1 Introduction

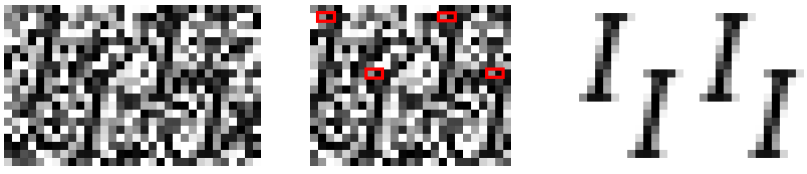
Frequent pattern mining [**?**] is the well-known subfield of data mining that aims to find and extract recurring substructures from data, as a form of knowledge discovery. The generic concept of pattern mining has been instantiated for many different types of patterns, e.g., for item sets (in Boolean transaction data), subgraphs (in graphs/networks), and episodes (in sequences). So far, however, little research has been done on pattern mining for raster-based data, i.e., geometric matrices in which the row and column orders are fixed. The exception is geometric tiling [**?**,**?**], but that problem only considers tiles, i.e., rectangular-shaped patterns, in Boolean data.

In this paper we generalise this setting in two important ways. First, we consider geometric patterns *of any shape* that are geometrically connected, i.e., it must be possible to reach any element from any other element in a pattern by only traversing elements in that pattern. Second, we consider *discrete geometric data* with any number of possible values (which includes the Boolean case). We call the resulting problem *geometric pattern mining*.

Figure 1 illustrates an example of geometric pattern mining. Figure 1a shows a 32×24 grayscale 'geometric matrix', with each element in $[0, 255]$, apparently filled with noise. If we take a closer look at all horizontal pairs of elements, however, we find that the pair $(146, 11)$ is, amongst others, more prevalent than expected from 'random noise' (Figure 1b). If we would continue to try all combinations of elements that 'stand out' from the background noise, we would eventually find four copies of the letter 'I' set in 16 point Garamond Italic (Figure 1c).

The 35 elements that make up a single 'I' in the example form what we call a *geometric pattern*. Since its four occurrences jointly cover a substantial part of the matrix, we could use this pattern to describe the matrix more succinctly than by 768 independent values. That is, we could describe it as the pattern 'I' at locations $(5, 4), (11, 11), (20, 3), (25, 10)$ plus 628 independent values, hereby separating structure from accidental (noise) data. Since the latter description is shorter, we have compressed the data. At the same time we have learned something about the data, namely that it contains four I's. This suggests that we can use compression as a criterion to find patterns that describe the data.

Approach and contributions. Our first contribution is that we introduce and formally define *geometric pattern mining*, i.e., the problem of finding recurring



(a) 32×24 'geometric matrix'. **(b)** Pair $(146, 11)$. **(c)** Pattern 'I' occurs four times.

Fig. 1: Geometric pattern mining example. Each element is in $[0, 255]$.

local structure in geometric, discrete matrices. Although we restrict the scope of this paper to two-dimensional data, the generic concept applies to higher dimensions. Potential applications include the analysis of satellite imagery, texture recognition, and (pattern-based) clustering.

We distinguish three types of geometric patterns: 1) *exact* patterns, which must appear exactly identical in the data to match; 2) *fault-tolerant* patterns, which may have noisy occurrences and are therefore better suited to noisy data; and 3) *transformation-equivalent* patterns, which are identical after some transformation (such as mirror, inverse, rotate, etc.). Each consecutive type makes the problem more expressive and hence more complex. In this initial paper we therefore restrict the scope to the first, exact type.

As many geometric patterns can be found in a typical matrix, it is crucial to find a compact set of patterns that together describe the structure in the data well. We regard this as a model selection problem, where a model is defined by a set of patterns. Following our observation above, that geometric patterns can be used to compress the data, our second contribution is the formalisation of the model selection problem by using the *Minimum Description Length (MDL) principle* [**?**,**?**]. Central to MDL is the notion that 'learning' can be thought of as 'finding regularity' and that regularity itself is a property of data that is exploited by *compressing* said data. This matches very well with the goals of pattern mining, as a result of which the MDL principle has proven very successful for MDL-based pattern mining [**?**,**?**].

Finally, our third contribution is Vouw, a heuristic algorithm for MDL-based geometric pattern mining that (1) finds compact yet descriptive sets of patterns, (2) requires no parameters, and (3) is tolerant to noise in the data (but not in the occurrences of the patterns). We empirically evaluate Vouw on synthetic data and demonstrate that it is able to accurately recover planted patterns.