Generalizing the majority voting scheme to spatially constrained voting

Andras Hajdu, *Member, IEEE,* Lajos Hajdu, Agnes Jonas, Laszlo Kovacs, and Henrietta Toman

Abstract—Generating ensembles from multiple individual classifiers is a usual appraoch to raise the accuracy of the decision. For decision majority voting is a popular rule. In this paper, we generalize classic majority voting by letting a further constraint to decide whether a correct or false decision is made if k correct votes is present among the total n ones. 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. The shape of the desired object define a geometric constraint the votes should obey to be able to decide together. Namely, the votes in this scenarion should fall inside a region matching the shape of the object. We give several theoretical result in this new model for both dependent/indipendent 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 on how our model is capable to characterize such a system and how the model can give a helping hand on the further improvability of the system, as well.

Index Terms—Computer Society, IEEEtran, journal, LaTeX, paper, template.

1 Introduction

Ensemble-based systems are rather popular to raise the classification accuracy by combining different sources (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. Strong focus is set to the combination of the labels of two (binary decision) or more classes. The combination of the votes 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 systems a cardinal issue is the assumptions on the dependency of the voters. Several results are gained for independent voters, but the minimum and maximum accuracy of such majority voting systems is also studied. In this paper, we investigate how such voting systems behave if we apply some further constraint on the votes. Namely, we generalize simple majority voting by introducing values $p_{n,k}$ for the probability that a good decision is made if k of n voters are correct. In other words, in our case it will be possible that a good decision is made even if the good votes are in minority.

The introduction of this new model is motivated by a medical image processing problem – the detection of the optic disc (OD) in retinal images. For an impression of the problem, see Figure 1 showing the optic disc, and the region of interest (ROI) of the retinal image.

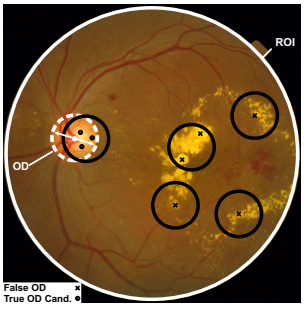

Fig. 1. The optic disc (OD) in a retinal image and some possible outputs of different OD detectors.

Organizing more individual OD detector algorithms into a voting system may raise detection accuracy [16]. In our approach, all of the OD algorithms return with the OD center as a single pixel. In this scenario, majority voting cannot be applied directly since besides the logical value of the votes their spatial placement are also important. Namely, we have considered discs of diameter of the OD (d_{OD}) covering the ouput of the detector algorithms. With this constraint, the circle having diameter d_{OD} with maximal number of candidates has been chosen for the optic disc. Note

- *A. Hajdu, L. Kovacs and H. Toman is with Department of Computer Graphics and Image Processing, University of Debrecen.*
- *L. Hajdu is with Institute of Mathematics, University of Debrecen.*

that, the diameter d_{OD} is a clinically predetermined constant.

The superiority of an ensemble over the individual algorithms motivated us to work out a corresponding theoretical model for the above constrained voting. Namely, the general $p_{n,k}$ term will be adjusted in this specific OD detection task by geometric constraints with requiring that the votes should fall inside a disc of a fixed diameter d_{OD}. In this combined system we can make a good decision even if the bad candidates have majority such as in the case illustrated in Figure 1. Bad decision can be made only when a subset of bad candidates with larger cardinality than the number of good ones can be bounded by a circle having diameter d_{OD}.

The rest of the paper is organized as follows. Section 2 recalls the basic concepts of the classical majority voting system as the basis for generalization. In section 3, we show how to incorporate constraints into this basic formulation. We present theoretical results and a demonstrative example for the case of indipendent voters. Since in applications independent detection algorithms are hardly expected, we also generalize to the dependent case in section 4 with including a corresponding demo example again. Especially, we investigate the possible lowest and highest accuracy of ensembles. Section 5 contains our empirical results regarding a true application (optic disc detection), where we apply our model to characterize of our current detector system and to analyze its further improvability. Finally, in section 6, we draw some conclusions.

2 Majority voting

Let $D = (D_1, D_2, \ldots, D_n)$ be a set of classifiers, $D_i : \mathbb{R}^k \to \Omega$ $(i = 1, \ldots, n)$, where Ω is a set of finite class labels. The majority voting rule assigns the class label supported by the majority of the classifiers D_i to x. 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 output is a binary vector) and all classifiers are independent and have the same classification accuracy p. An accurate class label is given by majority voting if at least $\lceil n/2 \rceil$ classifiers give correct answers. The majority vote method with independent classifier decisions gives an overall correct classification accuracy calculated by the following formula:

$$P = \sum_{k=0}^{\lfloor n/2 \rfloor} \binom{n}{k} p^{n-k}(1-p)^k. \tag{1}$$

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

3 Generalization to constrained voting

As we have already discussed in the introduction, we generalize the classic majority voting approach by considering some constraints that must be also fulfilled 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 5 we will adopt this general model to our practical problem with spatial constraints.

As we summarized in the introduction, several theoretical results are reached for indipendent voters, so we start with generalizing to this case first. However, in the vast majority of applications, we cannot expect indipendency among algorithms trying to detect the same object. Thus, we also generalize to the case of dependent voters with generalizing such formerly investigated concepts that have high practical impact, as well.

3.1 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 define the random variable ξ 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_i: if k out of the n classifiers make a good decision, then we make a good decision (i.e. we have $\xi = 1$) with probability $p_{n,k}$.

Note that in the special case where

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

we get the classical majority voting scheme.

The values $p_{n,k}$ for a fixed n as a function of k corresponding to the classic majority voting can be observed in Figure 2.

The ensemble accuracy of the classic majority voting system is enclosed in Table 3.1 for different number of classifiers (n) for some equal individual accuracies (p).

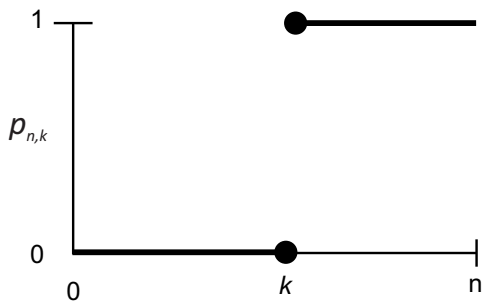

Fig. 2. The graph of $p_{n,k}$ for a fixed n for classic majority voting.

	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 1
Ensemble accuracy for classic majority voting.

First we show that ξ is of Bernoulli distribution, as well. We also provide the corresponding parameter q. In other words, in our model q represents the accuracy of the ensemble.

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

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

Proof: Since for any $k \in \{0,1,\dots,n\}$ we obviously 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, when we assume equal accuracy for the classifiers (i.e. $p = p_1 = \cdots = p_n$) received strong attention in the literature, so we generalize to this case first. In the rest of this subsection we assume $p = p_1 = \cdots = p_n$. Then,

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

Thus, by the particular choice (2) for the values of $p_{n,k}$, we get $q = P$ where P is given by (1). In order to have the majority voting be "better" than the individual decisions, we need only 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: Since by Lemma 3.1 ξ is of Bernoulli distribution with parameter q, we automatically have

$$E\xi = q.$$

Thus we need only to show that $q = p$ whenever $p_{n,k} = k/n$ $(k = 0,1,\dots,n)$. 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 of a random variable of binomial distribution with parameters (n,p). Thus we have

$$q = \frac{1}{n} np = p,$$

and the statement follows. □

Figure 3 also illustrates the special linear case for $p_{n,k}$ which assures equal ensemble q and individual accuracies p.

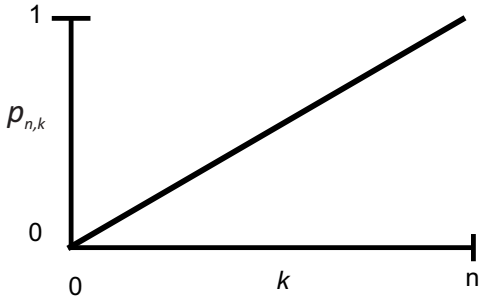

Fig. 3. The graph of $p_{n,k}$ for the linear case providing $p = q$.

As we mentioned already, $E\xi = q$ expresses that the "composite system" has accuracy q. Further, the above statement shows that if the probabilities $p_{n,k}$ increase uniformly (linearly), then the "composite system" 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 the new frame 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: We can write

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

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

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

Now by our assumptions $p \geq 1/2$, (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

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

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

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

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

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

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

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

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, in case of 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 $E\xi = q$, the inequality $E\xi \geq p$ immediately follows. □

As a simple consequence we obtain the following corollary concerning the classical majority voting scheme. Note that the next result is a theorem of Kuncheva et al. [13].

Corollary 3.2: Suppose that n is odd, $p \geq 1/2$ and for all $k = 0, 1, \ldots, n$ 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 of $p_{n,k}$, 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 where the composite system makes exclusively good decisions after t executions. So write $\xi^{\otimes t}$ for the random variable obtained by repeating ξ independently t times, and counting the number of 1 values (correct votes) obtained, 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 $i = 1, \ldots, n$, we certainly have $P(\eta_i^{\otimes t}) = p^t$. Here $\eta_i^{\otimes t}$ denotes the random variable obtained by repeating η_i independently t times, and counting the number of 1 values (correct votes) occured. To make the "combined system" 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 obviously, for any $x \in [0, 1]$ we 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: Let t and l be as given in the statement. Then we have

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

and

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

Thus by Lemma 3.2 we obtain that

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

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

3.2 Example 1 – demonstrating the independent case

We illustrate the results achieved for the independent case by a simple example for better understanding.

Suppose that n players play a game. Players can tell the truth with probability p or lie. Each player says a number. If one says $1/n$ that means telling the truth, if one says a number x_i independently from the interval $[-1/n, 0]$ that means telling a lie. Let k mean the number of true answers, in this way $n - k$ people tell lie. We get the final decision by adding

the numbers told by players. So we obtain the final decision by evaluating the expression below:

$$\sum_{i=1}^{n-k} x_i + \frac{k}{n}. \tag{5}$$

If (5) is positive, then we make a correct decision, otherwise we make a false one.

To characterize this simple game in our model note that the probability that a good decision is made in the case of k true answers can be calculated as:

$$p_{n,k} = P\left(\sum_{i=1}^{n-k} x_i + \frac{k}{n} > 0\right).$$

To have a closed formula for the above values $p_{n,k}$, we adopt some results regarding the distribution function of the sum of unifrom random variables from [18]. That is, we have

$$p_{n,k} = P\left(\sum_{i=1}^{n-k} x_i > -\frac{k}{n}\right) = 1 - P\left(\sum_{i=1}^{n-k} x_i < -\frac{k}{n}\right) =$$

$$P\left(\sum_{i=1}^{n-k} x_i < \frac{k}{n}\right) = \frac{\sum_{j=0}^{n-k} (-1)^j \binom{n-k}{j} \max\left(\frac{k-j}{n}, 0\right)^{n-k}}{(n-k)!\left(\frac{1}{n}\right)^{n-k}}.$$

The values $p_{n,k}$ for a fixed n as a function of k corresponding to Example 1 can be observed in Figure 4.

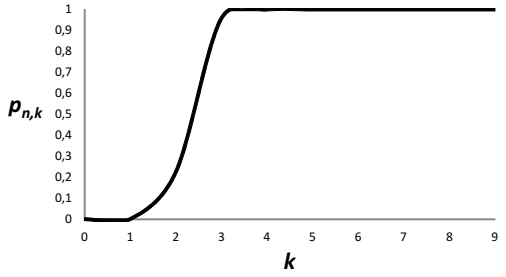

Fig. 4. The graph of $p_{n,k}$ for a fixed n for Example 1.

Now, the total accuracy q for this decision system can be calculated by the formula (4) we have already derived for independent classifiers. The ensemble accuracy q of Example 1 is enclosed in Table 3.2 for different number of classifiers (n) and for different probabilities for telling the truth by the players (p).

	n=3	n=5	n=7	n=9
p = 0.6	0.7914	0.8766	0.9134	0.9367
p = 0.7	0.8781	0.9476	0.9729	0.9846
p = 0.8	0.9438	0.9847	0.9951	0.9981
p = 0.9	0.9854	0.9982	0.9998	0.9999

TABLE 2
Ensemble accuracy for Example 1.

If we compare Table 3.1 with Table 3.2, we can see that for Example 1 the accuracy values of the decision system are greater than the corresponding ones for classic majority voting.

4 THE DEPENDENT CASE

In this section we investigate how dependencies among the voters influence the accuracy of the ensemble (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/maximal accuracy) of an ensemble. First we consider "pattern of success" and "pattern of failure" which are such realizations of the votes in a serie 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 discretization restriction of the model is needed to be made. Namely, not only the p_i 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 $E\eta_i = p_i$ $(i = 1, \dots, n)$), see e.g. [2], [20] and the references there, the vast majority of the results (such as e.g. in [13]) concern the case $p = p_1 = \cdots = p_n$. So in the next subsection we shall make the latter assumption, too. However, afterwards, in section 4.2, we give a much more general framework which handles both dependencies without the discretization restriction and also different accuracies of classifiers that makes the model realistic for applications. Finally, in section 4.3, we give an example similarly to Example 1, to illustrate the dependent behavior of the classifiers in our model.

4.1 Pattern of success and pattern of failure

Repeat the experiments $\eta_1, \dots, \eta_n$ t times, with some positive integer t, and write $\eta_i^{(j)}$ for the j-th realization of η_i. Suppose (as a rather strong, but standard assumption) that we have

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

Here r is a positive integer; to fit the previous more general model one can consider $r = np$. 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 above assumption. Write $\xi^{(j)}$ for the j-th realization of ξ $(j = 1, \dots, t)$. Then we clearly have $E\xi^{\otimes t} = E\xi^{(1)} + \cdots + E\xi^{(t)}$.

The number of 1 values are fixed for η_i, however, their positions are still free. 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 $\eta_i^{(j)}$ $(1 \le i \le n, 1 \le j \le t)$.

Our first result in this situation concerns the case of linear $p_{n,k}$.

Proposition 4.1: If $p_{n,k} = k/n$ for all $k = 0, 1, \dots, n$ 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 $j = 1, \dots, t$. Then we have $E\xi^{(j)} = u_j/n$. Thus

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

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

$$u_1 + \cdots + u_t = nr.$$

Combining the above equalities 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

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

where u_j is the number of ones in the j-th column of T. So to describe the pattern of success and the pattern of failure, 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. However, on the one hand the statement already in its form shows the essential method to be applied, and on the other hand, without these assumptions the statement would become rather technical. Further, as we have mentioned already, in the next subsection we describe a 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 $p_{n,k_1}/k_1 \ge p_{n,k}/k$ for all $k = 1, \dots, n$. Then $E\xi^{\otimes t} \le nrp_{n,k_1}/k_1$. Further, if $tk_1 = nr$ then the maximum can be attained.

Proof: As we noted already, we have

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

On the other hand, by our assumption $p_{n,k_1}/k_1 \ge p_{n,k}/k$ for all $k = 1, \dots, n$,

$$\sum_{j=1}^{t} p_{n,u_j} = \sum_{\substack{j=1 \\ u_j \neq 0}}^{t} u_j p_{n,u_j}/u_j \le$$

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

holds, 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 clearly have that

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

and the theorem follows. □

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 $p_{n,k_2}/k_2 \le p_{n,k}/k$ for all $k = 1, \dots, n$. Then $E\xi^{\otimes t} \ge nrp_{n,k_2}/k_2$. Further, if $tk_2 = nr$ then the minimum can be attained.

Proof: Since the proof of the statement follows the same lines as that of the previous theorem, we omit the details. □

We also consider the so-called multiplicative case for the pattern of success, which means to make only good decisions. 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 typically 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 $p_{n,k} = k/n$ we have the following result.

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

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

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: Let T be an arbitrary table having r ones in each row such that T has no column consisting only of zeros. Since $nr \ge t$, 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 $1 \le j_1, j_2 \le t$ we have $u_{j_1} - u_{j_2} \ge 2$, then clearly $(u_{j_1}-1)(u_{j_2}+1) > u_{j_1}u_{j_2}$. Hence "transiting" 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 $1 \le j_1, j_2 \le t$ we have $|u_{j_1} - u_{j_2}| \le 1$. Obviously, this is equivalent to

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

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

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.

In case of a general system $p_{n,k}$ 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 $0 < k \le n$. Let $k_0 \ne 0$ be an index such that $(\ln p_{n,k_0})/k_0 \ge (\ln p_{n,k})/k$ for all $k = 1, \dots, n$. Then $P(\xi^{\otimes t} = t) \le (1/n^t)p_{n,k_0}^{(nr/k_0)}$. Further, if $tk_0 = nr$ then the maximum can be attained.

Proof: We have

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

On the other hand, by our assumption $(\ln p_{n,k_1})/k_1 \ge (\ln p_{n,k})/k$ for all $k = 1, \dots, n$,

$$\sum_{j=1}^{t} \ln p_{n,u_j} = \sum_{\substack{j=1 \\ u_j \ne 0}}^{t} u_j(\ln p_{n,u_j})/u_j \le \sum_{j=1}^{t} u_j(\ln p_{n,k_0})/k_0 =$$

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

holds. Thus

$$P(\xi^{\otimes t} = t) \le (1/n^t)p_{n,k_0}^{(nr/k_0)},$$

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

4.2 Extremal accuracies by linear programming

We assumed earlier that the η_i $(i = 1, \dots, n)$ random variables (classifiers) are independent. In our application we consider different algorithms detecting the optic disc as classifiers. These algorithms can not be considered independent in all cases because it can happen that the performance of the algorithms is based on very similar conditions. In case of dependent algorithms we have to decide how to measure the dependencies of the algorithms. For this aim, we can consider the joint distribution of the outputs of the algorithms. So let

$$p_{a_1,\dots,a_n} = P(\eta_1 = a_1, \dots, \eta_n = a_n),$$

where $a_i = \{0, 1, *\}$. The star denotes any of the outputs: $* = 0$ or 1. The probabilities $p_{a_1,\dots,a_n}$ can be considered as the entries of the contingency table of $\eta_1, \dots, \eta_n$. The problem to determine the combination of classifiers achieving the best/the worst performance in classification is equivalent to solve the following linear optimization problem. Maximize/Minimize

$$q(p_{a_1,\dots,a_n}) = \sum_{k=0}^{n}(p_{n,k} \sum_{a_1+\dots+a_n=k} p_{a_1,\dots,a_n}) \qquad (6)$$

under the following conditions:

$$\sum_{a_i=1} p_{*,\dots,*,a_i,*,\dots,*} = p_i, (i = 1, \dots, n) \qquad (7)$$

$$\sum_{a_i=0} p_{*,\dots,*,a_i,*,\dots,*} = 1 - p_i, (i = 1, \dots, n)$$

$$\sum_{a_1,\dots,a_n} p_{a_1,\dots,a_n} = 1$$

$$p_{a_1,\dots,a_n} \ge 0, \ a_i \in \{0, 1\} \ \ (i = 1, \dots, n).$$

Here we assume that $E\eta_i = p_i$ $(i = 1, \dots, n)$ so the accuracy of the i-th detecting algorithm is p_i.

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

$$p_{a_1,\dots,a_n} = P(\eta_1 = a_1, \dots, \eta_n = a_n) =$$
$$= P(\eta_1 = a_1) \dots P(\eta_n = a_n).$$

That is, the entries of the contingency table can be determined by the probabilities $p_1, \dots, p_n$. In this case, the objective function $q(p_{a_1,\dots,a_n})$ in (6) can be written in the same form as in (3) for q.

We demonstrate our method by the following example. Take $n = 3$, and suppose that the accuracies of the algorithms (i.e. the expected values of η_1, η_2, η_3) are $p_1 = 1/2$, $p_2 = 2/3$, $p_3 = 3/4$, respectively. Further, let the $p_{n,k}$ values be given by

$$p_{3,0} = 0, \quad p_{3,1} = 2/3, \quad p_{3,2} = 1, \quad p_{3,0} = 1.$$

These values correspond to the following situation: if the "good" votes are in majority then we surely make a good decision, and already in case one "good vote" we can make a good decision with high probability.

Set, as before,

$$c_{a_1,a_2,a_3} = P(\eta_1 = a_1, \eta_2 = a_2, \eta_3 = a_3) \ \ (a_1, a_2, a_3 \in \{0, 1\}).$$

Then to maximize the accuracy q (the expected value of the composite system), we need to maximize the function

$$q := (2/3)(c_{1,0,0}+c_{0,1,0}+c_{0,0,1})+c_{1,1,0}+c_{1,0,1}+c_{0,1,1}+c_{1,1,1}$$

under the constraints

$$c_{1,0,0}+c_{1,0,1}+c_{1,1,0}+c_{1,1,1}=1/2,$$
$$c_{0,1,0}+c_{1,1,0}+c_{0,1,1}+c_{1,1,1}=2/3,$$
$$c_{0,0,1}+c_{1,0,1}+c_{0,1,1}+c_{1,1,1}=3/4,$$
$$c_{0,0,0}+c_{1,0,0}+c_{0,1,0}+c_{0,0,1}+c_{1,1,0}+c_{1,0,1}+c_{0,1,1}+c_{1,1,1}=1,$$
$$c_{a_1,a_2,a_3}\geq 0 \quad (a_1,a_2,a_3\in\{0,1\}).$$

Solving this standard linear programming problem (e.g. by Maple), we obtain that the maximum is given by $q=35/36$, taken at the values

$$c_{0,0,0}=c_{0,1,0}=c_{0,0,1}=c_{1,1,1}=0,$$
$$c_{1,0,0}=1/12,\ c_{1,1,0}=1/6,\ c_{1,0,1}=1/4,\ c_{0,1,1}=1/2.$$

So the "best choice" for the dependencies among η_1,η_2,η_3 correspond to the contingency table composed from the above values of c_{a_1,a_2,a_3}.

Finally we note that in a similar way one can easily compute the "worst choice" for the dependencies, too. For this purpose we need to minimize the value of the above given q, rather than maximize it, under the same constraints. Now a simple calculation by Maple gives that the "worst choice" of the contingency table entries is given by

$$c_{1,0,0}=c_{0,1,0}=c_{1,1,0}=c_{1,0,1}=0,$$
$$c_{0,0,0}=1/4,\ c_{0,0,1}=1/12,\ c_{0,1,1}=1/6,\ c_{1,1,1}=1/2$$

yielding the value $q=13/18$.

If a new classifier is added to an existing system, the accuracy of the new system is affected by two main properties of the new classifier: its accuracy and its correlation with the classifiers in the existing system. Let η_{n+1} be a random variable with $E\eta_{n+1}=p_{n+1}$.

If we consider that only the accuracies remain the same after adding η_{n+1} to the system, then a similar problem to the one defining in (6) has to be solved to determine $q(p_{a_1,\ldots,a_{n+1}}$.

Maximize/Minimize the function

$$q(p_{a_1,\ldots,a_{n+1}})=\sum_{k=0}^{n+1}(p_{n+1,k}\sum_{a_1+\ldots+a_{n+1}=k}p_{a_1,\ldots,a_{n+1}}) \tag{8}$$

with the following conditions:

$$\sum_{a_i=1}p_{*,\ldots,*,a_i,*,\ldots,*}=p_i,(i=1,\ldots,n+1)$$
$$\sum_{a_i=0}p_{*,\ldots,*,a_i,*,\ldots,*}=1-p_i,(i=1,\ldots,n+1)$$
$$\sum_{a_1,\ldots,a_{n+1}}p_{a_1,\ldots,a_{n+1}}=1,$$
$$p_{a_1,\ldots,a_{n+1}}\geq 0,\ a_i\in\{0,1\}\ \ (i=1,\ldots,n+1).$$

Here we assume that $E\eta_i=p_i$ $(i=1,\ldots,n+1)$ so the accuracy of the i-th detecting algorithm is p_i.

Since we have that

$$p_{n,k}\geq p_{n+1,k} \tag{9}$$

and

$$p_{n,k}\leq p_{n+1,k+1}, \tag{10}$$

so

$$Q_min\geq q(p_{a_1,\ldots,a_{n+1}}\geq Q_max$$

Where Q_min and Q_max is the solution of the problem with the condition $p_{n+1,k}=p_{n,k}$ and $p_{n+1,k}=p_{n,k-1}$, respectively.

If we consider that both the entries of the contingency table of $\eta_1,\ldots,\eta_n$ and the accuracies remain the same after adding η_{n+1} to the system, to determine the best/the worst choice for the new classifier to achieve the best/the worst performance for the new system the following linear optimization problem has to be solved. Maximize/Minimize the function

$$q(p_{a_1,\ldots,a_{n+1}})=\sum_{k=0}^{n+1}(p_{n+1,k}\sum_{a_1+\ldots+a_{n+1}=k}p_{a_1,\ldots,a_{n+1}}) \tag{11}$$

with the following conditions:

$$\sum_{a_i=1}p_{*,\ldots,*,a_i,*,\ldots,*}=p_i,(i=1,\ldots,n+1)$$
$$\sum_{a_i=0}p_{*,\ldots,*,a_i,*,\ldots,*}=1-p_i,(i=1,\ldots,n+1)$$
$$p_{a_1,\ldots,a_n}=\sum_{a_{n+1}=0}^{1}P(\eta_1=a_1,\ldots,\eta_n=a_n,\eta_{n+1}=a_{n+1}),$$
$$\sum_{a_1,\ldots,a_{n+1}}p_{a_1,\ldots,a_{n+1}}=1,$$
$$p_{a_1,\ldots,a_{n+1}}\geq 0,\ a_i\in\{0,1\}\ \ (i=1,\ldots,n+1).$$

Here we assume that $E\eta_i=p_i$ $(i=1,\ldots,n+1)$ so the accuracy of the i-th detecting algorithm is p_i.

In the special case, when η_{n+1} is totally independent from $(\eta_1,\ldots,\eta_n)$, the entries of the new contingency table can be determined by the probabilities $p_{a_1,\ldots,a_n}$ and p_{n+1}, where $E\eta_{n+1}=p_{n+1}$:

$$p_{a_1,\ldots,a_n,1}=P(\eta_1=a_1,\ldots,\eta_n=a_n,\eta_{n+1}=1)=$$
$$=p_{n+1}p_{a_1,\ldots,a_n}, \tag{12}$$
$$p_{a_1,\ldots,a_n,0}=P(\eta_1=a_1,\ldots,\eta_n=a_n,\eta_{n+1}=0)=$$
$$=(1-p_{n+1})p_{a_1,\ldots,a_n}.$$

Considering these equations, we get that the linear optimization problem to be solved can be written in a simpler form: maximize/minimize the function the

$$q(p_{a_1,\ldots,a_{n+1}})= \tag{13}$$
$$\sum_{k=0}^{n+1}(p_{n+1,k}(\sum_{a_1+\ldots+a_n=k}p_{a_1,\ldots,a_n,0}+\sum_{a_1+\ldots+a_n=k-1}p_{a_1,\ldots,a_n,1}))=$$

$$\sum_{k=0}^{n+1}(p_{n+1,k}(\sum_{a_1+\ldots+a_n=k}(1-p_{n+1})p_{a_1,\ldots,a_n}+$$

$$+\sum_{a_1+\ldots+a_n=k-1}p_{n+1}p_{a_1,\ldots,a_n}))$$

under the conditions defining in (7).
Here we assume that $E\eta_i = p_i$, $(i = 1,\ldots,n+1)$ so the accuracy of the i-th detecting algorithm is p_i.

If we consider that the entries of the contingency table of $\eta_1,\ldots,\eta_n$ remain the same after adding η_{n+1} to the system, the solution of the problem in (13) only depend on p_{n+1}. We get the following proposition for the system accuracies:
If

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

then

$$q(p_{a_1,\ldots,a_{n+1}}) \geq q(p_{a_1,\ldots,a_n}).$$

Since we have (9) and (10) that

$$\frac{\sum_{k=0}^{n}(p_{n,k}\sum_{a_1+\ldots+a_n=k}(p_{n,k}-p_{n+1,k}))}{\sum_{k=0}^{n}(p_{n,k}\sum_{a_1+\ldots+a_n=k}(p_{n+1,k+1}-p_{n+1,k}))} \geq 0.$$

In the particular case $p_{n,k} = k/n$ we obtain the following result.

Theorem 4.5: Let $\eta = (\eta_1,\ldots,\eta_n)$ be an n-dimensional random variable, where $E\eta_i = p_i$, $(i = 1,\ldots,n)$. We consider the joint distribution $p_{a_1,\ldots,a_n} = P(\eta_1 = a_1,\ldots,\eta_n = a_n)$. Let $p_{nk} = k/n$ $(k = 0,1,\ldots,n)$. Then we have $E\xi = \frac{p_1+\ldots+p_n}{n}$.

Proof: It follows from rearranging the sums in the following way:

$$E\xi = \sum_{k=0}^{n}\sum_{a_1+\ldots+a_n=k}\frac{k}{n}\cdot p_{a_1,\ldots,a_n} =$$

$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{a_i=1}p_{a_1,\ldots,a_n} = \frac{1}{n}\sum_{i=1}^{n}P(\eta_i = 1) = \frac{p_1+\ldots+p_n}{n}.$$

□

4.3 Example 2 – demonstrating the dependent case

Similarly to the independent case, we illustrate the results achieved for the dependent case by a corresponding variant of Example 1.

Suppose that n players play a game. Players can tell the truth with probability p or lie. Each player say a number. Truth-tellers say $1/n$, while liers a random x_i number from the interval $[-1/n, 0]$. k means the number of true answers, in this way $n-k$ people tell lie. However, unlike in Example 1, in this game the liers vote in a dependent way with saying the same random number, that is, $x_1 = x_2 = \cdots = x_{n-k} = x \in [-1/n, 0]$. Similarly to Example 1, for the final decision we evaluate the sum of the votes:

$$\frac{k}{n} + (n-k)x. \tag{14}$$

Namely, if (14) is positive then we make a correct decision, otherwise we make a false one.

For Example 2, the probability that we make a correct decision if k of the n votes are correct can be calculated as:

$$p_{n,k} = P\left(\frac{k}{n} + (n-k)x > 0\right) =$$

$$\begin{cases} P\left(x > -\frac{k}{n(n-k)}\right) = \min\left(\frac{k}{n-k}, 1\right), & \text{if } k < n, \\ 1, & \text{if } k = n. \end{cases}$$

The values $p_{n,k}$ for a fixed n as a function of k corresponding to Example 2 can be observed in Figure 5.

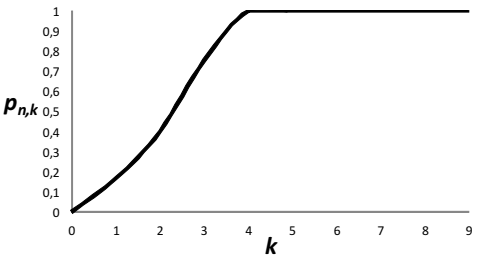

Fig. 5. The graph of $p_{n,k}$ for a fixed n for Example 2.

Now, the total accuracy q for this decision system can be calculated by the formula (4) we have already derived for independent classifiers. The ensemble accuracy q of Example 2 is enclosed in Table 4.3 for different number of classifiers (n) and for different probabilities for telling the truth by the players (p).

	n=3	n=5	n=7	n=9
p = 0.6	0.7929	0.8568	0.8886	0.9100
p = 0.7	0.8791	0.9331	0.9573	0.9712
p = 0.8	0.9443	0.9781	0.9899	0.9950
p = 0.9	0.9856	0.9970	0.9993	0.9997

TABLE 3
Ensemble accuracy for Example 2.

If we compare Table 3.2 with Table 4.3, we can see that **...@@@**

5 A CASE STUDY – OPTIC DISC DETECTION

In this section, we present a medical image processing application (optic disc detection) that motivated the creation of the constrained voting model. We start

with showing how the general formulation considering the $p_{n,k}$ probabilities 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 discussing on the possibilities of further improvement by exploiting some of our results.

5.1 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 $p_{n,k}$ the correct k votes must fall inside the true OD position, however, the $n-k$ false ones can fall within such a disc anywhere else within the ROI. 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 canidate of an algorithm is considered to be correct if its distance from the manually selected OD center is not larger than $d_{OD}/2$, see also Figure 7.

If we assume independency among the algorithms, for this case the behavior of the values $p_{n,k}$ as a function of k for a given n is shown in Figure 6.

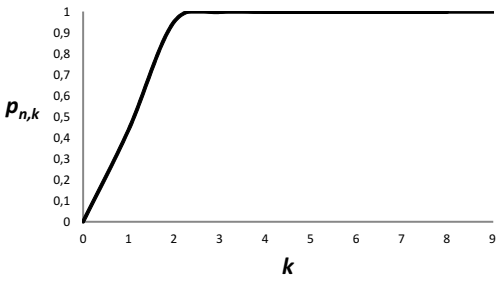

Fig. 6. The graph of $p_{n,k}$ for a fixed n with our geometric constraint to fall within a disc of diameter d_{OD}.

Figure 6 shows that the $p_{n,k}$ increase exponentially in k for a given n. This follows from the results of [21], [22] saying that the probability that the diameter of a point set is not less than a given constant decreases exponentially if 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 geometrically constrained system is enclosed in Table 5.1 for different number of independent classifiers (n) for some equal individual accuracies (p).

From Table 5.1 we can see a rapid increase in the ensemble accuracy e.g. in comparison with the accuracies found for Example 1 and 2, respectively. From trivial geometric considerations we can also see why the few number of votes (e.g. $n = 3$) performs bad.

Now, to describe the geometrically constrained case in detail, let us consider the probability $(1-p_i)r_i$ with $r_i \in [0,1]$ for the i-th independent classifier that means that the i-th classifier makes wrong classification and participates in making a bad decision. For the algorithm D_i with accuracy p_i giving a bad candidate (x_i, y_i) for the optic disc, we consider that the distribution of (x_i, y_i) is uniform outside the optic disc for all i $(i = 1, \ldots, n)$.

	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 4
Ensemble accuracy for applying our geometric constraint.

In this case, we have:

$$r_1 = \ldots = r_n = \frac{T_0}{T - T_0}, \tag{15}$$

where T_0 and T are the area of the optic disc and the ROI (region of interest in the image domain), respectively, so r_i is the same predetermined constant for all i $(i = 1, \ldots, n)$. For better understanding see also Figure 7.

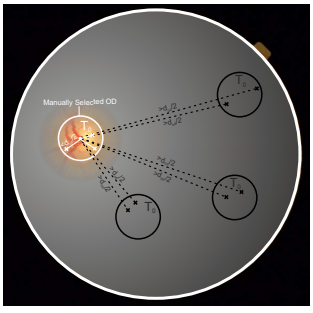

Fig. 7. The geometric constraint applied to the candidates of the algorithms: they should fall inside a disc of a fixed diameter d_{OD} to vote together.

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 bad votes are covered by the l disjoint discs of diameter d_{OD}, and k_i is the cardinality of the bad 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 k 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, bad 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 k votes among the circles, we can calculate the values $p_{n,k}$ as follows:

$$p_{n,k} = \sum_{k_1+\ldots+k_l=n-k, k>k_1} P(n,k,k_1,\ldots,k_l) \qquad (16)$$
$$+\frac{1}{2} \sum_{k_1+\ldots+k_l=n-k, k=k_1} P(n,k,k_1,\ldots,k_l).$$

The $p_{n,k}$ values calculated by (16) and the ones shown in Figure 6 are slightly differ. The reason of the difference is that in our geometric derivation to have a closed for, we have considered only disjoint discs that fall inside the ROI, as well. However, these differences are minor, and both approaches have exponential trends.

5.2 An ensemble-based OD detector

Progressive eye diseases can be caused by diabetic retinopathy (DR) which can even lead 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. The optic disc can be considered as a bright region with circular shape. In our automatic screening system originally we have collected eight OD detector algorithms to compose an ensemble from. Then, with a brute force approach (i.e. checking all the possible ensembles) we select such an ensemble which maximizes the accuracy of the combined system. In this way, we composed an ensemble of the following six OD-detectors:

- *Based on pyramidal decomposition:* Lalonde et al. [24] created an algorithm which generates a pyramid with simple Haar-based discrete wavelet transform. The pixel with highest intensity value in the low-resolution image (4th or 5th level of decomposition) is considered as the center of the OD. The individual accuracy of this algorithm was found to be $p_1 = 0.767$ on our dataset.
- *Based on edge detection:* This method [24] 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 lowest distance value is selected for OD center. The individual accuracy of this algorithm was found to be $p_2 = 0.958$ on our dataset.
- *Based on entropy measurement:* Sopharak et al. [25] 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 largest entropy value is selected as the OD center. The individual accuracy of this algorithm was found to be $p_3 = 0.315$ on our dataset.
- *Based on kNN classification:* Niemeijer et al. [26] 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 area. The centroid of the largest component found is considered as the OD center. The individual accuracy of this algorithm was found to be $p_4 = 0.759$ on our dataset.
- *Based on fuzzy convergence of blood vessels:* This method [27] 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 the center of the OD. The individual accuracy of this algorithm was found to be $p_5 = 0.977$ on our dataset.
- *Based on Hough transformation of vessels:* Ravishankar et al. [28] proposed to fit lines to the thinned vessel system by Hough transformation. The intersection of these lines results in a voting map. A weighting is also applied considering the intensity values corresponding to the intersection points. The pixel having highest voting result is considered as the center of the OD. The individual accuracy of this algorithm was found to be $p_6 = 0.647$ on our dataset.

For measuring the accuracy of both the individual algorithms and the ensembles we used the database Messidor [1] containig 1200 digital images, where the optic disc centers were manually selected by clinical experts. As for the decision of the ensemble, we select the disk with the OD-sized diameter containing the largest number of algorithm candidates. Then, as a final candidate we consider the centroid of these candidates. This final candidate is considered as correct, if it falls inside the disk centered at the manually selected OD center and havin the OD-diameter.

Using the theoretical foundations of the previous sections, we can characterize our current ensemble and can also check its improvability as given next.

5.3 Characterizing the OD-ensemble by the model

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

1. Kindly provided by the Messidor program partners (see http://messidor.crihan.fr).

The values $p_{n,k}$ for our application that are used to calculate the above characterizing ensemble accuracies as a function of k for a given n is shown in Figure 8.

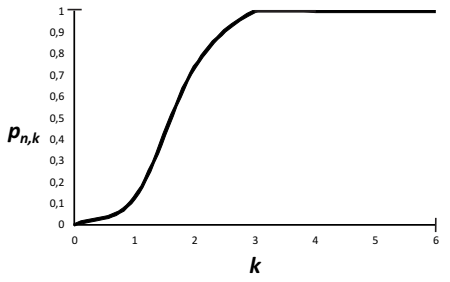

Fig. 8. The graph of $p_{n,k}$ for a fixed n with our geometric constraint.

Using the technique described in section 4.2, we have the following ensemble minimal and maximal accuracies, respectively:

$$q_{min} = 0.899, \qquad q_{max} = 1. \tag{17}$$

Based on our experiments the ensemble accuracy is:

$$q = 0.981, \tag{18}$$

which is quite close to the possible maximum accuracy. However, if we calculate the accuracy with the assumption of independency of the detectors, by formula @@@ we have:

$$q_{ind} = 0.998. \tag{19}$$

That is, we can see that an ensemble of independent algorithms would lead to nearly perfect results regarding accuracy. On the other hand, it is not suprising that our current system perform worse, since in this specific detection task it is quite challengint to find algorithms based on different (independent) principles.

5.4 On the improvability of the detector

Beside the above characterization, a more exciting issue is the improvability of an exisiting ensemble regarding its accuracy. For this study, we investigate to what extent the addition of a new algorithm D_{n+1} may improve the system. For this study we observe both the change of the system accuracy (q) and the interval for the minimal and maximal system accuracy. More precisely, we will consider the following cases:

A. we fix the 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,

B. we ignore the output of the algorithms of the current ensemble for an experiment, and
   1. add a new independent algorithm and check how the ensemble accuracy (q) changes by assuming that the ensemble members are independent,
   2. add a new dependent algorithm and check the minimal (q_{min}) and maximal (q_{max}) ensemble accuracy.

In section **??**, we have laid the theoretical background to extend the ensemble with adding a new classifier. Namely, we have formulated the ways of the calculation of ensemble accuracy for the cases, when the new classifier is dependent or independent from the ensemble, respectively. Besides the simple ensemble accuracy we have also explained how the minimal and maximal accuracy 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.

We start with the case (@@@), when the dependencies of the current ensemble members are considered as known in terms of a contingency table belonging to our test on the Messidor database and the new algorithm is considered to be independent from the ensemble. For this case we have the numerical results enclosed in Table 5.4. Note that, in this case we can check the interval where the ensemble accuracy will fall based on the lower and upper estimation that can be derived for $p_{n+1,k}$.

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

TABLE 5
The OD detector ensemble accuracy for the estimation of the values $p_{n+1,k}$ if a new independent algorithm of accuracy p_7 is added to a dependent system.

From Table 5.4 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 (18).

Next, we analyse the case (see @@@), when the dendencies of the algorithms are still considered, but the new algorithm should not be independent. In this setup, we can measure the accuracy interval for the minimal (q_{min}) and maximal ensemble accuracies (q_{max}), respectively, based on the estimation for the values $p_{n+1,k}$. The corresponding figures are included in Table 5.4.

Accuracy of a 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 6
Minimal and maximal OD detector ensemble accuracy for the estimation of the values $p_{n+1,k}$ if a new dependent algorithm of accuracy p_7 is added to a dependent system.

Table 5.4 also shows that it is quite natural that an individually very weak algorithm could lead to a remarkable improvement of the ensemble, however, this possibility is rather irrealistic. Moreover, since the current ensemble is not optimal regarding dependencies, even with a very diverse and accurate algorithm we cannot reach accuracy 1. It is also visible from the table that the original system accuracy (18) cannot be outperformed with the lowest $p_{n+1,k}$ estimation, and cannot be degraded with the highest $p_{n+1,k}$ estimation, as well.

Another point which is worth considering with the corresponding theoretical foundations are given in section **??** that since the retinal databases are quite heterogeneous, we cannot go for sure regarding the depencies of the algorithms already fixed in the ensemble found for a specific (in our case for the Messidor) database. Thus, if we keep the individual accuracies of the ensemble members, a useful information could be to see to what extent a new algorithm may ruin or improve the ensemble accuracy. In other words, we performed the analyses, where we ignored the contingency table found for the Messidor dataset. In Table 5.4 we enclosed the corresponding accuracy figures regarding the lower and upper estimation of the values $p_{n+1,k}$.

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

TABLE 7
The OD detector ensemble accuracy for the estimation of the values $p_{n+1,k}$ if a new independent algorithm of accuracy p_7 is added to an independent system.

By comparing Table 5.4 with Table 5.4 we can see that if we omit the depencies of the algorithms, we can expect higher ensemble accuracy. However, we have no information about the differences between datasets cauisng different dependencies among the ensemble members. Since the original ensemble would lead to very high accuracy with independent algorithms as given in (19), only in case of a very accurate new algorithm we can expect improvement.

Finally, we investigate the case, 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 minimal and maximal accuracy of the extended system regarding the lower and upper estimation of the values $p_{n+1,k}$. The current figures are enclosed in Table 5.4.

Accuracy of a 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 8
Minimal and maximal OD detector ensemble accuracy for the estimation of the values $p_{n+1,k}$ if a new dependent algorithm of accuracy p_7 is added to a system with no dependency constraints.

Table 5.4 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 $p_7 = 0.9$ will raise the minimal system accuracy given in (17) by all means. A comparison with Table 5.4 shows that if we do not assume any dependencies for the original ensemble, we can reach higher maximal and lower minimal system accuracies.

6 Conclusion

In this paper, we have worked out a new theoretical model that enables the investigation of majority voting systems being more general than the simple majority voting scheme. Namely, we have introduced a further constraint in the model to say the probability that a correct decision is made if k correct votes are among the total number n. We have derived several theoretical results for independent/dependent ensembles composed by classifiers have not necessarily the same individual accuracies. We have also embedded former concepts and results from the literature of classic majority voting, like the maximal (pattern of success) and minimal (pattern of failure) accuracy of an ensemble.

We have explained how a constraint may raise from shape characteristics of objects in a detection problem. Namely, we have presented an ensemble-based system for optic disc detection in retinal images, where the object has a circular anatomical geometry. In this case, the members of the ensembles are detector algorithms that give their results in terms of single pixels, as their candidate for the optic disc center. For this application we have shown how our results can be used for a quantitative and characterization of the combined system. In this specific scenario the probability constraint have an exponential behavior which

motivates the development of new algorithms, since in case of an independent (or even better, diverse) algorithm the system accuracy may be raised more rapidly then in classic majority voting. Again, we have given methods to see the influence of the addition of a new algorithm to the existing ensemble.

Our approach seems to be extendable to other detection problems (e.g. 3D organs, face components). However, for this aim further efforts are needed, since the geometric results presented in section 5.1 are primarily investigated for set diameter only, which is sufficient for the disc, but insufficient for more complex shapes.

ACKNOWLEDGMENTS

This work was supported in part by the Janos Bolyai grant of the Hungarian Academy of Sciences, and by the TECH08- 2 project DRSCREEN- Developing a computer based image processing system for diabetic retinopathy screening of the National Office for Research and Technology of Hungary (contract no.: OM-00194/2008, OM-00195/2008, OM-00196/ 2008). Research is supported in part by the OTKA grant NK101680 and by the TÁMOP 4.2.1./B-09/1/KONV-2010-0007 project. The project is implemented through the New Hungary Development Plan, cofinanced 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] 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.

[22] 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.

[23] A.E. Mahfouz and A.S. Fahmy, "Ultrafast Localization of the Optic Disc Using Dimensionality Reduction of the Search Space," Proc. 12th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI '09): Part II, pp. 985-992, 2009, doi:10.1007/978-3-642-04271-3_119.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

PLACE PHOTO HERE

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, 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 24 journal papers and 80 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 Biography text here.

Laszlo Kovacs Biography text here.

Henrietta Toman Biography text here.