

Conformal Anomaly Detection for visual reconstruction using gestalt principles

Ilia Nouretdinov*
Alexander Balinsky
Alex Gammerman

I.R.NOURETDINOV@RHUL.AC.UK
BALINSKYA@CARDIFF.AC.UK
A.GAMMERMAN@RHUL.AC.UK

Centre for Reliable Machine Learning, Royal Holloway University of London;
Department of Mathematics, Cardiff University

Abstract

In this paper, we combine a modern machine learning technique called conformal predictors (CP) with elements of gestalt detection and apply them to the problem of visual perception in digital images. Our main task is to quantify several gestalt principles of visual reconstruction. We interpret an image/shape as being perceivable (meaningful) if it sufficiently deviates from randomness - in other words, the image could hardly happen by chance. These deviations from randomness are measured by using conformal prediction technique that can guarantee the validity under certain assumptions. The technique describes the detection of perceivable images that allows to bound the number of false alarms, i.e. the proportion of non-perceivable images wrongly detected as perceivable.

Keywords: Conformal Anomaly Detection, classification, gestalt vision, visual reconstruction.

1. Introduction

In this paper, a novel approach to visual reconstruction of digital images is proposed and studied. By visual reconstruction we mean detection of meaningful parts of the image, without prior assumption or examples what is expected to be visible.

It is based on the framework of *conformal predictors* [V.Vovk et al. \(2005\)](#) or rather on Conformal Anomaly Detection (CAD) [R.Laxhammar \(2014\)](#); [J.Smith et al. \(2015\)](#). Its main advantage is that it guarantees bounds on the probability of error in the assumption that the data is i.i.d. or exchangeable. Our key assumption is that if an image looks like an anomaly when compared to the images of random noise, then it can be suspected to be “meaningful”. This is the reason to use CAD as the base of the work.

Two-dimensional digital images are made of discrete pixels. Stimuli arriving at a human retina are also discrete examples. How these separate measurements are organized in a human brain into geometrical examples like lines, cubes, circles, etc., is one of the main “enigmas of perception” [Kanizsa \(1997\)](#). Principles and rules that govern such visual organizations are called *principles of visual reconstruction* [Gombridge \(1971\)](#). Gestalt Theory is a single substantial scientific attempt to develop principles of visual reconstruction. Gestalt is a German word translatable as “whole”, “form”, “configuration” or “shape”.

* Corresponding author

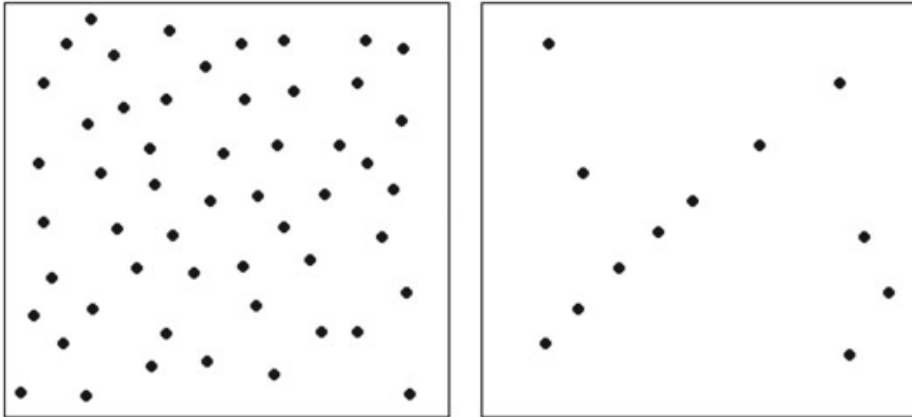


Figure 1: The Helmholtz principle in human perception.

The first rigorous approach to quantify basic principles of visual reconstruction was presented in [A.Desolneux et al. \(2008\)](#), and it is based on principles from image processing and especially on the Helmholtz Principle. According to the Helmholtz Principle, gestalts are sets of points organized by spatial arrangements. As a common-sense statement, this means that “events that could not happen by chance are immediately perceived”. For example, a group of seven aligned dots exists in both images in Fig. 1, but it can hardly be seen on the left-hand side image. Indeed, such a configuration is not exceptional given the total number of dots. In the right-hand image, we immediately perceive the alignment as a large deviation from the randomness that would be unlikely to happen by chance. The non-random element in the left image is not visible so clearly. Interestingly, human visual system is capable to understand these regular spatial arrangements almost effortlessly. Contrary to this, extracting structures by a computer is much more challenging.

Another form of the Helmholtz principle is called *the non-accidentalness principle*: an observed structure is considered meaningful (perceivable) when the relationship between its parts is too regular to be a result of an accidental arrangement of independent parts. In psychology [R.Arnheim \(1956\)](#), it is also found that “the overall structural features are the primary data of human perception, not the individual details”.

Gestalt theory of vision is based on the principal assumption that a meaningful (perceivable) visible structure can be described by a *hierarchy* of grouping operations. In this theory, the perception goes from the simplest elements (points) to lines (lines), then to the high-level ones (continued curves, line combinations, figures, etc.). The ways of grouping low-level examples into high-level ones are described by the list of special *gestalt principles* (grouping principles).

The general challenge for the Gestalt Theory of vision is that gestalt grouping may be applied in different orders, for many different parts and not reflecting real meaningful elements of the image. This can happen even starting from the first level where the points are organized into lines.

In [A.Desolneux et al. \(2008\)](#), the study of meaningful structures and groupings is based on hypothesis testing (also known as *a contrario* algorithms). Such algorithms use probability models not for the patterns to be recognized, but for random samples without the

pattern: the so-called null hypothesis. Instead of seeking patterns in an image that are highly likely in the model, we seek patterns in the image that are highly unlikely under the null hypothesis.

The main advantage of our approach is that we do not rely on a specific model of how the images are created. In our framework, the randomness of an image is tested with respect to a *data sample (training set)* and not to the distribution itself, that may be unknown. So, our approach is a machine learning type approach, i.e. data-driven and not model-driven.

In this article, we demonstrate that the framework of reliable learning (conformal prediction) provides guaranteed bounds on the probability of a *false alarms* i.e. the images that are random by the way of their generation but occasionally found to be perceivable. If an image is generated by the same distribution as the training set (of random images), it may be classified as an anomaly (i.e. a perceivable image) with bounded probability, that is equal to the pre-selected significance level.

We use ideas of the Gestalt Theory to define the *non-conformity measure (NCM)*. A non-conformity (strangeness) measure is a statistic used within the conformal framework for testing the hypothesis. In particular, we apply CAD not to the pictures themselves but to their *gestalt-profiles* that will be calculated according to the level of applicability of gestalt grouping operations within a concrete image.

In agreement with Gestalt Theory, we assume that the first level (combining the points into lines) is already important for the detection of perceivable images. It will be demonstrated that even by applying only this rule, we can get some interesting results. At the second level, we consider ones that apply to *pairs of lines*, i.e. describe pairwise geometrical relations. The statistic of such relationships may be collected immediately after the detection of the lines.

The paper is organized as follows. In Section 2, for the reader convenience, we remind key notions of the theories of gestalt vision and conformal prediction. In Section 3 we describe the data and main algorithms. The results are presented and compared in Section 4.

2. Related work

2.1. Gestalt principles

In this article, we use the gestalt grouping principles following their description in the book [A.Desolneux et al. \(2008\)](#). Usually, these principles are applied iteratively: a set of examples grouped on a lower level becomes an example that can be a member of a more high-level group. In our work, we focus our interest on 2–3 lowest levels: the points (zero level), the lines (the first level), and the relation of the lines (leading to the second level).

The *zero level* is the level of individual points that belong to the image. For simplicity, we look only on *binary* images, represented as square matrices of 0s and 1s, 1 for black, 0 for white. Therefore, we skip the principle of *color constancy*.

The *first level* is the level of *lines* (intervals) going through the individual *points*. The line is not necessarily a continuous one. it is just required to be straight and to connect some number of points (with the precision defined by the resolution level). We do not include variability of the width, leaving it for possible future extensions of the work.

The statistic of the number of first-level lines will be the prior information extracted from an image for further quantification of its meaningfulness.

However, we also use the elements of the *second level*, that is grouping of the first level lines into the figures by gestalt principles. In this work, we focus on those principles which apply to *pairs* of lines. Some review of used principles is given in Appendix A.

2.2. Conformal Prediction and Anomaly Detection

In this section, we briefly outline the main ideas of conformal predictors (CP) [V.Vovk et al. \(2005\)](#), [A.Gammerman and V.Vovk \(2007\)](#), and Conformal Anomaly Detection (CAD) that provides a framework for reconstructions of visual images. Originally, CPs were developed for classification or regression problems, where each example is divided into an *object* (typically, a feature vector) and a *label*. The output of conformal predictors is usually in the form of prediction set – that is a set of possible labels. The size of this *multi-label* prediction set depends on the required level of confidence (or significance level). The smaller the size (and ideally just one label) the more efficient CP are. If the actual true label is outside of the prediction set, this is a prediction error. This approach has guaranteed bounds on the probability of error given the required significance level under the assumption that the data is i.i.d. or exchangeable.

The central concept of CPs is a *non-conformity measure (NCM)*. NCM defines how strange a test example is with respect to the training set. There are many different NCM and the efficiency of CPs depends on how good the measure is. Following the CP theory we convert the NCM into the statistical notion of p-values. Using p-values, two parameters can be considered to assess how well a new, test example can be fit with the training set: *confidence*, which is (1-2nd largest p-value) and *credibility* which is just the largest p-value. The low credibility can be interpreted that the test example is not representative of the training set. In other words, the test example is not typical or abnormal. So, in principle, the CP approach allows us to detect an anomaly.

These ideas of CAD were developed further in several papers – see [R.Laxhammar \(2014\)](#), [I.Nouretdinov et al. \(2019\)](#), [Cherubin et al. \(2018\)](#). In this case, an example is *not* divided into an object and a label. The key idea of CAD is that when an example is outside the prediction set given a certain significance level, it leads to an alarm of its suspected *abnormality* with respect to the training set.

CAD is defined by Algorithm 1. It tests the hypothesis that an example is generated by the same distribution as a training set of examples. The core detail of CAD in the Non-Conformity Measure (NCM) function that is a kind of information distance between an object (typically, a feature vector) and a set of the same kind objects. For each (*i*th) example, NCM is applied to this example and the set of remaining examples, and get the output value α_i . The classification for each example is made by comparing its value NCM α_{l+1} with NCM α_j of the training examples, and obtaining the *p*-value. Note that adding a testing example to the training set may change NCM output values for the training examples. Therefore, if the algorithm is applied to many test examples, they have to be re-calculated each time. The prediction set is the set of the examples that would be assigned *p*-value above a significance level (threshold) ε . Low *p*-value (below the threshold) assigned to this

hypothesis means that the example is likely to be an *anomaly (outlier)* for the training data set.

Algorithm 1 CAD

```

1: INPUT: training data set  $\{z_1, \dots, z_l\} \subset Z$  (where  $z_i$  may be a feature vector).
2: INPUT: NCM function  $A : Z^{(*)} \times Z \rightarrow R$ 
3: INPUT: test example  $z_{l+1}$ 
4: INPUT: significance level(s)  $E = \{\varepsilon_1, \varepsilon_2, \dots, \varepsilon_{w-1}, \varepsilon_w\} \subset [0, 1]$ 
5: for  $j \in 1, 2, \dots, l + 1$  do
6:    $\alpha_j := A(\{z_1, \dots, z_{j-1}; z_{j+1}, \dots, z_{l+1}\}, z_j)$ 
7: end for
8:  $p := \frac{|\{j \in 1, 2, \dots, l+1 : \alpha_j \geq \alpha_{l+1}\}|}{l+1}$ 
9: for  $i \in 1, \dots, w$  do
10:   $e := \varepsilon_i$ 
11:  if  $p < e$  then
12:    OUTPUT anomaly alarm for  $z_{l+1}$  at level  $e$ 
13:  end if
14: end for

```

In the context of computer vision, we consider non-perceivable (random) images as typical ones, and anomalies are suspected perceivable images. If another random image is predicted as an anomaly, this is a mistake (‘false alarm’) because it was not perceivable. On the other hand, a perceivable image might be *not* marked as an anomaly, this may mean that the chosen significance level the test for anomalies is not set up low enough, and the CAD system makes an error.

CAD is valid in the sense that the probability of the *false alarm* is bounded by the given significance value (the ‘false alarm’ means that the anomaly alarm has been wrong). The validity means: if the example is generated by the same distribution as in the training set, the probability that it is outside of the prediction set is bounded by the threshold ε that was selected in advance. Like it is done in statistics, this threshold is usually set to 0.05 or 0.01.

Our aim here is to find whether an image is perceivable or non-perceivable. This problem can be interpreted as a binary classification. However, we treat it as anomaly detection since, for the binary classification, the examples of *both* classes are needed for training but they may not be always available. That is we may not have the perceivable images in our data set, while the non-perceivable images can be easily generated. In other words, it is often harder to have a set of *representative* perceivable images. A new image can be perceivable but does not have much in common with ones collected before. Therefore, the question can be formulated: ‘is the input image anomaly with respect to the set of non-perceivable images?’ If yes, then it can be marked as a perceivable one, even it does not show similarity with the set of perceivable images.

CAD heirs an important property of the Conformal Prediction framework that is the ability to adopt other machine learning methods inside. As known from the experience with conformal prediction, if a machine learning algorithm is efficient for the data, the same is mostly true for its conformal version.

The most commonly used underlying methods for CAD were k -Nearest-neighbors (kNN) [R.Laxhammar \(2014\)](#); [I.Nouretdinov et al. \(2019\)](#) or Kernel Density Estimation (KDE) [J.Smith et al. \(2015\)](#). In this paper, we also using Support Vector Machine (SVM). The SVM-based outlier detection was used e.g. in [C.Dawson](#) and [Mallinson and Gammerman \(2003\)](#).

3. Data preparation

3.1. Data set: USPS

As the example of perceivable images, we take hand-written digits from USPS (United States Postal Service) data set. It consists of 16x16 size handwritten digits, where a feature means the grey-scale intensity of a pixel from -1 to 1.

Originally, each image was supplied with a label of one of 10 classes (0,1,...,9). However, our task is not the distinction between them. What we wish is to separate them altogether from non-perceivable images. Therefore, the original labels are not used in the current work.

3.2. New classes

For different versions of this problem, we have created the following classes, that may play the role of perceivable or non-perceivable, dependently on the point of view.

1. The class of *real images* taken directly from the database in random order.
2. The class of *mixture images* where a noisy image is a mixture¹ of two randomly chosen images from the database.
3. The class of *random noise* i.e. points generated according to the uniform distribution on the square, independently on each other.

3.3. Pre-processing

To each of the images, we apply the following changes.

First, the images in the USPS data set are almost binary (the most pixels values are -1 or +1, intermediate values typically appear only on a digit's boundary), therefore we made them binary (positive values transformed to 1, negative ones to 0). After that, a pixel with value +1 will be also called a *black* pixel, and the whole image can be considered as a finite set of *points* (black pixels).

Second, for our aims, each image is *equalized* in such a way that they all have the same area (or, in other words, the same number of points). This area is set to be 1/8 of the whole image, that is 32 pixels of 256. The equalizing is done by adding points randomly to the image if the actual area is below, or by deleting some points randomly if it is above. Some examples of how this looks can be found in the Appendix. This way of equalizing was chosen also due to the domination of the extreme values (+1 and -1) in the data. Because

1. By the mixture of two images we mean that each pixel is randomly taken from one or the other of them, with independent processing of different pixels.

of that, thresholding can not always reach the goal. Another reason is to make the task a bit more hard and realistic by modeling the background noise level added to the pictures.

As we discussed in the Related Work Section 2, if an image is likely to be perceivable, then it contains the groups created with gestalt principles. They appear much more frequently than in a random (noisy) image with the same number of points.

At the first level, a group is a set of the points which are located along a straight line (or, in other words, within a narrow rectangle). The amount of points within such a group is an integer parameter that varies. This inspired us to start with creating so-called *gestalt-profiles* related to the picture: a gestalt profile is a vector whose indices are positive integers, and the content of a dimension is the number of lines which can be found in the picture. The indexing is needed because the definition of the line depends on what minimal amount of points a line is required to connect. This is the basic version of gestalt profiling, more complex ones can be created by taking into account the gestalt principles.² We detail this in Sections 3.4 and 3.5.

Then, an image is considered strange (anomalous) if its gestalt-profile is strange with respect to the set of gestalt-profiles of noisy images. At this stage, we apply the methods of CAD for anomaly detection as described in Section 3.6 to gestalt-profiles, not to the images themselves.

Finally, we will get p -values for the images that are low when an image is suspected to be non-random, and therefore possibly perceivable.

3.4. Basic analysis of images

Remind that an image U is finally presented as a $m \times m$ matrix of pixels, filled with 0s ('white') and 1s ('black'). So, we consider an image as a black drawing on a white background.

By a *point* (belonging to an image) we mean a 'black' position (p, q) (where $1 \leq p \leq m$, $1 \leq q \leq m$) so that

$$U_{pq} = 1.$$

By a *line* we mean an line connecting two points (u_1, v_1) and (u_2, v_2) , i.e. formally the set

$$\{(p_1\alpha + (1 - \alpha)p_2, q_1\alpha + (1 - \alpha)q_2) : 0 \leq \alpha \leq 1\}.$$

We say that it *connects* (*goes through*) another point (p, q) if it is matched by the rounded value of these coordinates for some concrete α . In other words, the precision matches the resolution: going of a line through a pixel is the same as including a point whose rounded coordinates are in the center of this pixel.

Each line provides three levels of information that can be extracted:

1. *vector* – the difference between the ends³
2. *line ends* – the coordinates of the ends within the image;
3. *line content* – the positions of the points within the line.

2. A similar language for image understanding is also provided in Scalar Vector Graphic.

3. To choose the vector direction in a unique way, we use the lexicographical ordering ($<$) as it will be done in Algorithm 2. This in fact means that the second end (p_2, q_2) lies above the first one (p_1, q_1) on the image.

3.5. Creating gestalt-profiles

Extracting gestalt-profiles from the images is done by the rule described with Algorithm 2. Here are some comments on it.

The set $Q = Q_i$ (it depends on i , but we will drop the index where possible) includes all lines going through i or more points (including the line ends). The notation V_i is used for simplification of Q_i where a line is interpreted only as a vector.

The parameter i is a variable integer threshold. Note that in our definition the ends should be points (pixels black with 1s), therefore looking at $i < 2$ does not make sense. Also, due to the equalizing pre-processing step (Sec. 3.3), the size of V_2 becomes a constant. Therefore, the smallest essential value is $i = 2$ in general, and $i = 3$ in our experiments. Its maximal value can be set to $2m + 1$ because a straight line can not intersect more pixels of a square image with resolution $m \times m$.

The vector $(|V_2|, \dots, |V_{2m+1}|)$ is considered as the basic version of gestalt-profile (the case $T = 1$ in Algorithm 2). It can be understood as the general statistic of the lines visible in the image, where a line means a series of points lying on a straight line (where straightness is understood to the existing resolution) but not necessarily in a continuous way.

The other values of the parameter T correspond to more advanced versions of gestalt-profile, involving more gestalt principles. The principle of similarity is involved in a slightly different way as the others, using another special parameter T_1 . The full list of combinations of parameters can be found in Appendix B.

3.6. Prediction

As mentioned earlier, we do not rely on having an ‘image’ class for training. One of three available data classes is selected as a background ‘non-perceivable’ class, and the other one works as a source of ‘perceivable’ examples for testing (or, in the control experiments, it may be the same class). However, at each step only one ‘perceivable’ item (called A_i) is compared against a set of ‘non-perceivable’ examples (called B_1, \dots, B_N).

The basic approach is CAD [R.Laxhammar \(2014\)](#) derived from algorithmic learning theory [V.Vovk et al. \(2005\)](#). The CAD framework is using as a meta-parameter for the NCM that is linked to an underlying anomaly detection (scoring) algorithm. As discussed in Section 2, we use two versions (k-Nearest-neighbors and Support Vector Machines). These two underlying methods are selected for being transparent and universal so that the conclusion would not depend too much on fitting the model to the data.

For convenience, we present CAD in two corresponding ready-for-use version: k -Nearest-neighbors CAD (Algorithm 3), and SVM CAD (Algorithm 4). Both of them apply to gestalt-profiles. We present them as separate pseudo-codes because, in addition to NCM they are also different in details of pre-processing (re-scaling).

3.7. Interpretation of the output

The output of Algorithm 3 or 4 can be interpreted as follows.

If A_i gets a low p-value p_i , this means that is detected as anomaly with respect to the training images B_1, \dots, B_N . In other words, if p_i is below the threshold, the image A_i is

Algorithm 2 Creating gestalt-profile of a image

```

INPUT: image  $U$  (binary, resolution  $m \times m$ )
INPUT: gestalt function type  $T$ 
INPUT: line analyser type  $T_1$ 
INPUT: resolution parameter  $K$  (for  $T_1 > 0$  only)
for  $i := 2$  TO  $2m + 1$  do
   $V := \emptyset$ 
   $Q := \emptyset$ 
  for  $(p_1, q_1)$  IN  $\{(p, q) : U_{p,q} = 1\}$  do
    for  $(p_2, q_2)$  IN  $\{(p, q) : U_{p,q} = 1, (p_1, q_1) < (p, q)\}$  do
      if the connecting line intersects at least  $i$  black pixels of  $U$  then
        for  $k := 0$  TO  $K$  do
           $\psi_k := U(\text{round}(p_1 \frac{1-k}{K} + \frac{k}{K} p_2), \text{round}(q_1 \frac{1-k}{K} + \frac{k}{K} q_2))$ 
        end for
         $\psi := (\psi_0, \psi_1, \dots, \psi_K)$ 
         $V := V \cup \{(p_2 - p_1, q_2 - q_1)\}$ 
         $Q := Q \cup \{(p_1, q_1; p_2, q_2; \psi)\}$ 
      end if
    end for
  end for
  define  $W_i = W_i(T, T_1, Q, V)$  according to Appendix B.
end for
OUTPUT vector  $(W_1, W_2, \dots, W_{2m+1})$ 

```

suspected to be perceivable. This corresponds to the usual interpretation of CAD, with a note that ‘abnormality’ in our case means that the image is more structured and less noisy.

A large number of low p-values assigned to A_i examples means that the test was efficient enough to ‘feel’ that it is unlikely to be from the same distribution as the ‘non-perceivable’ ones. The details of how it is measured for a series of examples will be provided further in the experimental Section 4.2.

4. Experiments

4.1. Problem statements

The task of our interest is the following: is it possible to detect a meaningful image as an anomaly, if it is inserted between the noisy (senseless) ones?

The possible tasks which we consider are listed in Table 1. In the detection problems, the classification is successful if we mostly able to detect the less noisy images as anomalies to a more noisy class.

In the control tasks, the goal is the opposite: if an image is classified as an anomaly, this is considered as an undesirable *false alarm*. The conformal prediction framework allows to bound their proportion by the pre-selected significance level.

Algorithm 3 Conformal assessment of images (k Nearest neighbors version)

INPUT: training (non-perceivable) gestalt-profiles B_1, \dots, B_N of length W .
 INPUT: gestalt-profiles A_1, \dots, A_n of length W for testing.
 INPUT: number of neighbors k
 INPUT: re-scaling (binary option ON/OFF)
for $i := 1, \dots, N$ **do**
 LET $B_{N+1} := A_i$
 if re-scaling is ON **then**
 for $j = 1, \dots, W$ **do**
 linearly rescale j -th dimension of B_1, \dots, B_{N+1} s.t. its min. is 0 and its max. is 1.
 end for
 end if
 for $u := 1, \dots, N + 1$ **do**
 for $v := 1, \dots, N + 1$ **do**
 LET $D(u, v)$ be Euclidean distance between B_u and B_v
 end for
 LET α_u be sum of k smallest values of $D(u, v)$ for $v = 1, \dots, u - 1, u + 1, \dots, N + 1$
 end for
 LET $p_i := \frac{|\{u=1, \dots, N+1: \alpha_u \geq \alpha_{N+1}\}|}{N+1}$
end for
 OUTPUT: p_1, \dots, p_n

'perceivable' class:	real images	mixture images	random noise
'non-perceivable' class			
real images			
mixture images	detection	control	
random noise	detection	detection	control

Table 1: Statements of the problem tasks

For the detection tasks, the smaller p -value is, the better for the efficiency of the algorithm, the ideal result is p -values about 0. For the control tasks, the ideal result is p -value distributed uniformly from 0 to 1.

Note that the 'internal background' class G created within Algorithm 4 is just an auxiliary artificial class used for the realization of one-class SVM, and is not related to any classes of the input data.

4.2. Evaluation criteria

We consider the experimental statements with the detection tasks 1–3 (numerated as in Table 1), and assess the quality of each of them by the following criteria of efficiency.

The evaluation criteria for conformal prediction were discussed in Vovk et al. (2017), but in the context of the full classification framework, for anomaly detection task we can use just the following simple performance measures:

Algorithm 4 Conformal assessment of images (SVM version)

INPUT: training (non-perceivable) gestalt-profiles B_1, \dots, B_N of length W .
 INPUT: gestalt-profiles A_1, \dots, A_n of length W for testing.
 INPUT: kernel function K (Polynomial of degree d , or RBF of radius γ)
 INPUT: parameter C for SVM (set to $C = 0$ in all the experiments)
 INPUT: re-scaling (binary option ON/OFF)
for $i := 1, \dots, N$ **do**
 LET $B_{N+1} := A_i$
 if re-scaling is ON **then**
 for $j = 1, \dots, W$ **do**
 linearly rescale j -th dimension of B_1, \dots, B_{N+1} s.t. its mean is 0 and its std. is 1.
 end for
 end if
 for $j = 1, \dots, W$ **do**
 find the average $M_j := \frac{B_1^j + \dots + B_{N+1}^j}{N+1}$
 find the variance $\sigma_j := \sqrt{\frac{(B_1^j - M_j)^2 + \dots + (B_{N+1}^j - M_j)^2}{N+1}}$
 end for
 creating the ‘internal background’ class:
 for $j := 1, \dots, N + 1$ **do**
 $G_j :=$ random normally distributed vector with average (M_1, \dots, M_W) and diagonal covariance matrix $(\sigma_1, \dots, \sigma_W)$
 end for
 run dual form SVM with kernel K and parameter C on two classes B and G
 get Lagrange multipliers $\alpha_1, \dots, \alpha_{N+1}$ for the class B
 LET $p_i := \frac{|\{u=1, \dots, N+1: \alpha_u \geq \alpha_{N+1}\}|}{N+1}$
 end for
 OUTPUT: p_1, \dots, p_n

1. median p-value (MPV);
2. average p-value (APV);
3. average log p-value (ALPV).

In all of these criteria, lower values reflect better efficiency (sensitivity of the tests, separability). The third criterion is considered as prior because of the special importance of low p-values for anomaly detection. We can refer to [I.Nouretdinov \(2007\)](#) where this criterion was suggested and justified.

The examples of performed control tasks 4 and 5 can be found in the Appendix. They are just confirming the validity of the method (bounded false alarm rate), and are not essential for evaluation of the efficiency.

4.3. Results

Tab. 2 compares the results according to Algorithms 3 and 4 with variations of settings.

Statement: Training on: Testing on:		1 (2) mixture images (1) real images			2 (3) random noise (1) real images			3 (3) random noise (2) mixture images			
T	T_1	method	MPV	APV	ALPV	MPV	APV	ALPV	MPV	APV	ALPV
vectors		kNN									
1	0	$k = 5$	0.213	0.292	-2.11	0.00387	0.0450	-5.39	0.0112	0.0660	-4.48
2	0	$k = 5$	0.187	0.298	-2.09	0.00728	0.0283	-4.95	0.0149	0.0587	-4.21
3	0	$k = 5$	0.223	0.296	-2.17	0.0100	0.0456	-4.82	0.0181	0.0549	-3.81
4	0	$k = 5$	0.178	0.315	-2.12	0.00528	0.0133	-5.32	0.0105	0.0232	-4.86
line ends		kNN									
5	0	$k = 5$	0.138	0.292	-2.23	0.00699	0.0293	-5.12	0.0109	0.0200	-4.63
6	0	$k = 5$	0.189	0.292	-2.18	0.0131	0.0487	-4.68	0.0195	0.0440	-3.84
7	0	$k = 5$	0.191	0.277	-2.09	0.0116	0.0901	-4.43	0.0777	0.1471	-2.89
3,5	0	$k = 5$	0.149	0.280	-2.31	0.00817	0.0232	-5.09	0.0113	0.0252	-4.39
3,5,7	0	$k = 5$	0.172	0.267	-2.27	0.00913	0.0246	-5.02	0.0157	0.0311	-4.12
3,6,7	0	$k = 5$	0.178	0.273	-2.24	0.0106	0.0361	-4.77	0.0255	0.0577	-3.60
5,6,7	0	$k = 5$	0.154	0.262	-2.29	0.00972	0.0268	-4.96	0.0157	0.0316	-4.11
3,5,6,7	0	$k = 5$	0.181	0.267	-2.31	0.00937	0.0265	-4.95	0.0157	0.0334	-4.08
line content		kNN									
3,5,6,7	1	$k = 5$	0.161	0.277	-2.23	0.0108	0.0304	-4.86	0.0185	0.0380	-3.94
3,5,6,7	0,1	$k = 5$	0.176	0.275	-2.26	0.0104	0.0282	-4.91	0.0162	0.0346	-4.03
3,5,6,7	2	$k = 5$	0.164	0.275	-2.24	0.0106	0.0293	-4.88	0.0171	0.0364	-3.98
3,5,6,7	0,2	$k = 5$	0.177	0.273	-2.27	0.0104	0.0275	-4.92	0.0161	0.0341	-4.04
3,5,6,7	0,1,2	$k = 5$	0.169	0.276	-2.25	0.0104	0.0283	-4.91	0.0162	0.0352	-4.01
line ends		RBF-SVM									
3,5,6,7	0	$\gamma = 1$	0.188	0.245	-2.21	0.00899	0.0323	-4.48	0.0349	0.0589	-3.52
3,5,6,7	0	$\gamma = 3$	0.128	0.245	-2.33	0.00655	0.0333	-4.69	0.0206	0.0457	-3.81
3,5,6,7	0	$\gamma = 5$	0.229	0.332	-1.85	0.0229	0.0691	-3.71	0.0468	0.0838	-2.97
line content		RBF-SVM									
3,5,6,7	0,2	$\gamma = 1$	0.195	0.268	-2.04	0.0188	0.0380	-4.14	0.0407	0.0695	-3.23
3,5,6,7	0,2	$\gamma = 3$	0.140	0.226	-2.22	0.00669	0.0307	-4.63	0.0327	0.0447	-3.61
3,5,6,7	0,2	$\gamma = 5$	0.121	0.258	-2.29	0.00851	0.0370	-4.55	0.0211	0.0509	-3.79
3,5,6,7	0,2	$\gamma = 10$	0.282	0.414	-1.58	0.0296	0.0775	-3.43	0.0526	0.0957	-2.80
3,5,6,7	0,1,2	$\gamma = 1$	0.199	0.279	-1.94	0.0270	0.0418	-4.38	0.0387	0.0728	-3.25
3,5,6,7	0,1,2	$\gamma = 3$	0.167	0.220	-2.20	0.00669	0.0316	-4.86	0.0368	0.0508	-3.48
3,5,6,7	0,1,2	$\gamma = 5$	0.123	0.246	-2.28	0.00655	0.0297	-4.88	0.0192	0.0449	-3.81
3,5,6,7	0,1,2	$\gamma = 10$	0.239	0.378	-1.71	0.0181	0.0653	-3.82	0.0436	0.0898	-3.01
the best settings	$T =$ $T_1 =$	method:	3,5,6,7 0,2	3,5,6,7 0,1,2	3,5,6,7 0	1 0	4 0	1 0	1 0	5 0	4 0
line interpretation		method:	RBF	RBF	RBF	kNN	kNN	kNN	kNN	kNN	kNN
usage of the principles			content	content	ends	vector	vector	vector	vector	ends	vector
constant width			+	+	+	-	+	-	-	+	+
continuity (V)			+	+	+	-	-	-	-	+	-
amodal completion (T)			+	+	+	-	-	-	-	-	-
vicinity (X)			+	+	+	-	-	-	-	-	-
similarity			+	+	-	-	-	-	-	-	-

Table 2: Efficiency of detection with various settings and criteria.

The numerical parameters $m = 16$, $k = 5$ (for kNN), $n = 100$, $N = 500$ are fixed for all of them. The reason for the choice is following: $m = 16$ is determined by USPS structure, $N = 500$ allows to get final p -values essentially (several times) below the standard statistical threshold of 1%, $n = 100$ is selected to make the results observable and visualizable, the number of neighbors $k = 5$ was chosen as an appropriate (not too large) one for the size of the training set $n = 100$. The re-scaling option is always on as well, to normalize gestalt profile dimensions.

The bottom part of Tab. 2 includes a summary. For each of the columns, it answers the following questions. First, in what settings the best results are achieved. Second, what kind of line interpretation (vector, ends, content) was deep enough for achieving the best results. Third, which gestalt principles were involved and which of them did not contribute to the best achievement. This allows us to compare them and to sort by their relative importance. Surely, these conclusions are preliminary and bounded by the experiments included in the table. We have mostly concentrated on the improvement of ALPV for the first problem statements, and less focused on the others.

In addition to that, the following conclusions can be made:

- The full combination aggregating all the considered principles at once is a good one but not always the very best.
- The combination $T = 3, 5, 6, 7$ (aggregating the gestalt principles: constant width, continuity, amodal completion, vicinity, but not including similarity as $T_1 = 0$) is the best (-2.33) for the prior ALPV criterion and the first (the most complex) task.
- However, almost the same quality (-2.31) is reachable with just $T = 3, 5$ (constant width, continuity), so these two principles are the most important ones.
- The results improved with replacing k -NN with SVM (RBF kernel). The best result for the ALPV criterion was achieved this way. This algorithm is sensible for the value of the parameter (the best value is about $\gamma = 3$) but more tolerable for over-extensions of the feature set.
- Involving similarity principle is not shown to contribute essentially to ALPV, but it allows us to achieve the best results according to the APV criterion (0.220).
- Statements 2 and 3 are much easier and require less complicated methods. However, they just give the information that an image is not random noise which is not so interesting as a distinction from the mixture images.
- Summarising, all the gestalt principles we tried can contribute to efficiency, but this is confirmed to a different degree. Sorted by priority, they are:
 - (0) point to line (always used as the fundamental platform);
 - (1) constant width (parallelism);
 - (2) continuity (V-junction);
 - (3-4) amodal completion and vicinity (T- and X- junctions);
 - (5) similarity of the content.

5. Conclusion and Future Work

In this work, we have shown how CAD can be used together with gestalt principles of vision (GPV) to separate perceivable and non-perceivable images. This combination of CAD with GPV allows to improve the accuracy of the detection of the images and interprets an image/shape as being perceivable (meaningful) if it sufficiently deviates from randomness - that is the image could hardly happen by chance. Practically, all considered gestalt principles have shown to contribute to improving the accuracy of detection, but to a different degree. The deviations from randomness are measured by using the conformal prediction technique that has a very valuable property of validity. The technique allows us to detect perceivable images and bound the number of false alarms.

Further work may concentrate on the application of complex gestalt principles like convexity, closure, common motion, etc. The principle of symmetry would be useful to apply for more complex (high-level) examples than just the lines. Also, at higher levels, the principles like similarity, vicinity, and constant width may have more interesting meaning

than when they are applied just to the line pairs. Also, the family of the grouping principles can be extended with the elements of the scalar vector graphics when they are relevant for the detection of meaningful elements.

The next step of our work is exploring the anomaly detection as a tool for visual reconstruction. Anomaly Detection itself is interesting for us as far is related on ‘being perceivable’ as the cause of the abnormality, and allows to explain it in terms of most meaningful region/elements of an image.

Acknowledgements

We are very grateful to the Amazon Research Awards (ARA) for their support of our project “Conformal Martingales for Change-Point Detection”. We would also like to thank AstraZeneca for their support of the project entitled “Machine Learning for Chemical Synthesis”. We are indebted to Vladimir Vovk and Lars Carlsson for their encouragement and useful discussions.

References

- A.Desolneux, L.Moisan, and J.-M. Morel. From gestalt theory to image analysis a probabilistic approach. *Springer*, 2008.
- A.Gammerman and V.Vovk. Hedging predictions in machine learning. *The Computer Journal*, 50:151–163, March 2007.
- C.Dawson. Outlier detection with one-class svms. *On-line tutorial*.
- Giovanni Cherubin, Adrian Baldwin, and Jonathan Griffin. Exchangeability martingales for selecting features in anomaly detection. In Alex Gammerman, Vladimir Vovk, Zhiyuan Luo, Evgueni Smirnov, and Ralf Peeters, editors, *Proceedings of the Seventh Workshop on Conformal and Probabilistic Prediction and Applications*, volume 91 of *Proceedings of Machine Learning Research*, pages 157–170. PMLR, 11–13 Jun 2018. URL <http://proceedings.mlr.press/v91/cherubin18a.html>.
- E. H. Gombridge. The story of art. *Phaidon, London*, 1971.
- I.Nouretdinov. Validity and efficiency of conformal anomaly detection on big distributed data. *Advances in Science, Technology and Engineering Systems Journal*, 2:254–267, 3 2007.
- I.Nouretdinov, J.Gammerman, L.Shemilt, and D.Rehal. Multi-level conformal clustering: A distribution-free technique for clustering and anomaly detection. *Neurocomputing*, pages 1–13, 2019. doi: <https://doi.org/10.1016/j.neucom.2019.07.114>.
- J.Smith, I.Nouretdinov, R.Craddock, C.Offer, and A.Gammerman. Conformal anomaly detection of trajectories with a multi-class hierarchy. *Statistical Learning and Data Sciences: Third International Symposium*, pages 281–290, 2015.
- G. Kanizsa. Grammatica del vedere/la grammaire du voir. *Il Mulino, Bologna/Editions Diterot Art et Sciences*, 1997.

- H. Mallinson and A. Gammerman. Imputation using support vector machines. 2003.
- B. Rajaei and R.G. von Gioi. Gestaltic grouping of line segments. *Image Processing On Line*, 2018.
- R. Arnheim. Art and visual perception: A psychology of the creative eye. *Faber & Faber*, 1956.
- R. Laxhammar. Anomaly detection. *Chapter 4 In: Vineeth N. Balasubramanian, Shen-Shyang Ho and Vladimir Vovk (Ed.), Conformal Prediction for Reliable Machine Learning: Theory, Adaptations and Applications*, pages 71–97, 2014.
- V. Vovk, I. Nourtdinov, V. Fedorova, I. Petej, and A. Gammerman. Criteria of efficiency for set-valued classification. *Annals of Mathematics and Artificial Intelligence*, 81:21–46, October 2017.
- V. Vovk, A. Gammerman, and G. Shafer. Algorithmic learning in random world. *Springer*, 2005.

Appendix A. Review of gestalt principles

In Table 3 we collect some of the gestalt grouping principles. We also mention what kind of information is extracted from a line. The principle of *constant width* (parallelism) needs the smallest amount of information, only vectors of differences between the line ends. For *continuity of direction* (V-junction), *amodal completion* (T-junction) and *vicinity* (X-junction)⁴, the location of the lines in the space becomes essential, but it is enough to use the line ends. The *similarity* principle is the only one where we looking at the *content* of the lines, i.e. how the lines are black by the points internally.

Some other gestalt principles are left for future work, as they need at least 3 lines to be applied. When applied just to 2 lines, *convexity* becomes the same as continuity of direction, *closure* is never satisfied, *common motion* is always satisfied.

Our interpretation of the gestalt principles was inspired by [Rajaei and von Gioi. \(2018\)](#). In that work, the continuity level was measured by the angle between the connected lines, through their scalar product. The same can be applied for quantifying parallelism between the lines. Therefore, we assume that the vectors are still ‘parallel’ to some degree if their scalar product is non-zero. This becomes our core way of involving gestalt principles: instead of calculating the number of lines, we consider the overall scalar product of their oriented version. To apply T-, V-, or X-junction, we restrict this summing only to the pairs of lines following one of these relations. Finally, the similarity principle means that the scalar products are additionally multiplied by the proportion of matching the content of these lines. This is how quantification of parallelism and similarity is joined together for the pairs of first-level lines.

4. We apply the vicinity principle in a simplified form, checking whether the (Hausdorff) distance between the example is zero. For the lines, this means that they are intersecting at a common point.

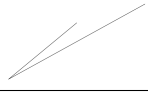
Name	Grouping idea	Applicability
<i>points to line</i>	grouping points on a straight line 	points
<i>color constancy</i>	same color within connected regions (applicable in non-binary grayscale)	1 line (content)
<i>constant width</i> (parallelism)	grouping parallel curves 	2 lines (vectors)
<i>continuity of direction</i> (V-junction)	creating non-linear curves 	2 lines (ends)
<i>amodal completion</i> (T-junction)	applies when a curve stops on another curve, tends to interpret the interrupted curve as the boundary of some example 	2 lines (ends)
<i>vicinity</i> (including X-junction)	distance between examples is small enough with respect to the rest; 	2 lines (ends)
<i>similarity</i>	group similar examples into higher-scale examples 	2 lines (content)
<i>convexity</i>	any convex curve (even if not closed) suggests itself as the boundary of a convex body 	3 lines at least
<i>closure</i>	leads us to see as an example the part of the plane surrounded by a closed contour 	3 lines at least
<i>common motion,</i> or perspective law (Y-junction)	several concurring lines appear in an image, with common meeting point (possibly, in continuation) 	3 lines at least
<i>symmetry</i>	group any set of examples that is symmetric with respect to some straight line 	higher levels
<i>past experience</i>	(applicable w.r. to the training data)	

Table 3: Gestalt principles

Appendix B. Parametrised versions of gestalt-profiles

Table 4 shows how the gestalt profiles are obtained for different values of parameters T and T_1 , with comments which of the gestalt principles (Appendix A) are involved. In accordance to it, we define the function $W_i = W_i(T, T_1, Q, V)$ needed for Algorithm 2. Note that more than one value of T may be applied simultaneously. This would mean that the gestalt-profile is the union of the profiles for the different values of T . The same is true for the second parameter T_1 .

- case $T = 1$, any T_1 (quantity): $W_i := |V|$
- case $T = 2$, any T_1 (length): $W_i := \sum_{v \in V} \langle v, v \rangle$
- case $T = 3$, $T_1 = 0$ (constant width, parallelism):

$$W_i := \sum_{v_1, v_2 \in V} \langle v_1, v_2 \rangle$$

- case $T = 4$, $T_1 = 0$ (variation of case 3): $W_i := \sum_{v_1 \in V} \max_{v_2 \in V} \langle v_1, v_2 \rangle$
- case $T = 5$, $T_1 = 0$ (continuity, V-junction):

$$W_i := \sum_{(p_1, q_1; p_2, q_2; \psi), (p'_1, q'_1; p'_2, q'_2; \psi') \in Q: \exists a \exists b: (p_a, q_a) = (p'_b, q'_b)} (p_2 - p_1)(p'_2 - p'_1) + (q_2 - q_1)(q'_2 - q'_1)$$

- case $T = 6$, $T_1 = 0$ (amodal completion, T-junction):

$$W_i := \sum_{(p_1, q_1; p_2, q_2; \psi), (p'_1, q'_1; p'_2, q'_2; \psi') \in Q: \exists a \exists b \exists c \in (0, 1): (p_a, q_a) = (p'_b, q'_b)^c + (1-c)(p'_{3-b}, q'_{3-b})} (p_2 - p_1)(p'_2 - p'_1) + (q_2 - q_1)(q'_2 - q'_1)$$

- case $T = 7$, $T_1 = 0$ (vicinity, understood as X-junction):

$$W_i := \sum_{(p_1, q_1; p_2, q_2; \psi), (p'_1, q'_1; p'_2, q'_2; \psi') \in Q: \exists c, d \in (0, 1): (p_1, q_1)^c + (1-c)(p_2, q_2) = (p'_1, q'_1)^d + (1-d)(p'_2, q'_2)} (p_2 - p_1)(p'_2 - p'_1) + (q_2 - q_1)(q'_2 - q'_1)$$

- case $T = 5$, $T_1 = 1$ (with similarity, 'one-sided' version; is defined by analogy for $T_1 = 1$ with other $T \geq 3$):

$$W_i := \sum_{(p_1, q_1; p_2, q_2; \psi), (p'_1, q'_1; p'_2, q'_2; \psi') \in Q: \exists a \exists b: (p_a, q_a) = (p'_b, q'_b)} ((p_2 - p_1)(p'_2 - p'_1) + (q_2 - q_1)(q'_2 - q'_1)) \mid \{k : \psi_k = \psi'_k\}$$

T	T_1	the idea how i -th dimension is defined	gestalt principles
1	0	$ V_i $, or the total number of lines in V_i	points to line
2	0	The summarised squared length of the lines in V_i (the length of a line here means the distance between its edges)	points to line
3	0	The summarised pairwise scalar product of the lines.	constant width
4	0	The sum-max scalar product.	constant width
5	0	Same as $T = 3$, restricted to V-junctions, the pairs of lines with a common edge.	constant width + continuity
6	0	Same as $T = 3$, restricted to T-junctions, one of the lines goes through an edge of the other	constant width + amodal
7	0	Same as $T = 3$, restricted to X-junctions, the pairs of intersecting lines.	constant width + vicinity
3	1	The summarised pairwise scalar product of the lines, multiplied by the proportion of content coincidence.	constant width + similarity
3	2	The summarised pairwise scalar product of the lines, multiplied by the proportion of content coincidence (2-sided).	constant width + similarity
5	1	Same as $T = 3$, restricted to V-junctions, the pairs of lines with a common edge, multiplied by the proportion of content coincidence.	constant width + continuity + similarity
5	2	Same as $T = 3$, restricted to V-junctions, the pairs of lines with a common edge, multiplied by the proportion of content coincidence (2-sided).	constant width + continuity + similarity
6	1	Same as $T = 3$, restricted to T-junctions, one of the lines goes through an edge of the other, multiplied by the proportion of content coincidence.	constant width + amodal + similarity
6	2	Same as $T = 3$, restricted to T-junctions, one of the lines goes through an edge of the other, multiplied by the proportion of content coincidence (2-sided).	constant width + amodal + similarity
7	1	Same as $T = 3$, restricted to X-junctions, the pairs of intersecting lines, multiplied by the proportion of content coincidence.	constant width + vicinity + similarity
7	2	Same as $T = 3$, restricted to X-junctions, the pairs of intersecting lines, multiplied by the proportion of content coincidence (2-sided).	constant width + vicinity + similarity

Table 4: Types of gestalt-profiles producible with Algorithm 2.

- case $T = 5, T_1 = 2$ (with similarity, ‘two-sided’ version; is defined by analogy for $T_1 = 2$ with other $T \geq 3$):

$$W_i := \sum_{(p_1, q_1; p_2, q_2; \psi), (p'_1, q'_1; p'_2, q'_2; \psi') \in Q: \exists a \exists b: (p_a, q_a) = (p'_b, q'_b)} ((p_2 - p_1)(p'_2 - p'_1) + (q_2 - q_1)(q'_2 - q'_1)) \max \{ |\{k : \psi_k = \psi'_k\}|, |\{k : \psi_{K+1-k} = \psi'_{K+1-k}\}| \}$$

Appendix C. Examples of predictions

As the example for illustration, we use the settings $T = 3, T_1 = 0$. They are best for ALPV on the first statement amongst the cases when the lines are interpreted as only vectors. The formula of gestalt-profiling is:

$$W_i := \sum_{v_1, v_2 \in V} \langle v_1, v_2 \rangle .$$

Sensitivity check

Fig. 2–4 show the distribution of p -values, and the images sorted by p -values. This is done for three proper tasks when the training is done on a more noisy class than the testing. These figures show which of the meaningful images are easier to be detected (on the left side), and which of them are harder (on the right side). Usually, the problem is caused either by a high amount of the noise, imputed on the stage of equalizing or by circle elements (in such digits as 0,6,9). This can be taken into account in planning the future work: prior attention to convexity, closure, and possibly non-linear circles/arcs.

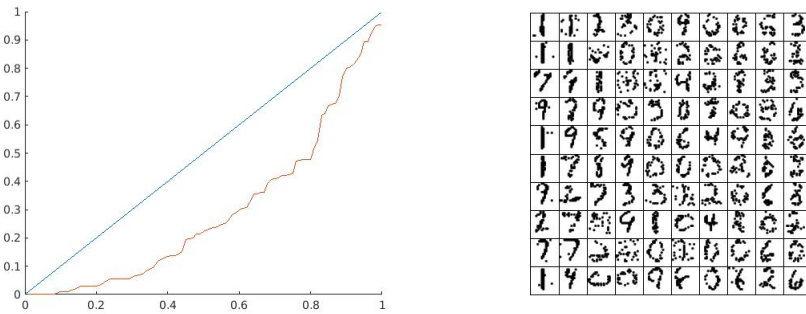


Figure 2: Distribution of p-values (sorted) assigned to real images, trained on mixture images, and real images sorted by ascending p-values (from left to right, from up to down within a column)

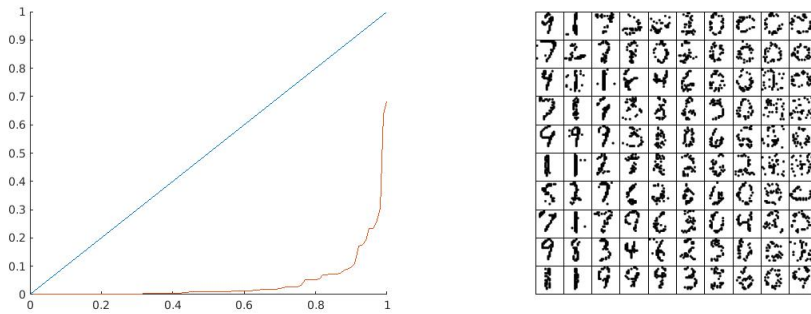


Figure 3: Distribution of p-values (sorted) assigned to real images, trained on random noise, and real images sorted by ascending p-values (from left to right, from up to down within a column)

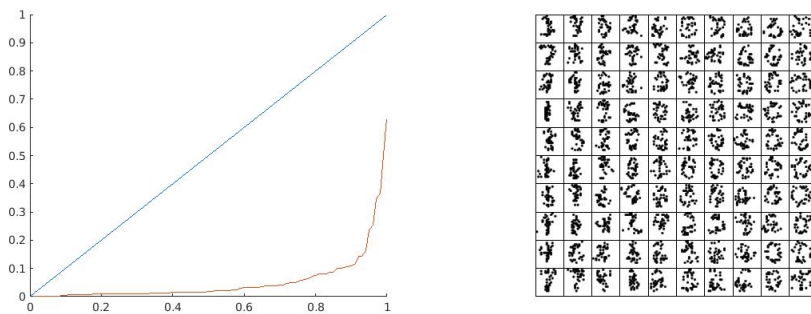


Figure 4: Distribution of p-values (sorted) assigned to mixture images, trained on random noise, and mixture images sorted by ascending p-values (from left to right, from up to down within a column)

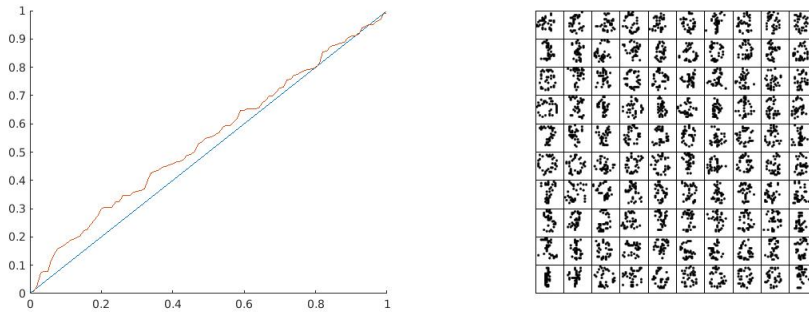


Figure 5: Distribution of p -values (sorted) assigned to mixture images, trained on mixture images, and mixture images sorted by ascending p -values (from left to right, from up to down within a column)

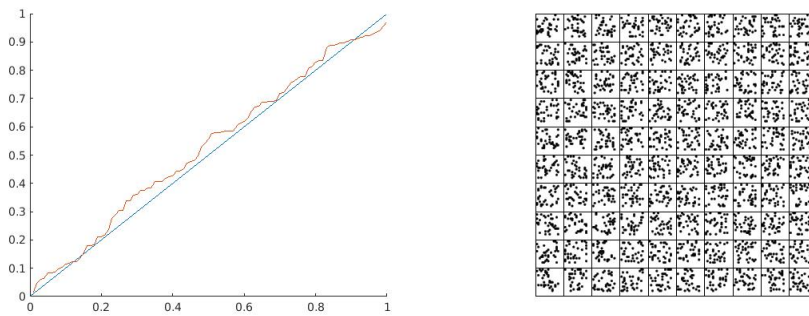


Figure 6: Distribution of p -values (sorted) assigned to random noise, trained on random noise, and random noise images sorted by ascending p -values (from left to right, from up to down within a column)

Control tests

Fig. 5–6 are the control tests, they show the distribution when the training is done on the same class to which it is applied. As expected, the distribution of p -value is close to the uniform, so the required bounds on the false alarm rate are satisfied. The figures on the left sides correspond to cases of ‘false alarms’: the noisy images are suspected to be meaningful because of the occasional elements of the structure visible in them.