pavel Škrabánek and Filip Majerík. *Artificial Intelligence Perspectives in Intelligent Systems: Proceedings of the 5th Computer Science On-line Conference 2016 (CSOC2016), Vol 1*, chapter Simplified version of white wine grape berries detector based on SVM and HOG features, pages 35–45. Springer International Publishing, 2016.

[10] ITU-R Recommendation BT.601. Studio encoding parameters of digital television for standard 4:3 and wide screen 16:9 aspect ratios, March 2011.

[11] Christopher Kanan and Garrison W. Cottrell. Color-to-Grayscale: Does the Method Matter in Image Recognition? *PLoS ONE*, 7(1):1–7, 01 2012.

[12] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886 – 893, June 2005.

[13] Christopher J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.*, 2(2):121–167, June 1998.

[14] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, 13:281 – 305, February 2012.

[15] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, 45(4):427 – 437, July 2009.

[16] Christoph H. Lampert. Kernel methods in computer vision. *Foundations and Trends® in Computer Graphics and Vision*, 4(3):193 – 285, 2008.

[17] Scott Krig. *Computer Vision Metrics: Survey, Taxonomy, and Analysis*. Apress, Berkely, CA, USA, 1st edition, 2014.

[18] Richard J. Howarth. Sources for a history of the ternary diagram. *The British Journal for the History of Science*, 29:337–356, 9 1996.

[19] Pavel Škrabánek. Editor for marking and labeling of object images for binary supervised classification in matlab environment. In *Proceedings of the 21st International Conference on Soft Computing MENDEL 2015*, pages 151 – 158, Brno, Czech Republic, 23–25 June 2015. Brno University of Technology.
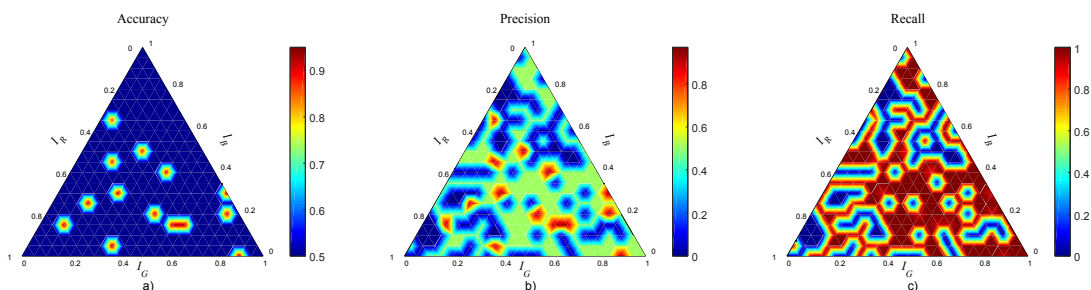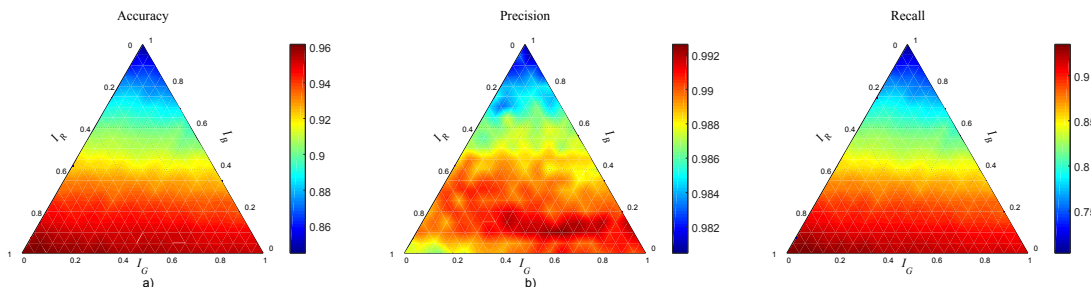
Figure 4. Performance of the robust detector with the RBF kernel for $C = 10$ and $\sigma = 1$ according to: a) accuracy, b) precision, c) recall.
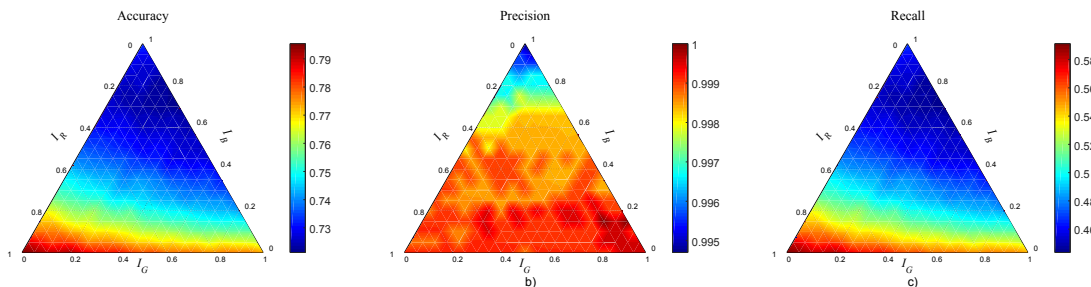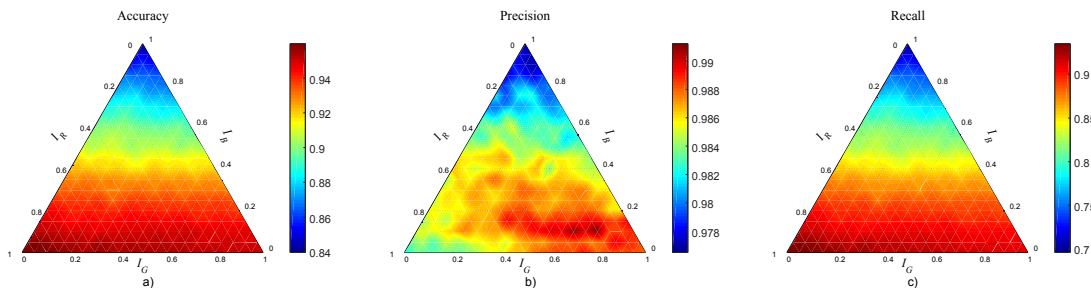

Figure 5. Performance of the robust detector with the RBF kernel for $C = 10$ and $\sigma = 10$ according to: a) accuracy, b) precision, c) recall.
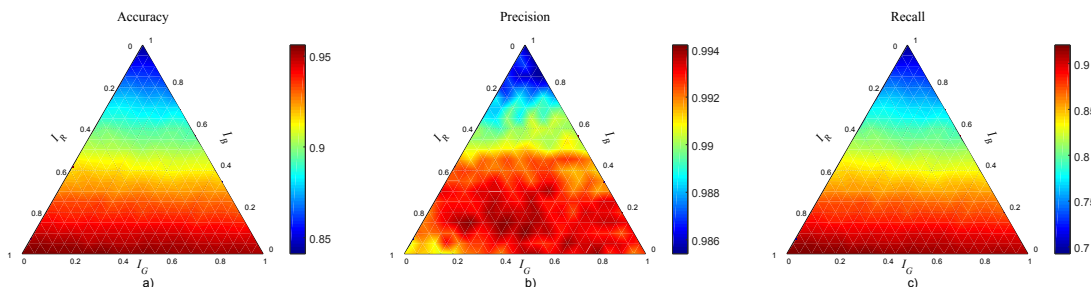
Figure 6. Performance of the robust detector with the RBF kernel for $C = 10$ and $\sigma = 20$ according to: a) accuracy, b) precision, c) recall.
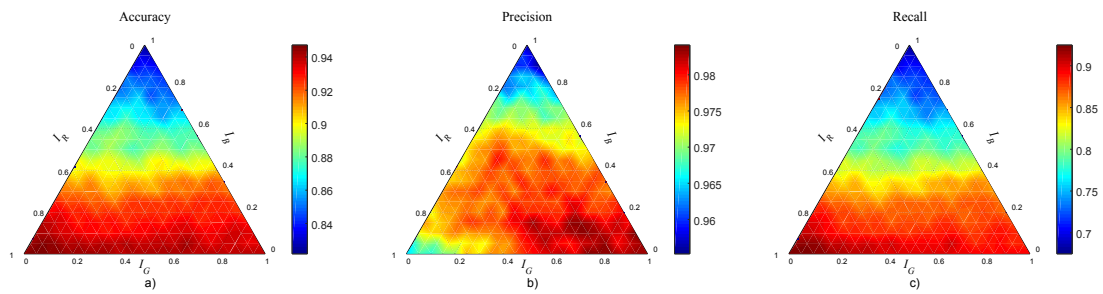

Figure 7. Performance of the robust detector with the RBF kernel for $C = 10$ and $\sigma = 30$ according to: a) accuracy, b) precision, c) recall.

Figure 8. Performance of the robust detector with the RBF kernel for $C = 10$ and $\sigma = 40$ according to: a) accuracy, b) precision, c) recall.

Figure 9. Performance of the robust detector with the linear kernel for $C = 1$ according to: a) accuracy, b) precision, c) recall.