Minimal 2-dimensional Periodicities and Maximal Space Coverings

Mireille Régnier
INRIA Rocquencourt

May 29, 1995

[summary by Mireille Régnier]

1. Introduction

String searching can be generalized to "multidimensional search" or "multidimensional pattern matching": a multidimensional pattern, p, most often an array and usually connected and convex, is searched in a multidimensional array, the text, t. A strong interest appeared recently [3, 2, 4]. Notably, the duel paradigm improves average and worst-case complexity of pattern matching. Knowing for each position of self-overlap a mismatching position—the *witness*—allows to eliminate one of the two candidates by one question—the *duel*. One studies here pattern periodicities and space coverings. One proposes a period definition valid in any dimension and consistent with the more general definitions in dimension 1, i.e. on words. We prove here that a periodic pattern is generated by a subpattern, and the subpattern, as well as the generating law and the link to the regular distribution of periods, is exhibited. The exceptions to this regularity , the degenerated periods, are interpreted as "border effects". They derive from some regularity of the generating subpattern, a basic phenomenon in dimension 1. This allows for a classification of periodicities valid in any dimension, and detailed in dimension 2. Notably, the number of periodicity classes appear linear in the dimension. Also, one provides a full characterization of sources positions, including the degenerated ones that are essential to the design and correctness of 2D pattern matching algorithms. This considerably refines and achieves the previous classification by [1], and even the extended results in [4], and allows for a classification of space coverings, where non-degenerated periodicities appear essential. One exhibits relationship between the periods of a pattern and the possible space coverings by the same pattern. This is relevant both to the derivation of the theoretical complexity of d-dimensional pattern matching and to algorithmic issues.

The simple remark that the set of invariance vectors almost has a monoïd structure provides the link to the well studied periodic functions in $\mathbb{Z}^d$. Using their properties leads to a great simplification of the proof of previous results in the area. Additionally, it provides tools for a generalization to any dimension. Finally, the paper provides knowledge to derive efficient pattern preprocessing. In particular, the characterization of minimal generating sub-patterns reduces (partially) periodicity and witness computation to well known problems on words. This allows for using the large toolkit of 1D algorithms to determine periodicities. A preliminary version of this work appeared in [6].

2. Formalism

Basic Notations. A d-dimensional pattern p is a d-dimensional array whose values range on some alphabet A. Given a vector $\vec{u}$, we denote $\vec{u}[i]$ or $\vec{u}_i$ its i-th coordinate. Let P be the set of vectors $\vec{u}$ such that $|\vec{u}[i]| \leq l_i$ where l_i is some integer, called the i-th dimension of p.

B	G	c	d	e	f	g	h	a	*b*	c	d	e	f	g	h	a
C	D	k	l	m	n	i	j	k	l	m	n	i	j	k	l	m
g	h	*a*	b	c	d	e	f	g	h	a	*b*	c	d	e	f	g
i	j	k	l	m	n	i	j	k	l	m	n	i	j	k	l	m
e	f	g	h	**a**	b	c	d	e	f	g	h	a	*b*	c	d	e
i	j	k	l	m	n	i	j	k	l	m	n	i	j	*k*	l	m
c	d	e	f	g	h	**a**	b	c	d	e	f	g	h	a	X	c
i	j	k	l	m	n	i	j	k	l	m	n	i	j	k	l	m
a	b	c	d	e	f	g	h	*a*	b	c	d	e	E	g	G	X

FIGURE 1. Radiant biperiodic pattern

DEFINITION 1. Two vectors $\vec{u}$ and $\vec{v}$ are said in the same direction if and only if, for any i: $\vec{u}_i.\vec{v}_i \geq 0$. A vector $\vec{u}$ *dominates* a vector $\vec{v}$ in the same direction if and only if, for any i, $|\vec{v}_i| \leq |\vec{u}_i|$. A vector $\vec{u}$ is minimal if it does not dominate any vector.

We are interested in shifts such that the two copies are consistent in the overlapping area.

DEFINITION 2. A vector $\vec{u}$ is an *invariance* vector for p if and only if, for any $\vec{v} \in P$, one has $p[\vec{v} + \vec{u}] = p[\vec{v}]$. A couple $(\vec{u}, \vec{v})$ of invariance vectors is said an *invariance couple* if and only if $\forall j : \sum_j |u_j| + |v_j| \leq l_j$. It is simple if $\vec{u}$ and $\vec{v}$ are collinear. These invariance vectors are said *simple*. We note I the set of invariance vectors.

3. Main results

Lattice distribution of invariance vectors. If a pattern p admits a non-simple invariance couple, it is said *biperiodic* (see Figure 1). We have:

DEFINITION 3. Given a lattice L with basis $(\vec{u}, \vec{v})$, we denote $FC_{\vec{u},\vec{v}} = \{\lambda\vec{u} + \mu\vec{v}; 0 \leq \lambda, \mu < 1\}$. A S-path is a chain $\vec{w}_1 \ldots \vec{w}_k$ of vectors in p such that, for any i, either $\vec{w}_{i+1} - \vec{w}_i$ or $\vec{w}_i - \vec{w}_{i+1}$ is in S. Given two vectors $(\vec{u}, \vec{v})$, the *free zone* $FZ_{\vec{u},\vec{v}}$ is the set of points $\vec{w}$ in p such that there exists no $(\vec{u}, \vec{v})$-path interior to P to $FC_{\vec{u},\vec{v}}$. The *periodicity domain* is $p - FZ_{\vec{u},\vec{v}}$. The *border* is:

$$B = p - \cup_{\vec{x},\vec{y} \in L^2} \{\vec{w} \mid (\vec{w} + \vec{x}, \vec{w} + \vec{y}) \in p^2, dir(\vec{w}) = dir(\vec{x}) = dir(\vec{y})\}.$$

THEOREM 1. *Let p be biperiodic. For any invariance couple $(\vec{u}, \vec{v})$ exists a lattice L such that:*

$$I \subseteq L \cup B \cup FZ_{\vec{u},\vec{v}}, \tag{1}$$

where B is the border of L. If L admits two simple vectors, then (1) reduces to:

$$I \subseteq L \cup B. \tag{2}$$

$(\vec{u}, \vec{v})$ is said a non-degenerated invariance couple *and L is said a non-degenerated lattice. A pattern admits at most one non-degenerated lattice, called the* canonical lattice *and denoted $L_{\vec{E},\vec{F}}$ where $(\vec{E}, \vec{F})$ is a basis. The invariance vectors in $\tilde{I} = I - B_{\vec{E},\vec{F}}$ are named the non-degenerated invariance vectors. p is said a non-degenerated biperiodic pattern.*

Figure 1 provides an example where $\vec{E} = [4, 4]$ and $\vec{F} = [6, 2]$. It is worth noticing that a basis is not necessarily made of invariance vectors: this is intrinsically 2D. Similar phenomena occur on any set of collinear vectors: e.g. a regular distribution of invariance vectors and a degeneracy paradigm.

Periodicity classification. A pattern is:

(1) *non-periodic:* no invariance couple
(2) *monoperiodic:* exists one simple invariance couple; all invariance couples are simple.
(3) *biperiodic:* exist one non-simple invariance couple. If the associated lattice is non-degenerated, the pattern is said non-degenerated biperiodic. It divides into two subclasses:
   (a) *fundamental biperiodic* or *lattice periodic:* all lattice vectors are invariance vectors.
   (b) *non fundamental biperiodic* or *radiant periodic:* all invariant lattice vectors are in the same direction.

Word properties. It appears from our example kind of a word repetition. In 1D, minimal generators and periods are based on primality notion on words. Extending this *primality* notion to dimension 2. provides an alternative point of view to the characterization of I as a subset of a lattice $L_{(\vec{E},\vec{F})}$ plus its border $B_{(\vec{E},\vec{F})}$. As a major algorithmic consequence, it allows for using 1D algorithms to search for periodicities, hence witnesses. Also, it simplifies the proofs [7].

Definition 4. Let p be a non-degenerated biperiodic pattern. Let $(\vec{E},\vec{F})$ be a fundamental basis such that a fundamental lattice cell $FC_{\vec{E},\vec{F}}$ is in the periodicity domain. Denote i its direction, and j the other direction. Let $\delta = GCD(\vec{E_j},\vec{F_j})$ and $L = \inf\{k \geq 0; k\vec{e_i} \in L_{\vec{E},\vec{F}}\}$. Define for any (λ,μ) in $[0 \ldots L-1] \times [0 \ldots \delta - 1]$, $\vec{w}_{\lambda,\mu}$ as the only vector in $FC_{\vec{E},\vec{F}}$ such that:

$$\vec{w}_{\lambda,\mu} - (\lambda\vec{e_i} + \mu\vec{e_j}) \in L_{\vec{E},\vec{F}}.$$

Let $p_{\lambda,\mu}$ be $p[\vec{w}_{\lambda,\mu}]$; let s_μ be the primitive word associated to the word $p_{0,\mu-1} \ldots p_{L-1,\mu-1}$. The sequence $(s_\mu)_{1\leq\mu\leq\delta}$ is the *linear canonical generator* in direction i.

Remark that the existence and uniqueness of $\vec{w}_{\lambda,\mu}$ is a direct consequence of Euclid's theorem and that $(\vec{E},\vec{F})$-periodicity implies that (s_i) is independent of the fundamental basis chosen.

Theorem 2. *Let p be a non-degenerated biperiodic pattern, and $(s_i)_{1\leq i<\delta}$ be the associated linear generator. Then, any vector $\vec{w}$ in the periodicity domain and in the same direction satisfies:*

$$p[\vec{w}] = s_\mu(\lambda \bmod |s_\mu|) \tag{3}$$

where (λ,μ) is defined by the equation $\vec{w} - \vec{w}_{\lambda,\mu} \in L_{\vec{E},\vec{F}}$. One has $L = GCM(|s_\mu|) = \frac{|FC_{\vec{E},\vec{F}}|}{\delta}$.

Intuitively, a biperiodic pattern p is made of δ patterns that repeat indefinitely, except maybe for the borders: rows (or columns) i, $i \in \{1 \ldots \delta\}$ are linear concatenations of strings s_i^* and row $j+\delta$ is equal to row j shifted by some value α. In Figure 1, we have $\delta_1 = \delta_2 = 2$ and $s_1 = abcdefgh$ and $s_2 = ijklmn$.

Position of sources. . We remark that (3) holds for any $\vec{w}$ if p is fundamental biperiodic. We show that if a vector $\vec{w}$ in $L_{\vec{E},\vec{F}} \cap T$ is not an invariance vector then $P_{\vec{w}}$ contains a point that *violates $(\vec{E},\vec{F})$ periodicity*: capital characters in Figure 1. Extremal such points, $[15,2], [16,0]$ and $[1,8]$, lead to the exclusion of $[8,0], [6,2]$ and $[0,8]$ from I (represented by bolded **a**).

Maximal Coverings Classification. One proves that two copies of p shifted by $\vec{u}$ and $\vec{v}$ are mutually consistent if and only if $\vec{u} - \vec{v}$ is an invariance vector or $p_{\vec{u}} \cap p_{\vec{v}} = \emptyset$. One defines a $(\vec{u},\vec{v})$-lattice covering as a set of interleaved $\vec{u}$-overlapping sequences where two neighbouring sequences are shifted by $\vec{v}$. It is *regular* if $\vec{u} + \vec{v} \in T$, else it is said *extended.* It steadily follows:

Theorem 3. *A maximal covering of the 2-dimensional space by a pattern p is either of the three:*

(1) *tiling,*

(2) *a tiling of $\vec{u}$-overlapping sequences where $\vec{u}$ is a minimal invariance vector.*

(3) *a $(\vec{u},\vec{v})$-lattice coverings. It is* regular *and $(\vec{u},\vec{v})$ is a basis of the canonical lattice if p is biperiodic; otherwise it is extended.*

Remark that extended lattice coverings are an extension of the covering notion, where some "holes" appear in the representation. This is pertinent for algorithmic issues as it allows to determine the *maximum* number of occurrences of a given pattern, a parameter related to the worst-case complexity.

4. Hints for the proofs

One shows that the sum of invariance vectors is an invariance vector *almost everywhere*, and characterizes this zone of non-invariance, the free zone, that creates "border effects". This additive property allows to use general results on biperiodic functions on Z^2 and prove a lattice distribution of almost all invariance vectors. Notice this vectorial approach provides a very short proof of the previous results in [1, 4]. Many proofs rely on the Factorisation Theorem [5]: equation $ab = ba$ implies that a and b are powers of a same primitive word. For example, in Theorem 2, equation (3) implies that, for any μ, $|s_\mu|$ divides L. Otherwise, for some j, one has $L \bmod |s_j| = \alpha \neq 0$. With $a = s_j[1]\dots s_j[\alpha]$ and $b = s_j[\alpha+1]\dots s_j[|s_j|]$, s_j factors as $s_j = ab = ba$ which contradicts the primitivity. Hence, $GCM(|s_j|)$ divides L. Also, (3) implies that $GCM(|s_j|)\vec{e_i}$ is a lattice vector, hence $L\vec{e_i}$, by the minimality property.

A major consequence of these word properties is the possibility to compute the linear generator, hence the fundamental basis, from any fundamental parallelogram. One initially computes δ as $GCD(\vec{u}_j, \vec{v}_j)$ and L as $\inf\{k; k\vec{e_i} \in L_{\vec{u},\vec{v}}\}$. For each of the δ sequences $p_{\lambda,\mu}$ defined, one can extract the associated primitive word s_j. One may use the well known 1D algorithm that searches for the primitive seed of a word (for instance the preprocessing of Knuth-Morris-Pratt). Then, one can compute all witnesses between two sequences s_j and s_k. This determines whether the set is cyclic (not minimal) and $(\vec{E},\vec{F})$ steadily follows. An implementation and other applications are described in [7].

Bibliography

[1] Amir (A.) and Benson (G.). – Two-dimensional periodicity and its application. In *SODA'92*. – 1992. Proceedings of the 3rd Symposium on Discrete Algorithms, Orlando, FL.

[2] Amir (A.), Benson (G.), and Farach (M.). – Alphabet independent two dimensional matching. In *STOC'92*, pp. 59–67. – 1992. Victoria, BC.

[3] Baeza-Yates (R.) and Régnier (M.). – Fast algorithms for two dimensional and multiple pattern matching. *Information Processing Letters*, vol. 45, n° 1, 1993, pp. 51–57.

[4] Galil (Z.) and Park (K.). – Truly alphabet independent two-dimensional pattern matching. In *FOCS'92*. – IEEE, 1992. Proceedings of the 33rd IEEE Conference on Foundations of Computer Science, Pittsburgh, USA.

[5] Lothaire (M.). – *Combinatorics on Words*. – Addison-Wesley, 1983, *Encyclopedia of Mathematics and its Applications*, vol. 17.

[6] Régnier (M.) and Rostami (L.). – A unifying look at *d*-dimensional periodicities and space coverings. In *CPM'93. Lecture Notes in Computer Science*, vol. 684, pp. 215–227. – Springer-Verlag, 1993. Proceedings of the 4th Symposium on Combinatorial Pattern Matching, Padova, Italy.

[7] Régnier (M.) and Rostami (L.). – Minimal d-dimensional and Maximal Space Coverings, 1995.