Generalizing the majority voting scheme to spatially constrained voting

András Hajdu*, *Member, IEEE,* Lajos Hajdu, Ágnes Jónás, László Kovács, and Henrietta Tomán

***Abstract*—Generating ensembles from multiple individual classifiers is a popular approach to raise the accuracy of the decision. As a rule for decision making, majority voting is a usually applied model. In this paper, we generalize classical majority voting by incorporating probability terms $p_{n,k}$ to constrain the basic framework. These terms control whether a correct or false decision is made if k correct votes are present among the total number of n. This generalization is motivated by object detection problems, where the members of the ensemble are image processing algorithms giving their votes as pixels in the image domain. In this scenario, the terms $p_{n,k}$ can be specialized by a geometric constraint. Namely, the votes should fall inside a region matching the size and shape of the object to vote together. We give several theoretical results in this new model for both dependent and independent classifiers, whose individual accuracies may also differ. As a real world example, we present our ensemble-based system developed for the detection of the optic disc in retinal images. For this problem, experimental results are shown to demonstrate the characterization capability of this system. We also investigate how the generalized model can help us to improve an ensemble with extending it by adding a new algorithm.**

***Index Terms*—generalized majority voting, classifier combination, independence and dependence, pattern recognition, object detection.**

EDICS Category: ARS-IVA, ARS-RBS, TEC-BIP, SMR-SMD, SMR-REP

I. Introduction

ENSEMBLE-BASED systems are rather popular to raise the decision accuracy by combining the responses of different sources (voters, classifiers). Regarding pattern recognition, the idea of combining the decisions of multiple classifiers has also been studied [1]. As corresponding examples, we can mention neural networks [2], [3], decision trees [4], sets of rules [5] and other models [6], [7], [8]. As a specific application field, now we will focus on object detection in digital images which is a vivid field [9], [10], [11], as well.

A usual way for information fusion is to consider the majority of the votes of the classifiers as the basis of the decision. The current literature is quite rich regarding both theoretical results and applications of such systems (ensembles). Strong focus is set to the combination of votes of binary (correct/false) values. The related decision may take place based on simple majority [2], [12], [13], weighted majority [12], or using some other variants [14], [15].

In the research of majority voting, a cardinal issue is the assumptions on the dependency of the voters. Several results are achieved for independent voters, and the minimal and maximal accuracies of such majority voting systems are also studied for the dependent case. In this paper, we investigate how such voting systems behave if we apply some further constraints on the votes. Namely, we generalize the classical majority voting scheme by introducing real values $0 \leq p_{n,k} \leq 1$ for the probability that a good decision is made if we have k correct votes out of the n ones. In other words, in our case it will be possible that a good decision is made even if the good votes are in minority (less than half).

The creation of this new model is motivated by a retinal image processing problem – the detection of the optic disc (OD), which appears as a bright circular patch within the region of interest (ROI) in a retinal image (see Figure 1). Namely, in a former work we observed that organizing more individual OD detector algorithms into an ensemble may raise detection accuracy [16]. In the voting system applied here, each individual OD algorithm votes in terms of a single pixel as its candidate for the OD center. The application of existing majority voting models are not adequate here, since they consider only the correctness of the votes, which concerns falling into the true OD region in this scenario. However, in our case, the spatial behavior of the votes is also important, since they vote together for a specific location of the OD, only if they fall within a region matching the OD geometry. Consequently, we should consider discs of diameter of the OD $d_{OD} \in \mathbb{R}_{\geq 0}$ covering the candidates of the individual detector algorithms as shown in Figure 1. The diameter d_{OD} can be derived by averaging the manual annotations made by clinical experts on a dataset and can be adjusted to the resolution of the image. As a final decision, the disc having diameter d_{OD} with maximal number of candidates included is chosen for the OD location. In this combined system, we can make a good decision even if the false candidates have majority such as in the case illustrated in Figure 1. A bad decision is made only when a subset of false candidates with larger cardinality than the number of correct ones can be covered by a disc having diameter d_{OD}.

In this paper, we propose the generalization of the classical majority voting model by incorporating the probability terms $p_{n,k}$ mentioned before. With an appropriate geometric

A. Hajdu, A. Jónás, L. Kovács and H. Tomán are with the Department of Computer Graphics and Image Processing, Faculty of Informatics, University of Debrecen, POB 12, 4010 Debrecen, Hungary. *Corresponding author A. Hajdu: Tel.: +36-52-512900/62830, Fax: +36-52-512900/62822, email: hajdu.andras@inf.unideb.hu.

L. Hajdu is with the Institute of Mathematics, University of Debrecen, POB 12, 4010 Debrecen, Hungary.

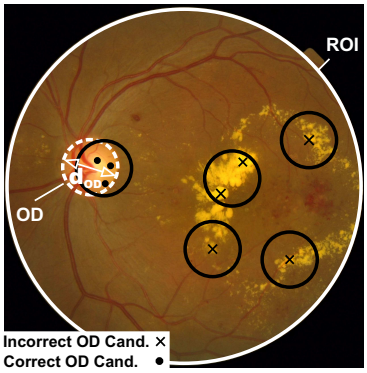


Fig. 1. The optic disc (OD) of diameter d_{OD} in a retinal image and the OD center candidates (3 correct, 5 false) of individual detector algorithms. Candidates inside the black circles can vote together for possible OD locations.

constraint, our generalized model can be specialized to be applicable for the above detection scenario, as well. Namely, the corresponding values $p_{n,k}$ will be adjusted by requiring that the candidates should fall inside a disc of a fixed diameter d_{OD} to vote together. With the help of this model, we can characterize our detector ensemble and gain information on further improvability issues, as well. As a different approach, it would be possible to require more than half of the votes to fall inside such a disc. However, this strict majority voting rule is rather unnatural in the spatial domain, which impression has been also confirmed during our empirical studies.

The rest of the paper is organized as follows. Section II recalls the basic concepts of classical majority voting, which will provide fundamentals for our more general framework. In section III, we show how to incorporate the probability terms $p_{n,k}$ to constrain the basic formulation. We present theoretical results for the case of independent voters. Since in applications independent detector algorithms can hardly be expected, we also generalize the method to the dependent case in section IV. As a main focus, we investigate the possible lowest and highest accuracy of constrained ensembles. Moreover, we both consider equal and different individual accuracies for the members of the ensemble. From a practical point of view, the further improvability of an ensemble is of great importance, so in section V we give the theoretical background on how an ensemble behaves if a new classifier is added to it. Section VI contains our empirical results regarding a real world application (optic disc detection), where we apply this new model to characterize our current OD detector ensemble and to analyse its further improvability by adding a new algorithm. In section VII, we discuss our results and draw some conclusions regarding other test datasets and detection problems, and the improvability of the proposed method.

II. Majority voting

Let $D_1, D_2, \ldots, D_n$ be a set of classifiers (voters), $D_i : \Lambda \to \Omega$ $(i = 1, \ldots, n)$, where Λ can be any domain, and Ω is a set of finite class labels. The majority voting rule assigns the class label supported by the majority of the classifiers $D_1, \ldots, D_n$ to $\alpha \in \Lambda$. Usually, ties (same number of different votes) are broken randomly.

In [13] Kuncheva et al. discuss exhaustively the following special case. Let n be odd, $|\Omega| = 2$ (each classifier has a binary (correct/false) output value) and all classifiers are independent and have the same classification accuracy p. A correct class label is given by majority voting if at least $\lceil n/2 \rceil$ classifiers give correct answers. The majority voting rule with independent classifier decisions gives an overall correct classification accuracy calculated by the following formula:

$$P = \sum_{k=\lceil n/2 \rceil}^{n} \binom{n}{k} p^k (1-p)^{n-k}. \tag{1}$$

Several interesting results can be found in [1] applying majority voting to pattern recognition tasks. This method is guaranteed to give a higher accuracy than the individual classifiers if the classifiers are independent and $p > 0.5$ holds for their individual accuracies.

III. Generalization to constrained voting

As it has been discussed in the introduction, we generalize the classical majority voting approach by considering some constraints that must be also met by the votes. To give a more general methodology beyond geometric considerations, we model this type of constrained voting by introducing values $0 \leq p_{n,k} \leq 1$ describing the probability of making a good decision, when we have exactly k good votes from the n voters. Then, in section VI we will adopt this general model to our practical problem with spatial constraints.

As we have summarized in the introduction, several theoretical results are achieved for independent voters in the current literature, so we start with generalizing them to this case. However, in the vast majority of applications, we cannot expect independency among algorithms trying to detect the same object. Thus, later we extend the model to the case of dependent voters with generalizing such formerly investigated concepts that have high practical impact, as well.

A. The independent case

In our model, we consider a classifier D_i with accuracy p_i as a random variable η_i of Bernoulli distribution, i.e.:

$$P(\eta_i = 1) = p_i, \quad P(\eta_i = 0) = 1 - p_i \quad (i = 1, \ldots, n).$$

Here $\eta_i = 1$ means correct classification by D_i. In particular, the accuracy of D_i is just the expected value of η_i, that is, $E\eta_i = p_i$ $(i = 1, \ldots, n)$.

Let $p_{n,k}$ $(k = 0, 1, \ldots, n)$ be given real numbers with $0 \leq p_{n,0} \leq p_{n,1} \leq \cdots \leq p_{n,n} \leq 1$, and let the random variable ξ be such that:

$$P(\xi = 1) = p_{n,k} \quad \text{and} \quad P(\xi = 0) = 1 - p_{n,k},$$

where $k = |\{i \ : \ \eta_i = 1\}|$. That is, ξ represents the modified majority voting of the classifiers $D_1, \ldots, D_n$: if k out of the n classifiers give a correct vote, then we make a good decision (i.e. we have $\xi = 1$) with probability $p_{n,k}$.

Note that, in the special case, where:

$$p_{n,k} = \begin{cases} 1, & \text{if } k > n/2, \\ 1/2, & \text{if } k = n/2, \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

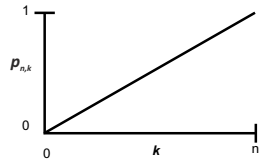
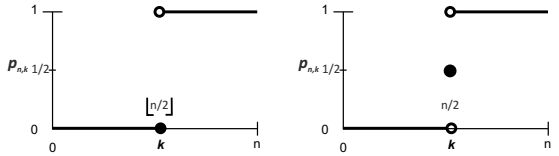


Fig. 3. The graph of $p_{n,k} = k/n$ providing $p = q$.

we get back the classical majority voting scheme.

The values $p_{n,k}$ as a function of k corresponding to the classical majority voting can be observed in Figure 2 for both an odd and even n, respectively.



(a) n is odd (b) n is even

Fig. 2. The graph of $p_{n,k}$ for classical majority voting for (a) an odd, and (b) an even number of voters n.

The ensemble accuracy of the classical majority voting system is shown in Table I for different number of classifiers n for some equal individual accuracies p (see also [12]).

	n=3	n=5	n=7	n=9
p = 0.6	0.6480	0.6826	0.7102	0.7334
p = 0.7	0.7840	0.8369	0.8740	0.9012
p = 0.8	0.8960	0.9421	0.9667	0.9804
p = 0.9	0.9720	0.9914	0.9973	0.9991

TABLE I
ENSEMBLE ACCURACY FOR CLASSICAL MAJORITY VOTING.

As the very first step of our generalization, we show that similarly to the individual voters, ξ is of Bernoulli distribution, as well. We also provide its corresponding parameter q, that represents the accuracy of the ensemble in our model.

Lemma 3.1: The random variable ξ is of Bernoulli distribution with parameter q, where:

$$q = \sum_{k=0}^{n} p_{n,k} \Bigg(\sum_{\substack{I \subseteq \{1,\dots,n\} \\ |I|=k}} \prod_{i \in I} p_i \prod_{j \in \{1,\dots,n\} \setminus I} (1-p_j) \Bigg). \quad (3)$$

Proof: Since for any $k \in \{0, 1, \dots, n\}$ we have:

$$P(|\{i \,:\, \eta_i = 1\}| = k) = \sum_{\substack{I \subseteq \{1,\dots,n\} \\ |I|=k}} \prod_{i \in I} p_i \prod_{j \in \{1,\dots,n\} \setminus I} (1-p_j),$$

the statement immediately follows from the definition of ξ. ■

The special case assuming equal accuracy for the classifiers received strong attention in the literature, so we investigate this case first. That is, in the rest of section III, we suppose that $p = p_1 = \ldots = p_n$. Then, (3) reads as:

$$q = \sum_{k=0}^{n} p_{n,k} \binom{n}{k} p^k (1-p)^{n-k}. \quad (4)$$

Thus, if n is odd then by the particular choice (2) for the values $p_{n,k}$, we get $q = P$, where P is given in (1). In order to have our generalized majority voting model be more accurate than the individual decisions, we have to guarantee that $q \geq p$. The next statement yields a guideline along this way.

Proposition 3.1: Let $p_{n,k} = k/n$ $(k = 0, 1, \dots, n)$. Then, we have $q = p$, and consequently $E\xi = p$.

Proof: See Appendix. ■

Figure 3 also illustrates the special linear case for $p_{n,k} = k/n$.

The above statement shows that if the probabilities $p_{n,k}$ increase uniformly (linearly), then the ensemble has the same accuracy as the individual classifiers. As a trivial consequence we obtain the following corollary.

Corollary 3.1: Suppose that for all $k = 0, 1, \dots, n$ we have $p_{n,k} \geq k/n$. Then $q \geq p$, and consequently $E\xi \geq p$.

The next result helps us to compare our model constrained by $p_{n,k}$ with the classical majority voting scheme.

Theorem 3.1: Suppose that $p \geq 1/2$ and for any k with $0 \leq k \leq n/2$ we have:

(i) $p_{n,k} + p_{n,n-k} \geq 1$,
(ii) $p_{n,n-k} \geq (n-k)/n$.

Let q be given by (4). Then, $q \geq p$, and consequently $E\xi \geq p$.

Proof: See Appendix. ■

As a specific case, we obtain the following corollary concerning the classical majority voting scheme [13].

Corollary 3.2: Suppose that n is odd, $p \geq 1/2$ and for all $k = 0, 1, \dots, n$ we have:

$$p_{n,k} = \begin{cases} 1, & \text{if } k > n/2, \\ 0, & \text{otherwise.} \end{cases}$$

Then, $q \geq p$, and consequently $E\xi \geq p$.

Proof: Observing that by the above choice for the values $p_{n,k}$ both properties (i) and (ii) of Theorem 3.1 are satisfied, the statement immediately follows from Theorem 3.1. ■

Of particular interest is the case, when the ensemble makes exclusively good decisions after t executions. That is, we are curious to know the conditions to have a system with accuracy 100%. So write $\xi^{\otimes t}$ for the random variable obtained by repeating ξ independently t times, and counting the number of one values (correct decisions) received, where t is a positive integer. Then, as it is well-known, $\xi^{\otimes t}$ is a random variable of binomial distribution with parameters (t, q) with q given by (4). Now we are interested in the probability $P(\xi^{\otimes t} = t)$. In case of using an individual classifier D_i (that is, a random variable η_i) with any $i = 1, \dots, n$, we certainly have $P(\eta_i^{\otimes t} = t) = p^t$. To make the ensemble better than the individual classifiers, we need to choose the probabilities $p_{n,k}$ so that $P(\xi^{\otimes t} = t) \geq p^t$. In fact, we can characterize a much more general case. For this purpose we need the following lemma, due to Gilat [17].

Lemma 3.2: For any integers t and l with $1 \leq l \leq t$ the function:

$$f(x) = \sum_{k=l}^{t} \binom{t}{k} x^k (1-x)^{t-k}$$

is strictly monotone increasing on $[0, 1]$.

Note that, for any $x \in [0,1]$ we obviously have:

$$\sum_{k=0}^{t} \binom{t}{k} x^k (1-x)^{t-k} = 1.$$

As a simple consequence of Lemma 3.2, we obtain the following result.

Theorem 3.2: Let t and l be integers with $1 \leq l \leq t$. Then, $P(\xi^{\otimes t} \geq l) \geq P(\eta_1^{\otimes t} \geq l)$, if and only if, $q \geq p$, i.e. $E\xi^{\otimes t} \geq tp$.

Proof: See Appendix. ■

IV. The dependent case

In this section, we investigate how dependencies among the voters influence the accuracy of the ensemble; for related results, see e.g. [12], [19]. For this purpose, we generalize some concepts that were introduced for classical majority voting to measure the extremal behavior (minimal and maximal accuracies) of an ensemble. First we consider *pattern of success* and *pattern of failure* [12] which are such realizations of the votes in a series of experiments that lead to the possible highest and lowest accuracy of the ensemble, respectively. It is worth noting that to define these measures, a rather serious restriction considering discretization of the model is needed to be applied. Namely, not only the accuracies of the individual classifiers are given, but also the precise numbers of successful decisions during the experiment are fixed. E.g. for a classifier having accuracy $p = 0.6$ we consider 6 correct votes in 10 experimental runs.

Though there are some results in the literature for the case of different accuracies p_i of the classifiers D_i (or, in other words, for the case $E\eta_i = p_i$ $(i = 1, \ldots, n)$), see e.g. [2], [20], [21] and the references there, the vast majority of the results (such as e.g. in [13]) concern the case $p = p_1 = \ldots = p_n$. So in section IV-A, we shall make the latter assumption, too. However, in section IV-B, we give a much more general framework which handles both dependencies without the restriction considering discretization, and also different accuracies of classifiers that makes the model realistic for applications.

A. Pattern of success and pattern of failure

In this section, we suppose that the individual classifier accuracies coincide ($p = p_1 = \ldots = p_n$). Repeat the experiments $\eta_1, \ldots, \eta_n$ t times, with some positive integer t, and write $\eta_i^{(j)}$ for the j-th realization of η_i $(j = 1, \ldots, t)$. Suppose (as a rather strong, but standard assumption) that we have:

$$|\{j \ : \ \eta_i^{(j)} = 1\}| = r \quad \text{for all } i = 1, \ldots, n. \tag{5}$$

Here r is a positive integer with $r = np$. We are interested in the behavior (accuracy) of ξ repeated t times, or in other words in the value $E\xi^{\otimes t}$, under the condition (5). Write $\xi^{(j)}$ for the j-th realization of ξ $(j = 1, \ldots, t)$. Then, we clearly have $E\xi^{\otimes t} = E\xi^{(1)} + \ldots + E\xi^{(t)}$.

The number of one values is fixed for η_i, however, their positions can freely change. For simplicity, we shall describe the situation by a table T of size $n \times t$: in the (i,j)-th entry $T(i,j)$ of T we write 0 or 1, according to the actual value of $\eta_i^{(j)}$ $(1 \leq i \leq n, 1 \leq j \leq t)$. Our first result in this interpretation concerns the case of linear $p_{n,k}$.

Proposition 4.1: If $p_{n,k} = k/n$ for all $k = 0, 1, \ldots, n$, then $E\xi^{\otimes t} = r$.

Proof: Denote by u_j the number of ones in the j-th column of the table T for $j = 1, \ldots, t$. Then, we have $E\xi^{(j)} = u_j/n$. Thus:

$$E\xi^{\otimes t} = E\xi^{(1)} + \ldots + E\xi^{(t)} = u_1/n + \ldots + u_t/n. \tag{6}$$

Since $u_1 + \ldots + u_t$ is just the total number of ones in T, we have:

$$u_1 + \ldots + u_t = nr. \tag{7}$$

Combining (6) and (7) we obtain $E\xi^{\otimes t} = r$, and the statement follows. ■

In view of the proof of Proposition 4.1, we see that in case of a general system $p_{n,k}$ we have:

$$E\xi^{\otimes t} = \sum_{j=1}^{t} p_{n,u_j}. \tag{8}$$

So to describe the pattern of success (highest accuracy) and the pattern of failure (lowest accuracy), we need to maximize and minimize the above quantity, respectively.

Our next result concerns the pattern of success. Here we consider the problem only under some further assumptions, which in fact are not necessary to study and describe the situation as it will be shown in section IV-B. However, on the one hand, the statement together with its proof already show the basic idea for construction. On the other hand, former results usually consider these assumptions, so in this way our model can be fitted to the existing literature, as well. In section IV-B, we describe the general method, which works without any technical restrictions.

Theorem 4.1: Let the probabilities $p_{n,k}$ be arbitrary, up to $p_{n,0} = 0$. Let $k_1 \neq 0$ be an index such that $p_{n,k_1}/k_1 \geq p_{n,k}/k$ for all $k = 1, \ldots, n$. Then, $E\xi^{\otimes t} \leq nrp_{n,k_1}/k_1$. Further, if $tk_1 = nr$ then the maximum can be attained.

Proof: See Appendix. ■

Our next theorem describes the pattern of failure, in a similar fashion as the previous statement.

Theorem 4.2: Let the probabilities $p_{n,k}$ be arbitrary, up to $p_{n,0} = 0$. Let $k_2 \neq 0$ be an index such that $p_{n,k_2}/k_2 \leq p_{n,k}/k$ for all $k = 1, \ldots, n$. Then, $E\xi^{\otimes t} \geq nrp_{n,k_2}/k_2$. Further, if $tk_2 = nr$ then the minimum can be attained.

Proof: Since the proof follows the same lines as that of Theorem 4.1 (see Appendix), we omit the details. ■

Similarly to the independent case in section III-A, we also investigate the case, when only good decision is made by the ensemble. In other words, we would like to describe the situation, where:

$$P(\xi^{\otimes t} = t) = \prod_{j=1}^{t} p_{n,u_j}$$

is maximal. Note that, in this case one can easily obtain a table T with $P(\xi^{\otimes t} = t) = 0$. So now finding the minimum (i.e. investigating the pattern of failure) does not make sense.

For the special case of $p_{n,k} = k/n$, we have the following result.

Theorem 4.3: Let $p_{n,k} = k/n$ for all $k = 0, 1, \ldots, n$, and assume that $nr \geq t$. Then $P(\xi^{\otimes t} = t)$ is maximal for the tables T in which:

$$\lfloor nr/t \rfloor \leq u_j \leq \lceil nr/t \rceil \quad (1 \leq j \leq t),$$

where u_j denotes the number of ones in the j-th column of T. Further, all these tables T can be explicitly constructed.

Proof: See Appendix. ∎

Note that, if $t > nr$ then T necessarily has a column with all zero entries, whence $P(\xi^{\otimes t} = t) = 0$ in this case. For general values $p_{n,k}$, we have the following result.

Theorem 4.4: Let the probabilities $p_{n,k}$ be arbitrary, up to $p_{n,0} = 0$ and $p_{n,k} > 0$ for $0 < k \leq n$. Let $k_0 \neq 0$ be an index such that $(\ln p_{n,k_0})/k_0 \geq (\ln p_{n,k})/k$ for all $k = 1, \ldots, n$. Then, $P(\xi^{\otimes t} = t) \leq p_{n,k_0}^{(nr/k_0)}$. Further, if $tk_0 = nr$ then the maximum can be attained.

Proof: See Appendix. ∎

B. Extremal accuracies by linear programming

In this section, we drop the condition (5), and give a compact tool based on linear programming to calculate the minimal and maximal ensemble accuracies. We assumed earlier that the random variables η_i ($i = 1, \ldots, n$) are independent. In our application, we consider different algorithms detecting the optic disc as voters. These algorithms cannot be assumed to be independent in all cases, because it can happen that the operations of the algorithms are based on very similar principles. In case of dependent algorithms, we have to decide how to measure the dependencies of the algorithms. For this aim, we can investigate the joint distribution of the outputs of the algorithms. So let:

$$c_{a_1,\ldots,a_n} = P(\eta_1 = a_1, \ldots, \eta_n = a_n), \tag{9}$$

where $a_i \in \{0, 1, *\}$. The star denotes any of the possible correctness values, that is, $* = 0$ or 1. The probabilities $c_{a_1,\ldots,a_n}$ can be considered as the entries of the contingency table of $\eta_1, \ldots, \eta_n$. The problem to determine the combination of voters achieving the best/worst ensemble performance is equivalent to maximize/minimize the function:

$$q(c_{a_1,\ldots,a_n}) = \sum_{k=0}^{n} \left(p_{n,k} \sum_{a_1+\ldots+a_n=k} c_{a_1,\ldots,a_n} \right) \tag{10}$$

under the following conditions:

$$\begin{aligned} &\sum_{a_i=1} c_{*,\ldots,*,a_i,*,\ldots,*} = p_i \quad (i = 1, \ldots, n), \\ &\sum_{a_1,\ldots,a_n} c_{a_1,\ldots,a_n} = 1, \\ &c_{a_1,\ldots,a_n} \geq 0, a_i \in \{0, 1\} \quad (i = 1, \ldots, n), \end{aligned} \tag{11}$$

where $E\eta_i = p_i$ ($i = 1, \ldots, n$) is the accuracy of the i-th detecting algorithm. Observe that this is just a linear programming problem for the variables $c_{a_1,\ldots,a_n}$, which can be solved by standard tools.

In the special case, when $(\eta_1, \ldots, \eta_n)$ are totally independent, we have:

$$c_{a_1,\ldots,a_n} = P(\eta_1 = a_1) \ldots P(\eta_n = a_n). \tag{12}$$

That is, the entries of the contingency table can be determined by the probabilities $p_1, \ldots, p_n$. In this case, the ensemble performance q is simply given by (3).

V. Extending the ensemble with adding a new classifier

From a practical point of view, it is very important to study the improvability of an existing ensemble regarding its accuracy. To address this issue, we investigate to what extent the addition of a new classifier D_{n+1} with accuracy p_{n+1} may improve the system. For this study, we observe both the change of the system accuracy q and the interval $[q_{min}, q_{max}]$ for the minimal and maximal system accuracy. More precisely, we will consider the following cases:

A. we fix the individual accuracies and output of the algorithms of the current ensemble for an experiment in terms of a contingency table, and:
   1. add a new independent algorithm and check how the ensemble accuracy (q) changes,
   2. add a new dependent algorithm and check how the minimal (q_{min}) and maximal (q_{max}) ensemble accuracy change, respectively,

B. we fix the individual accuracies, but ignore the output of the algorithms of the current ensemble for an experiment, add a new algorithm and check the minimal (q_{min}) and maximal (q_{max}) ensemble accuracy.

After adding a new algorithm to the existing system, the new system accuracy depends not only on the accuracies $p_1, \ldots, p_{n+1}$, but also on the values $p_{n+1,k}$. As an estimation for $p_{n+1,k}$, from the definition of $p_{n,k}$ we have:

$$p_{n,k} \geq p_{n+1,k}, \tag{13}$$

$$p_{n,k} \leq p_{n+1,k+1}. \tag{14}$$

In (13), the added vote is supposed to be false, so the probability of good decision after the extension cannot be greater than in the existing system. The estimation (14) describes the case of adding a correct vote to the system. To sum up (13) and (14), we get the following properties for $p_{n+1,k}$:

$$p_{n,k-1} \leq p_{n+1,k} \leq p_{n,k}. \tag{15}$$

Applying inequalities (15), the values $p_{n+1,k}$ can be estimated from the values $p_{n,k}$.

If a new member is added to an existing ensemble, the accuracy of the extended ensemble is affected by two main properties of the new voter: its accuracy and its correlation with the members of the existing system. Let η_{n+1} be a random variable with $E\eta_{n+1} = p_{n+1}$. To determine the best/worst choice for the new member to achieve the best (q_{max})/worst (q_{min}) performance for the extended ensemble the following linear optimization problem has to be solved in the general case B. Maximize/Minimize the function:

$$q(c_{a_1,\ldots,a_{n+1}}) = \sum_{k=0}^{n+1} \left(p_{n+1,k} \sum_{\substack{a_1+\ldots+a_n+ \\ +a_{n+1}=k}} c_{a_1,\ldots,a_{n+1}} \right) \tag{16}$$

under the following conditions:

$$\begin{aligned}&\sum_{a_i=1} c_{*,\dots,*,a_i,*,\dots,*} = p_i \quad (i=1,\dots,n+1),\\ &\sum_{a_1,\dots,a_{n+1}} c_{a_1,\dots,a_{n+1}} = 1,\\ &c_{a_1,\dots,a_{n+1}} \geq 0,\ a_i \in \{0,1\} \quad (i=1,\dots,n+1),\end{aligned} \tag{17}$$

where $E\eta_i = p_i$ $(i = 1,\dots,n+1)$, so the accuracy of the i-th classifier is p_i.

In case A.2., besides the objective function in (16) and the conditions in (17) are the same, we have an extra condition:

$$c_{a_1,\dots,a_n} = c_{a_1,\dots,a_n,0} + c_{a_1,\dots,a_n,1}. \tag{18}$$

From the definition of $c_{a_1,\dots,a_n}$ given in (9) it follows that the term containing $c_{a_1,\dots,a_{n+1}}$ in (16) can be split as:

$$\begin{aligned}&\sum_{a_1+\dots+a_{n+1}=k} c_{a_1,\dots,a_{n+1}} =\\ &\sum_{a_1+\dots+a_n=k} c_{a_1,\dots,a_n,0} + \sum_{a_1+\dots+a_n=k-1} c_{a_1,\dots,a_n,1}.\end{aligned} \tag{19}$$

Without having any further information about $p_{n+1,k}$, we can give an interval for q_{min} and q_{max}. Let $q^{\ominus}_{min}/q^{\ominus}_{max}$ and $q^{\oplus}_{min}/q^{\oplus}_{max}$ be the minimal/maximal value of the objective function (16) if we consider the estimations $p_{n,k-1} = p_{n+1,k}$ and $p_{n+1,k} = p_{n,k}$, respectively. From (15), we get:

$$q^{\ominus}_{min} \leq q_{min} \leq q^{\oplus}_{min}, \quad \text{and} \quad q^{\ominus}_{max} \leq q_{max} \leq q^{\oplus}_{max}. \tag{20}$$

In the special case, when η_{n+1} is totally independent from $(\eta_1,\dots,\eta_n)$, the entries of the extended contingency table can be determined by $c_{a_1,\dots,a_n}$ and p_{n+1}:

$$\begin{aligned}c_{a_1,\dots,a_n,1} &= p_{n+1} c_{a_1,\dots,a_n},\\ c_{a_1,\dots,a_n,0} &= (1-p_{n+1}) c_{a_1,\dots,a_n}.\end{aligned} \tag{21}$$

Considering the equations (16), (19) and (21) we get that the linear optimization problem can be solved by maximizing/minimizing the function:

$$\begin{aligned}&q(c_{a_1,\dots,a_{n+1}}) =\\ &\sum_{k=0}^{n+1} p_{n+1,k}\Bigg(\sum_{a_1+\dots+a_n=k}(1-p_{n+1})c_{a_1,\dots,a_n} +\\ &\qquad + \sum_{a_1+\dots+a_n=k-1} p_{n+1} c_{a_1,\dots,a_n}\Bigg)\end{aligned} \tag{22}$$

under the conditions given in (11).

If we consider that the entries of the contingency table of $\eta_1,\dots,\eta_n$ remain the same after adding an independent variable η_{n+1} to the ensemble (case A.1.), the solution of the problem in (22) under the conditions (11) depends only on p_{n+1} and $p_{n+1,k}$.

In the same way as in (20), from (15) we get:

$$q^{\ominus} \leq q \leq q^{\oplus}, \tag{23}$$

where $q^{\ominus}$ and $q^{\oplus}$ denote the minimal/maximal value of the objective function (22) for a fixed p_{n+1} if we consider the estimations $p_{n,k-1} = p_{n+1,k}$ and $p_{n+1,k} = p_{n,k}$, respectively.

For the improvability of the system, we have the following proposition.

Proposition 5.1: For the accuracy of the extended ensemble we have:

$$q(c_{a_1,\dots,a_{n+1}}) \geq q(c_{a_1,\dots,a_n}),$$

if:

$$p_{n+1} \geq \frac{\sum_{k=0}^{n}\left(\sum_{a_1+\dots+a_n=k} c_{a_1,\dots,a_n}(p_{n,k}-p_{n+1,k})\right)}{\sum_{k=0}^{n}\left(\sum_{a_1+\dots+a_n=k} c_{a_1,\dots,a_n}(p_{n+1,k+1}-p_{n+1,k})\right)}$$

holds for the accuracy of the added member.

Proof: First, note that the value of this fraction is non-negative, since $p_{n,k} \geq p_{n+1,k}$ and $p_{n+1,k} \leq p_{n+1,k+1}$. Moreover, from (15) and (21) the statement follows. ■

In section VI-D, we will show some experimental results for the improvability of the accuracy of our OD detector ensemble with adding a new algorithm.

VI. Application – optic disc detection

Now we turn to show, how our generalized model supports real-world problems in a clinical field. Progressive eye diseases can be caused by diabetic retinopathy (DR) which can lead even to blindness. One of the first essential steps in automatic grading of the retinal images is to determine the exact location of the main anatomical features, such as the optic disc. The locations of these features play important role in making diagnosis in the clinical protocol. In this section, for the OD detection task, we start with showing how the general formulation considering the probabilities $p_{n,k}$ is restricted for this specific challenge using geometric constraints defined by anatomic rules. Then, we present the accuracy of our current ensemble, characterize it by the achieved results and discuss the possibilities of its further improvement.

A. Constraining by shape characteristics

In our application, the votes are required to fall inside a disc of diameter d_{OD} to vote together. For the calculation of the values $p_{n,k}$ for our proposed method, the k correct votes must fall inside the true OD region, however, the $n-k$ false ones can fall within discs with diameter d_{OD} anywhere else within the ROI (region of interest in the image). That is, more false regions are possible to be formed which gives the possibility to make a correct decision even if the true votes are in minority. Note that, a candidate of an algorithm is considered to be correct if its distance from the manually selected OD center is not larger than $d_{OD}/2$. For this configuration, see Figure 5.

If we assume independency among the algorithms, for our application the behavior of the values $p_{n,k}$ as a function of k for a given n is shown in Figure 4 for $n = 9$ and $p = 0.9$.

This function has been determined empirically by dropping random pixels on the disc in a large number of experiments. Figure 4 shows that $p_{n,k}$ increases exponentially in k for a given n. This fact is also suggested by the results in [22], [23] saying that the probability that the diameter of a point set is not less than a given constant decreases exponentially if

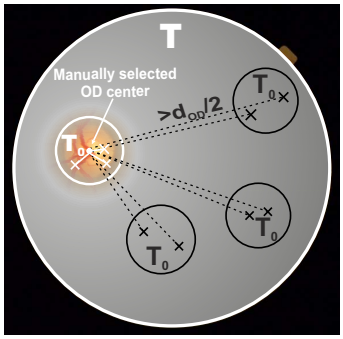


Fig. 5. The geometric constraint applied to the candidates of the algorithms: they should fall inside a disc of a fixed diameter d_{OD} to vote together.

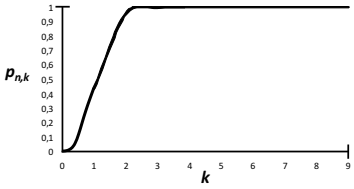


Fig. 4. The graph of $p_{n,k}$ for a fixed $n = 9$ and $p = 0.9$ with our geometric constraint to fall within a disc of diameter d_{OD}.

the number of points tends to infinity. Note that, this diameter corresponds again to the diameter d_{OD} of the OD.

The ensemble accuracy for our spatially constrained system is measured empirically by the help of a set of test images. The obtained data are enclosed in Table II for different number of independent classifiers (n) for some equal individual accuracies (p).

	n=3	n=5	n=7	n=9
p = 0.6	0.6435	0.9076	0.9654	0.9893
p = 0.7	0.7889	0.9631	0.9938	0.9985
p = 0.8	0.9029	0.9906	0.9986	0.9997
p = 0.9	0.9697	0.9994	1.0000	1.0000

TABLE II
MEASURED ENSEMBLE ACCURACY UNDER THE GEOMETRIC CONSTRAINT.

From Table II we can see a rapid increase in the ensemble accuracy. From trivial geometric considerations, it can be also seen why an ensemble with few members (e.g. $n = 3$) performs bad.

Now, to describe the spatially constrained case in detail, let us assign the probability $(1 - p_i)s_i$ with $s_i \in [0, 1]$ to the i-th independent classifier. This probability means that the i-th voter makes false individual decision (term $1 - p_i$) and participates in making a false ensemble decision (term s_i). For the algorithm D_i with accuracy p_i giving a false candidate having coordinates (x_i, y_i) for the OD center, we consider that the distribution of (x_i, y_i) is uniform outside the true OD region for all $i = 1, \ldots, n$. With this setup, we have:

$$s_1 = \ldots = s_n = \frac{T_0}{T - T_0}, \tag{24}$$

where T_0 and T are the area of the OD and the ROI, respectively, so in this case s_i is the same predetermined constant for all $i = 1, \ldots, n$. For better understanding, see also Figure 5.

For the interpretation of the values $p_{n,k}$ for this case, let us consider the decomposition of the number of false candidates $n-k = k_1 + \ldots + k_l$, where all the false votes are covered by the l disjoint discs of diameter d_{OD}, and k_i is the cardinality of the false votes covered by the i-th disc. Without the loss of generality, we may assume that $k_1 \geq \ldots \geq k_l$. To determine the values $p_{n,k}$, we introduce the probability $P(n, k, k_1, \ldots, k_l)$ for the good decision in case of a concrete realization of the n votes:

$$P(n, k, k_1, \ldots, k_l) = \frac{n!}{k!k_1! \ldots k_l!} p_1 \ldots p_k (1 - p_{k+1}) \ldots (1 - p_n) \cdot \cdot \left(1 - \frac{T_0}{T}\right)^{k_1} \ldots \left(1 - \frac{lT_0}{T}\right)^{k_l}.$$

Applying the geometric constraint, false decision is made only when $k_1 > k$ so $p_{n,k} = 0$ for $k_1 > k$, while $p_{n,k} = 1$ for $k > k_1$ should hold. The case $k_1 = k$ is broken randomly. Based on these considerations and summing for the possible distribution of the $n - k$ false votes among the discs, we can calculate the corresponding values $p_{n,k}$ as follows:

$$p_{n,k} = \sum_{k_1 + \ldots + k_l = n-k, k > k_1} P(n, k, k_1, \ldots, k_l) + \frac{1}{2} \sum_{k_1 + \ldots + k_l = n-k, k = k_1} P(n, k, k_1, \ldots, k_l) \tag{25}$$

The values $p_{n,k}$ calculated by (25) and the ones shown in Figure 4 slightly differ. The reason for this difference is that in our geometric derivation to have the closed form (25), we have considered only disjoint discs that completely fall inside the ROI, as well. However, these differences are minor, and both approaches have exponential trends.

From the basic results and concepts introduced in section II, strict majority voting scheme could be also applied as a decision rule, which means that at least $\lfloor n/2 \rfloor + 1$ votes should fall within a disc of diameter d_{OD} to make a good decision. However, this strict approach is much more unnatural than the proposed one confirmed by the experimental results presented in the next sections, as well.

B. An ensemble-based OD detector

To take advantage of the theoretical foundations of the previous sections for efficient OD detection, we have collected eight corresponding individual algorithms to create an ensemble from. Then, with a brute force approach (i.e. checking all the possible combinations) we select such an ensemble which maximizes the accuracy of the combined system. For measuring the accuracy of both the individual algorithms and the ensembles, we used the dataset MESSIDOR [24] containing 1200 digital images, where the OD centers were manually labelled by clinical experts. The images are losslessly compressed with 45° FOV and of different resolutions (1440×960, 2240×1488, and 2304×1536 pixels) that were

re-scaled to 1500×1152 for normalization. For this specific resolution, we get $d_{OD} = 184$ pixels from averaging the manual annotations of clinical experts for this dataset. As a result of brute force selection, we composed an ensemble from six OD-detectors. To have an impression about the similarities and differences between these approaches, next we give a short description for each of them. Each individual accuracy (p_i) has been measured on the dataset MESSIDOR.

- *Based on pyramidal decomposition:* Lalonde et al. [25] created an algorithm which generates a pyramid with simple Haar-based discrete wavelet transform. The pixel with the highest intensity value in the low-resolution image (4th or 5th level of decomposition) is considered as the center of the OD. $p_1 = 0.767$
- *Based on edge detection:* This method [25] uses edge detection algorithm which is based on Rayleight-based CFAR threshold. Next, Hausdorff distance is calculated between the set of edge points and a circular template like the average OD. The pixel with the lowest distance value is selected for OD center. $p_2 = 0.958$
- *Based on entropy measurement:* Sopharak et al. [26] proposed this method which applies a median and a CLAHE filter on the retinal image. In a neighborhood of each pixel, the entropy of intensity is calculated; the pixel with the largest entropy value is selected as the OD center. $p_3 = 0.315$
- *Based on kNN classification:* Niemeijer et al. [27] extracted features (number, width, orientation and density of vessels and their combination), and applied a kNN classifier to decide whether a pixel belongs to the OD region. The centroid of the largest component found is considered as the OD center. $p_4 = 0.759$
- *Based on fuzzy convergence of blood vessels:* This method [28] thins the vessel system and models each line-shape segment with a fuzzy segment. A voting map of these fuzzy segments is created and the pixel receiving the most votes is considered as OD center. $p_5 = 0.977$
- *Based on Hough transformation of vessels:* Ravishankar et al. [29] proposed to fit lines to the thinned vessel system by Hough transformation. The intersection of these lines results in a probability map. A weighting is also applied considering the intensity values corresponding to the intersection points. The pixel having the highest probability is considered as OD center. $p_6 = 0.647$

As for the decision of the ensemble, we select the disc of the fixed diameter d_{OD} containing the largest number of algorithm candidates. Then, as the final OD center, we consider the centroid of these candidates. The final OD center is correctly found, if it falls inside the disc aligned to the manually selected OD center and having diameter d_{OD}.

C. Characterizing and comparing OD-ensemble accuracies

A natural question regarding the ensemble of the detectors is what accuracies we can expect as the best or worst based on the given individual detector accuracies. Then, we can see where the accuracy of our current ensemble falls within this interval, and can also check how it relates to a system which would contain independent ensemble members.

In our application, the values $p_{n,k}$ for calculating the above characterizing ensemble accuracies as a function of k for $n = 6$ is calculated empirically and shown in Figure 6. Note that, though our system naturally contains dependencies among its members, the exponential behavior of the independent ensemble (see Figure 4) can be observed here, as well.

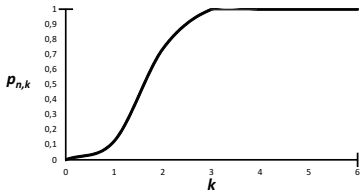


Fig. 6. The graph of $p_{n,k}$ for a fixed $n = 6$ in our OD detector ensemble.

Using the linear programming technique described in section IV-B, we have the following minimal and maximal ensemble accuracies, respectively:

$$q_{min} = 0.899, \qquad q_{max} = 1 \tag{26}$$

for the given individual accuracies.

Based on our experimental tests, the ensemble accuracy for our system has been found to be:

$$q = 0.981, \tag{27}$$

which is quite close to the possible maximal accuracy $q_{max} = 1$. However, if we calculate the system accuracy using (10) under the conditions (11) and with the assumption (12) on the independency of the detectors, we have:

$$q_{ind} = 0.998. \tag{28}$$

That is, an ensemble of independent algorithms with the given individual accuracies $p_1, \ldots, p_6$ would lead to nearly perfect results regarding accuracy. On the other hand, it is not surprising that our current system performs worse, since in this specific detection task it is quite challenging to find algorithms based on different (independent/diverse) principles.

Similarly to our proposed method, we have also determined the highest ensemble accuracy regarding the strict majority voting scheme. In this case, the brute force search provided the highest accuracy for the ensemble of the five members having individual accuracies p_1, p_2, p_4, p_5, and p_6, respectively. The ensemble accuracy measured by following the strict decision rule (at least three votes should fall within a disc of diameter d_{OD}) has been found to be:

$$q_{strict} = 0.944. \tag{29}$$

Comparing (29) with (27) confirms that the proposed spatially constrained voting model leads to remarkably higher accuracy than by simply extending the classic majority vote rule.

Accuracy of the new algorithm	$q^{\ominus}$	$q^{\oplus}$
p_7 = 0.6	0.957	0.989
p_7 = 0.9	0.975	0.995

TABLE IV
THE INTERVAL FOR THE OD DETECTOR ENSEMBLE ACCURACY IF A NEW INDEPENDENT ALGORITHM IS ADDED TO A DEPENDENT SYSTEM.

D. On adding algorithms to the detector

In section V, we have laid the theoretical background to extend the ensemble with adding a new algorithm. Namely, we have formulated the ways of the calculation of ensemble accuracy for the cases, when the new member is dependent or independent from the ensemble, respectively. Besides the simple ensemble accuracy, we have also explained how the minimal and maximal accuracies of the ensemble would change. Now, we adopt these results to our specific application and investigate how our current OD detector ensemble is going to behave if a new detector algorithm is added. Besides the simple ensemble accuracy, we have also explained how the minimal and maximal accuracies of the ensemble would change.

To start our experimental discussion on this topic, we check the behavior of our current OD detector ensemble during its compilation. Namely, we measure the change of the ensemble accuracy, when the sixth member is added to the ensemble of five members. For this aim, we calculate the accuracy of each ensemble of five individual algorithms with the corresponding figures enclosed in Table III. Thus, Table III contains the accuracies of the six possible ensembles of five members, where in the ith column the ith member is excluded having individual accuracy p_i, for $i = 1, \ldots, 6$.

Index of excluded member	1	2	3	4	5	6
Ensemble accuracy (5 members)	0.980	0.957	0.979	0.976	0.961	0.976
Ensemble accuracy (6 members)	0.981					

TABLE III
CHANGE OF THE ENSEMBLE ACCURACY, WHEN THE SIXTH MEMBER IS ADDED TO THE ENSEMBLE OF FIVE ALGORITHMS.

From Table III we can see that the largest increase in accuracy (from 0.957 to 0.981) is reached not by adding the most accurate ($p_5 = 0.977$) member, but a slightly less accurate ($p_2 = 0.958$) one. Similarly, the smallest improvement (from 0.980 to 0.981) is found not by adding the least accurate ($p_3 = 0.315$) member, but by adding an individually more accurate ($p_1 = 0.767$) one. To understand these results we should realize that there are specific dependencies among the members. Thus, in general, it is not sufficient to simply compose an ensemble based on the individual accuracies.

Next, we adopt the results from section V to investigate how our current OD detector ensemble consisting of six algorithms is going to behave if a new detector algorithm is added. We start with the case A.1 from section V, when the dependencies of the current ensemble members are considered as known in terms of a contingency table belonging to our experimental test on the dataset MESSIDOR and the new algorithm is considered to be independent from the ensemble. For this case, through the solution of (22), we gain the numeric results enclosed in Table IV. Note that, in this case we can check the interval $[q^{\ominus}, q^{\oplus}]$ introduced in (23) where the ensemble accuracy will fall based on the lower and upper estimation that can be derived for $p_{n+1,k}$ as given in (15).

From Table IV, we can see that in our application a new (independent) algorithm with accuracy approximately 0.9 is highly expected to improve the current system accuracy given in (27). The case A.1 in section V also includes the special scenario, when the existing ensemble contains independent members and we add an independent algorithm, as well. For this scenario, we can investigate the minimal and maximal accuracies of the new system by solving the problem in (22) under the extra condition (12). In Table V, we enclosed the respective accuracy figures regarding the lower and upper estimations of the values $p_{n+1,k}$.

Accuracy of the new algorithm	$q^{\ominus}$	$q^{\oplus}$
p_7 = 0.6	0.975	0.997
p_7 = 0.9	0.984	0.999

TABLE V
THE INTERVAL FOR THE OD DETECTOR ENSEMBLE ACCURACY IF A NEW INDEPENDENT ALGORITHM IS ADDED TO AN INDEPENDENT SYSTEM.

By comparing Table IV with Table V, we can see that if we assume total independency among the algorithms, we can expect higher ensemble accuracy. Since the original ensemble would lead to very high accuracy with independent algorithms as given in (28), only in case of a very accurate new algorithm we can expect improvement.

Next, we analyse the case A.2 from section V, when the dependencies of the algorithms are still considered, but the new algorithm should not be independent. In this setup, we can determine the accuracy interval introduced in (20) for the minimal (q_{min}) and maximal (q_{max}) ensemble accuracies, respectively, based on the estimation for the values $p_{n+1,k}$ as given in (15). The corresponding figures presented in Table VI can be determined by the solution of (16) under the conditions (17), (18).

Accuracy of the new algorithm	$q^{\ominus}_{min}$	$q^{\oplus}_{min}$	$q^{\ominus}_{max}$	$q^{\oplus}_{max}$
p_7 = 0.1	0.920	0.981	0.981	0.995
p_7 = 0.7	0.920	0.981	0.981	0.995
p_7 = 0.9	0.942	0.981	0.981	0.995

TABLE VI
THE INTERVAL FOR THE MINIMAL AND MAXIMAL OD DETECTOR ENSEMBLE ACCURACY IF A NEW DEPENDENT ALGORITHM IS ADDED TO A DEPENDENT SYSTEM.

Table VI shows that an individually very weak, but diverse algorithm could lead to a remarkable improvement of the ensemble, however, this possibility is rather unrealistic. Moreover, since the current ensemble is not optimal regarding dependencies, even with a very diverse and accurate algorithm we cannot reach accuracy 100%. It is also visible from Table VI that the original system accuracy (27) cannot be outperformed with the lower estimation for $p_{n+1,k}$, and cannot be degraded with its upper estimation, either.

Another point which is worth considering is that since the retinal databases are quite heterogeneous, we cannot go for sure regarding the dependencies of the algorithms of the ensemble found for a specific (in our case for the MESSIDOR) database. Thus, if we keep the individual accuracies of the ensemble members, but drop the dependency relations, it would be useful to know to what extent a new algorithm may ruin or improve the ensemble accuracy. Consequently, we investigate the case B in section V, when a new algorithm with accuracy p_7 is added to our current ensemble with no constraints are given for the dependencies. In other words, we check the intervals for the minimal and maximal accuracies of the extended system regarding the lower and upper estimation of the values $p_{n+1,k}$, respectively. The corresponding figures enclosed in Table VII can be determined by the solution of (16) under the conditions (17).

Accuracy of the new algorithm	$q^{\ominus}_{min}$	$q^{\oplus}_{min}$	$q^{\ominus}_{max}$	$q^{\oplus}_{max}$
p_7 = 0.7	0.764	0.899	1	1
p_7 = 0.9	0.908	0.934	1	1

TABLE VII
THE INTERVAL FOR THE MINIMAL AND MAXIMAL OD DETECTOR ENSEMBLE ACCURACY IF A NEW DEPENDENT ALGORITHM IS ADDED TO A SYSTEM WITH NO DEPENDENCY CONSTRAINTS.

Table VII indicates the natural fact that if the dependencies are unknown, the minimal and maximal accuracy can highly differ, and e.g. the ensemble performance can be worse than that of some of its members. However, it is also worth considering for our specific OD detector ensemble that a new algorithm of accuracy $p_7 = 0.9$ by all means will raise the minimal system accuracy given in (26). A comparison with Table VI shows that if we do not assume any dependencies for the original ensemble, we can reach higher maximal and lower minimal system accuracies.

For the strict majority voting approach, an ensemble with even number of members is meaningless, since as it is also known from classic theory [1] ensemble accuracy always drops for even number L of members regarding the $L-1$ case. So we have analyzed the change in accuracy, when the ensemble containing five members is extended to seven members. First of all, we have determined the most accurate ensemble with seven members from all the implemented eight algorithms. This ensemble includes the same six algorithms as listed before plus the one described in [30] having individual accuracy $p_7 = 0.320$. Then, we have selected the most/least accurate ensembles with five members, respectively, and checked which members were added to compile the ensemble with seven members. The corresponding quantitative results are given in Table VIII.

Indices of excluded members	2,5 (lowest acc.)	3,7 (highest acc.)
Ensemble accuracy (5 members)	0.626	0.944
Ensemble accuracy (7 members)	0.853	

TABLE VIII
CHANGE OF THE ENSEMBLE ACCURACY FOR STRICT MAJORITY, WHEN THE SIXTH AND SEVENTH MEMBER IS ADDED TO THE ENSEMBLE OF FIVE ALGORITHMS.

The results of Table VIII are quite obvious, since two individually highly accurate (p_2, p_5) and also two rather inaccurate (p_3, p_7) algorithms are present. Thus, their joint removal leads to a strong drop/increment regarding the ensemble accuracy, respectively.

VII. DISCUSSION AND CONCLUSIONS

In this paper, we have introduced a new model that enables the investigation of majority voting systems in the spatial domain. We have considered independent/dependent ensembles composed by classifiers having not necessarily the same individual accuracies. We have described how a constraint may raise from shape characteristics, and presented an ensemble-based system for optic disc detection in retinal images, where the object has a circular anatomical geometry. The general theory of ensemble-based systems describes several voting methodologies. However, for spatial voting, corresponding models have not been presented yet, and their adaptation is rather challenging to this domain. For instance, the extension of the approach proposed in this paper is currently under study for weighted spatial majority voting, but for several cases (e.g. dependent voters) it is far from being trivial. At this point, we were able to show the superiority of our proposed method over the strict version of majority voting (see section VI-C) which is a simple, but rather unnatural and less efficient extension.

Our detailed experimental studies have been performed on the image dataset MESSIDOR [24]. However, it is well-known that we can expect high variance among retinal image databases (see e.g. [31]), so tests on different datasets are recommended. Thus, to validate more its efficiency, we have tested the proposed ensemble-based approach on a database containing 327 images provided by the Moorfields Eye Hospital, London from a real mass screening process. The highest accuracy $q = 0.921$ has been found for the ensemble containing the four members having individual accuracies $p_1 = 0.798$, $p_3 = 0.150$, $p_4 = 0.801$, $p_5 = 0.835$, respectively (for the remaining three algorithms we have measured $p_2 = 0.780$, $p_6 = 0.342$, and $p_7 = 0.297$, respectively). Similarly to MESSIDOR, the ensemble performed better than any of its members for the Moorfields dataset, as well. Moreover, we can observe that the individual accuracies have been varied more among the different datasets than that of the ensemble. This observation suggests that we can expect a more stable and calculable behavior if we work with ensembles.

Our approach can be extended to other detection problems with keeping in mind that the presented results are suitable to handle such shapes that can be described by set diameter. To demonstrate the efficiency of our method, we considered another detection problem: the localization of the macula, which is the center of the sharp vision in the retina and appears as a dark, disc-like object of diameter approximately 6mm. That is, we have a very similar scenario to that of the OD detection problem. We have set up an ensemble of five macula detectors [32], [33], [34], [35], [36] having individual accuracies 0.583, 0.870, 0.714, 0.624, 0.962, respectively. By applying the proposed spatially constrained decision scheme, we have found 0.968 for the accuracy of the ensemble for

the dataset MESSIDOR. From this result we can see that our ensemble-based approach has led to improvement in this field, as well.

APPENDIX

Proof of Proposition 3.1: Since by Lemma 3.1 ξ is of Bernoulli distribution with parameter q, we have: $E\xi = q$. Thus, we just need to show that $q = p$ whenever $p_{n,k} = k/n$ $(k = 0, 1, \ldots, n)$. By our settings, from (4) we have:

$$q = \sum_{k=0}^{n} \frac{k}{n}\binom{n}{k}p^k(1-p)^{n-k} = \frac{1}{n}\sum_{k=0}^{n} k\binom{n}{k}p^k(1-p)^{n-k}.$$

Observe that the last sum just expresses the expected value np of a random variable of binomial distribution with parameters (n, p). Thus, we have $q = p$, and the statement follows. ∎

Proof of Theorem 3.1: We can write:

$$q = \sum_{k=0}^{n} p_{n,k}\binom{n}{k}p^k(1-p)^{n-k} = \sum_{k=0}^{\lfloor n/2 \rfloor}\left(p_{n,k}\binom{n}{k}p^k(1-p)^{n-k} + \right.$$

$$\left. + p_{n,n-k}\binom{n}{n-k}p^{n-k}(1-p)^k\right) + p_{n,n/2}\binom{n}{n/2}p^{n/2}(1-p)^{n/2}.$$

Here if n is odd, the last term should be considered to be zero.

Now by our assumptions $p \geq 1/2$, together with (i) and (ii), using also the identities $\binom{n}{k} = \binom{n}{n-k}$ and $k/n + (n-k)/n = 1$, for any k with $0 \leq k < n/2$ we have:

$$p_{n,k}\binom{n}{k}p^k(1-p)^{n-k} + p_{n,n-k}\binom{n}{n-k}p^{n-k}(1-p)^k \geq$$

$$(1-p_{n,n-k})\binom{n}{k}p^k(1-p)^{n-k} + p_{n,n-k}\binom{n}{n-k}p^{n-k}(1-p)^k =$$

$$= \frac{k}{n}\binom{n}{k}p^k(1-p)^{n-k} + \frac{n-k}{n}\binom{n}{n-k}p^k(1-p)^{n-k} +$$

$$+ p_{n,n-k}\binom{n}{n-k}\left(p^{n-k}(1-p)^k - p^k(1-p)^{n-k}\right) \geq$$

$$\geq \frac{k}{n}\binom{n}{k}p^k(1-p)^{n-k} + \frac{n-k}{n}\binom{n}{n-k}p^k(1-p)^{n-k} +$$

$$+ \frac{n-k}{n}\binom{n}{n-k}\left(p^{n-k}(1-p)^k - p^k(1-p)^{n-k}\right) =$$

$$= \frac{k}{n}\binom{n}{k}p^k(1-p)^{n-k} + \frac{n-k}{n}\binom{n}{n-k}p^{n-k}(1-p)^k.$$

In the last inequality, we use (ii) and the fact that $p^{n-k}(1-p)^k - p^k(1-p)^{n-k}$ is non-negative. Furthermore, if n is even, by (ii) we also have:

$$p_{n,n/2}\binom{n}{n/2}p^{n/2}(1-p)^{n/2} \geq \frac{n/2}{n}\binom{n}{n/2}p^{n/2}(1-p)^{n/2}.$$

Thus, we obtain:

$$q \geq \sum_{k=0}^{n} \frac{k}{n}\binom{n}{k}p^k(1-p)^{n-k} = p.$$

Here, the last equality follows from the proof of Proposition 3.1. Since $E\xi = q$, we have the inequality $E\xi \geq p$. ∎

Proof of Theorem 3.2: Let t and l be as given in the statement. Then, we have:

$$P(\xi^{\otimes t} \geq l) = \sum_{k=l}^{t}\binom{t}{k}q^k(1-q)^{t-k},$$

$$P(\eta_1^{\otimes t} \geq l) = \sum_{k=l}^{t}\binom{t}{k}p^k(1-p)^{t-k}.$$

Thus, by Lemma 3.2, we obtain:

$$P(\xi^{\otimes t} \geq l) \geq P(\eta_1^{\otimes t} \geq l),$$

if and only if, $q \geq p$, and the theorem follows. ∎

Proof of Theorem 4.1: Using (8) and our assumption $p_{n,k_1}/k_1 \geq p_{n,k}/k$ for all $k = 1, \ldots, n$, we get:

$$E\xi^{\otimes t} = \sum_{j=1}^{t} p_{n,u_j} = \sum_{\substack{j=1 \\ u_j \neq 0}}^{t} u_j p_{n,u_j}/u_j \leq$$

$$\leq \sum_{j=1}^{t} u_j p_{n,k_1}/k_1 = (p_{n,k_1}/k_1)\sum_{j=1}^{t} u_j = nrp_{n,k_1}/k_1,$$

which implies the first part of the statement.

Assume now that we also have $tk_1 = nr$. Fill in the $n \times t$ table T with zeros and ones arbitrarily, such that we have r ones in each row. If there is a column containing less than k_1 ones, then by $tk_1 = nr$ there is another column with more than k_1 ones. Write j_1 and j_2 for the indices of these columns, respectively. Then there exists a row say with index i, such that $T(i, j_1) = 0$ and $T(i, j_2) = 1$. Change these zero and one values, and continue this process as long as possible. Since $tk_1 = nr$, finally we end up with a table T containing r ones in each row and k_1 ones in each column. Then, we have:

$$E\xi^{\otimes t} = \sum_{j=1}^{t} p_{n,k_1} = tp_{n,k_1} = tk_1 p_{n,k_1}/k_1 = nrp_{n,k_1}/k_1,$$

and the theorem follows. ∎

Proof of Theorem 4.3: Let T be an arbitrary table having r ones in each row such that T has no column consisting only of zeros. Since $nr \geq t$, such a T exists (and can be easily constructed). In view of the proof of Proposition 4.1, for the corresponding $\xi^{\otimes t}$ we have:

$$P(\xi^{\otimes t} = t) = (1/n^t)\prod_{j=1}^{t} u_j.$$

If for some indices $1 \leq j_1, j_2 \leq t$ we have $u_{j_1} - u_{j_2} \geq 2$, then $(u_{j_1} - 1)(u_{j_2} + 1) > u_{j_1}u_{j_2}$ clearly holds. Hence moving a one from the j_1-th column to the j_2-th column (keeping its row; just as at the end of the proof of Theorem 4.1), the new value for $P(\xi^{\otimes t} = t)$ will be larger than the previous one. Continuing this process as long as possible, finally we obtain a table T, where for any indices $1 \leq j_1, j_2 \leq t$ we have $|u_{j_1} - u_{j_2}| \leq 1$. Obviously, this is equivalent to:

$$\lfloor nr/t \rfloor \leq u_j \leq \lceil nr/t \rceil \quad (1 \leq j \leq t).$$

Observing that for all such tables T the values $P(\xi^{\otimes t} = t)$ coincide, and these tables differ from each other only by a permutation of their columns, the theorem follows. ∎

Proof of Theorem 4.4: First, we have:

$$P(\xi^{\otimes t}=t)=\prod_{j=1}^{t}p_{n,u_j}=\exp\left(\sum_{j=1}^{t}\ln p_{n,u_j}\right).$$

On the other hand, by our assumption $(\ln p_{n,k_0})/k_0 \geq (\ln p_{n,k})/k$ for all $k=1,\ldots,n$,

$$\sum_{j=1}^{t}\ln p_{n,u_j}=\sum_{\substack{j=1\\u_j\neq 0}}^{t}\frac{u_j\ln p_{n,u_j}}{u_j}\leq\sum_{j=1}^{t}\frac{u_j\ln p_{n,k_0}}{k_0}=$$

$$=\frac{\ln p_{n,k_0}}{k_0}\sum_{j=1}^{t}u_j=\frac{nr\ln p_{n,k_0}}{k_0}$$

holds. Thus:

$$P(\xi^{\otimes t}=t)\leq p_{n,k_0}^{(nr/k_0)},$$

which implies the first part of the statement. The second part can be proved by following the argument at the end of the proof of Theorem 4.1. ■

ACKNOWLEDGMENT

The authors are grateful to the referees for their work and useful remarks to improve the quality of the content. This work was supported in part by the János Bolyai grant of the Hungarian Academy of Sciences; the project TÁMOP-4.2.2.C-11/1/KONV-2012-0001 supported by the European Union, co-financed by the European Social Fund; the OTKA grant NK101680; the project TÁMOP 4.2.1./B-09/1/KONV-2010-0007 implemented through the New Hungary Development Plan, co-financed by the European Social Fund and the European Regional Development Fund.

REFERENCES

[1] L. Lam and C.Y. Suen, "Application of Majority Voting to Pattern Recognition: An Analysis of Its Behavior and Performance," IEEE Trans. on Systems, Man, and Cybernetics, Part A: Systems and Humans, vol. 27, no. 5, pp. 553-568, Sep. 1997, doi:10.1109/3468.618255.

[2] L.K. Hansen and P. Salamon, "Neural Network Ensembles," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 12, no. 10, pp. 993-1001, Oct. 1990, 10.1109/34.58871.

[3] S. Cho and J. Kim, "Combining Multiple Neural Networks by Fuzzy Integral for Robust Classification," IEEE Trans. Systems, Man and Cybernetics, vol. 25, no. 2, pp. 380-384, Feb. 1995, doi:10.1109/21.364825.

[4] E.B. Kong and T. Diettrich, "Error-Correcting Output Coding Corrects Bias and Variance," Proc. 12th International Conference on Machine Learning (ICML 1995), pp. 313-321, 1995, doi:10.1.1.57.5909.

[5] K.M. Ali and M.J. Pazzani, "Error Reduction Through Learning Multiple Descriptions," Machine Learning, vol. 24, no. 3, pp. 173-202, Sept. 1996, doi:10.1023/A:1018249309965.

[6] T.K. Ho, J. Hull and S. Srihari, "Decision Combination in Multiple Classifier Systems," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 16, no. 1, pp. 66-75, Jan. 1994, doi:10.1109/34.273716.

[7] Y.S. Huang and C.Y. Suen, "A Method of Combining Multiple Experts for the Recognition of Unconstrained Handwritten Numerals,", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 17, no. 1, pp. 90-94, Jan. 1995, doi:10.1109/34.368145.

[8] L. Xu, A. Krzyzak and C.Y. Suen, "Several Methods for Combining Multiple Classifiers and Their Applications in Handwritten Character Recognition", IEEE Trans. on System, Man and Cybernetics, vol. 22, no. 3, pp. 418-435, May 1992, doi:10.1109/21.155943.

[9] K. Sirlantzis, S. Hoque, M.C. Fairhurst, "Diversity in Multiple Classifier Ensembles Based on Binary Feature Quantisation with Application to Face Recognition", Appl. Soft Comput., vol. 8, no. 1, pp. 437-445, Jan. 2008, doi:10.1016/j.asoc.2005.08.002.

[10] A. Perez-Rovira and E. Trucco, "Robust Optic Disc Location via Combination of Weak Detectors," Proc. IEEE Engineering in Medicine and Biology Society, 30th Annual International Conference (EMBS 2008), pp. 3542-3545, Aug. 2008, doi:10.1109/IEMBS.2008.4649970.

[11] T.J. Fuchs, J. Haybaeck, P.J. Wild, M. Heikenwalder, H. Moch, A. Aguzzi and J.M. Buhmann, "Randomized Tree Ensembles for Object Detection in Computational Pathology," Proc. 5th International Symposium on Advances in Visual Computing (ISVC 2009): Part I, pp. 367-378, 2009, doi:10.1007/978-3-642-10331-5_35.

[12] L.I. Kuncheva, *Combining Pattern Classifiers, Methods and Algorithms*, New Jersey: John Wiley & Sons, Inc., 2004, doi:10.1002/0471660264.

[13] L.I. Kuncheva C.J. Whitaker and C.A. Shipp, "Limits on the Majority Vote Accuracy in Classifier Fusion," Pattern Analysis and Applications, vol. 6, no. 1, pp. 22-31, Apr. 2003, doi:10.1007/s10044-002-0173-7.

[14] N. Littlestone and M. Warmuth, "The weighted majority algorithm," Proc. 30th Symposium on Foundations of Computer Science (SFCS), pp. 256-261, Nov. 1989, doi:10.1109/SFCS.1989.63487.

[15] J.Z. Kolter and M.A. Maloof, "Dynamic weighted majority: An ensemble method for drifting concepts" The Journal of Machine Learning Research, vol. 8, pp. 2755-2790, Dec. 2007.

[16] B. Harangi, J.R. Qureshi, A. Csutak, T. Peto, A. Hajdu, "Automatic Detection of the Optic Disc Using Majority Voting in a Collection of Optic Disc Detectors," Proc. 7th IEEE International Symposium on Biomedical Imaging (ISBI 2010), pp. 1329-1332, Apr. 2010, doi:10.1109/ISBI.2010.5490242.

[17] D. Gilat, "Monotonicity of a Power Function: An Elementary Probabilistic Proof," The American Statistician, vol. 31, no. 2, pp. 91-93, May 1977, doi:10.2307/2683050.

[18] A. Buonocorea, E. Pirozzi and L. Caputo, "A Note on the Sum of Uniform Random Variables," Statistics and Probability Letters, vol. 79, no. 19, pp. 2092-2097, Oct. 2009, doi:10.1016/j.spl.2009.06.020.

[19] H. Altincay, "On Naive Bayesian Fusion of Dependent Classifiers," Pattern Recognition Letters vol. 26, no. 15, pp. 2463-2473, Nov. 2005, doi:10.1016/j.patrec.2005.05.003.

[20] X. Wang and N.J. Davidson, "The Upper and Lower Bounds of the Prediction Accuracies of Ensemble Methods for Binary Classification," Proc. 9th International Conference on Machine Learning and Applications (ICMLA '10), pp. 373-378, Dec. 2010, doi:10.1109/ICMLA.2010.62.

[21] O. Matan, "On voting ensembles of classifiers," Proc. AAAI-96 workshop on Integrating Multiple Learned Models, pp. 84-88, Apr. 1996.

[22] M.J. Appel and R.P. Russo, "On the h-diameter of a Random Point Set," Technical Report 370, The University of Iowa, Jul. 2008.

[23] M.J. Appel, C.A. Najim and R.P. Russo, "Limit Laws for the Diameter of a Random Point Set," Advances in Applied Probability, vol. 34, no. 1, pp. 1-10, Mar. 2002, doi:10.1239/aap/1019160946.

[24] Dataset MESSIDOR provided by the Messidor program partners [Online]. Available: http://messidor.crihan.fr.

[25] M. Lalonde, M. Beaulieu and L. Gagnon, "Fast and Robust Optic Disc Detection Using Pyramidal Decomposition and Hausdorff-based Template Matching," IEEE Trans. Medical Imaging, vol. 20, no. 11, pp. 1193-1200, Nov. 2001, doi:10.1109/42.963823.

[26] A. Sopharak, K. Thet New, Y. Aye Moe, M.N. Dailey and B. Uyyanonvara, "Automatic Exudate Detection with a Naive Bayes Classifier," Proc. International Conference on Embedded Systems and Intelligent Technology (ICESIT2008), pp. 139-142, Feb. 2008.

[27] M. Niemeijer, M.D. Abramoff and B. van Ginneken, "Fast Detection of the Optic Disc and Fovea in Color Fundus Photographs," Medical Image Analysis, vol. 13, no. 6, pp. 859870, Sep. 2009, doi:10.1016/j.media.2009.08.003.

[28] A. Hoover and M. Goldbaum, "Locating the Optic Nerve in a Retinal Image using the Fuzzy Convergence of the Blood Vessels," IEEE Trans. Medical Imaging, Vol. 22, no. 8, pp. 951-958, Aug. 2003, doi:10.1109/TMI.2003.815900.

[29] S. Ravishankar, A. Jain and A. Mittal, "Automated Feature Extraction for Early Detection of Diabetic Retinopathy in Fundus Images," Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009), pp. 210-217, Jun. 2009, doi:10.1109/CVPR.2009.5206763.

[30] M. Lalonde, M. Beaulieu and L. Gagnon, "Fast and robust optic disc detection using pyramidal decomposition and Hausdorff-based template matching," IEEE Transactions on Medical Imaging, vol.20, no.11, pp.1193-1200, Nov. 2001, doi: 10.1109/42.963823.

[31] B. Antal and A. Hajdu, "An Ensemble-Based System for Microaneurysm Detection and Diabetic Retinopathy Grading," IEEE Transactions on Biomedical Engineering, vol.59, no.6, pp.1720-1726, Jun. 2012, doi: 10.1109/TBME.2012.2193126.

[32] T. Petsatodis, A. Diamantis and G.P. Syrcos, ”A Complete Algorithm for Automatic Human Recognition based on Retina Vascular Network Characteristics,” Proc. 1st International Scientific Conference e RA, pp. 41-46, Sep. 2004.
[33] S. Sekhar, W. Al-Nuaimy and A. K. Nandi, ”Automated localization of optic disc and fovea in retinal fundus images,” Proc. 16th European Signal Processing Conference (EUSIPCO), Aug. 2008.
[34] A.D. Fleming, S. Philip, K.A. Goatman, J.A. Olson and P.F. Sharp, ”Automated Assessment of Diabetic Retinal Image Quality Based on Clarity and Field Definition,” Investigative Ophthalmology and Visual Science, vol. 47, no. 3, pp. 1120-1125, Mar. 2006, doi: 10.1167/iovs.05-1155.
[35] L. Kovacs, R.J. Qureshi, B. Nagy, B. Harangi and A. Hajdu, ”Graph based detection of optic disc and fovea in retinal images,” Proc. 4th International Workshop on Soft Computing Applications (SOFA), pp. 143-148, Jul. 2010, doi: 10.1109/SOFA.2010.5565610.
[36] F. Zana , I. Meunier and J.C. Klein, ”A region merging algorithm using mathematical morphology: application to macula detection,” Proc. 4th International Symposium on Mathematical Morphology and its Applications to Image and Signal Processing, pp. 423-430, Jun. 1998.

Andras Hajdu received his MSc degree in Mathematics from the Lajos Kossuth University, Hungary, in 1996. He obtained his PhD degree in Mathematics and Computer Science from the University of Debrecen, Hungary, in 2003. From 2001 he served as Assistant Lecturer, since 2003 he has been an Assistant Professor, since 2009 an Associate Professor, and since 2011 he has been the Head of Department of Computer Graphics and Image Processing at the Faculty of Informatics, University of Debrecen. He is a member of the IEEE, the Janos Bolyai Mathematical Society, John von Neumann Computer Society (Hungary), Public Body of the Hungarian Academy of Sciences, member of the steering committee the Hungarian Association for Image Analysis and Pattern Recognition. He has authored or co-authored 26 journal papers and 88 conference papers. His main interest lies in discrete mathematics with applications in digital image processing.

Lajos Hajdu received his MSc degree in Mathematics from the Lajos Kossuth University, Hungary, in 1992. He obtained his PhD degree in Mathematics from the University of Debrecen, Hungary, in 1998. In 2011 he obtained the Doctoral degree of the Hungarian Academy of Sciences. From 1996 he served as Assistant Lecturer, since 1999 he has been an Assistant Professor and since 2003 an Associate Professor of Department of Algebra and Number Theory at the Faculty of Science and Technology Informatics, University of Debrecen. He is a member of the János Bolyai Mathematical Society, and the Mathematical Committee of the Hungarian Academy of Sciences. He has authored or co-authored 70 journal papers and 10 conference papers. His main interest lies in diophantine number theory, discrete tomography and discrete mathematics with applications in digital image processing.

Agnes Jonas graduated as an Applied Mathematician at the University of Debrecen, Faculty of Science in 2010. Her main fields of interests are probability theory, statistics and bioinformatics. Currently, she is a PhD student at the University of Veterinary Medicine Vienna.

Laszlo Kovacs received his MSc degree in Computer Science from University of Debrecen, Hungary in 2010. Currently, he is a PhD student at the Faculty of Informatics, University of Debrecen, Hungary. He is a student member of the IEEE and the John von Neumann Computer Society (Hungary). He has authored or co-authored 3 journal papers and 7 conference papers. His main interest lies in discrete mathematics with applications in digital image processing and high performance computing.

Henrietta Toman received her MSc degree in Mathematics from the University of Debrecen, Hungary in 2000. From 2001 she served as Assistant Lecturer, since 2009 she has been an Assistant Professor at the Department of Computer Graphics and Image Processing, Faculty of Informatics, University of Debrecen. She has authored or co-authored 5 journal papers and 6 conference papers. Her main interest lies in discrete mathematics and geometry.