GRAMMAR APPROACH TO PRINTED NOTES RECOGNITION*

Schlesinger M.I.† *Savchynskyy B.D.*‡ *Anochina M.A.*
International Research and Training Center UNESCO
for Informational Technologies and Systems.
Acad. Glushkov str. 40, Kiev.

Abstract. Structural approach to printed notes recognition is considered. Two-dimensional generalizations of context-free grammars have been used as a base of recognition algorithms. Special subclass named fixed nonterminals' size grammars of two-dimensional context-free grammars is defined and described. Effective algorithm for images described with the subclass' grammars analysis is presented.

1 Introduction

Printed notes recognition in addition to it's doubtless cultural importance has wide range of use. Typical uses are creation of electronic musical libraries, automation of mechanical work in such processes as rewriting of score to separate parts, musical composition's transposition to another tonalities and many other practical works.

An interest to the problem of printed notes recognition is stimulated also by it's internal scientific content. An image of musical scores is a striking example of objects with difficult, but completely defined internal structure, what gives a possibility to use the newest methods of structural recognition. Well-known works [1, 2, 3] are devoted to theoretical and applied problems of music scores recognition. More complete bibliography is cited in [4].

This work includes five sections. The second section describes main ideas of a base for developed recognition algorithms. The third and the fourth sections are devoted to formalization of these ideas. Two-dimensional generalizations of context-free languages and grammars by Chomsky serve us a base of formalization. Results of experimental testing are presented in the final section.

2 Informal description of the structure of music score image

Processes forming the technology of recognition are based on solution of such tasks.

Let X be a set of all possible images, E – its known subset, considered later as the set of certain ideal, etalon images; $f : E \times X \mapsto R$ is given function such, that $f(e, x)$ characterizes difference between real image x and ideal one e. Current task is to build for given set E and function $f : E \times X \mapsto R$ an algorithm, which points out for every input image x such etalon image e^* from the set E which is differ minimally from x:

$$e^* = \arg\min_{e \in E} f(e, x). \tag{1}$$

We will describe the set E of ideal images of music scores. It's appropriate to describe this set using imaginary process of drawing of such images. This is well-known and prevalent in structural recognition way when the set of objects is defined by generative model (see [5]).

The first stage lies in sequential subdividing of clear page to horizontal stripes of two types. Stripes of the first type define places at a paper where music lines will be placed. Stripes of the second type are the intervals between lines. Thus, certain sketch of future image is the result of the first stage. Stripes of the second type determine the part of an image, which will not be changed later, whereas stripes of the first type determine those places, which will be drawn in at the next stages.

*The work was supported with Ukrainian State Program "Pattern computer" and with DAAD German State Program.
†schles@image.kiev.ua
‡bogdan@image.kiev.ua

At the second stage every stripe of the first type built at the first stage is to be processed. Firstly an image of five lines of the staff is to be drawn in every stripe. Further processing of the stripe consists in its subdividing along its length in horizontal direction to sequence of rectangles that are snug against one another. These rectangles are also subdivided into two types. Rectangles of the first type are labelled in some way to determine that an image inside them will be created at the third stage, whereas images that will not be changed later are drawn inside rectangles of the second type.

Images of elementary musical symbols that are solid and do not consist of more simple ones are created in rectangles of the second type. These are such symbols as clefs, pauses, bar lines and so on. The set of such symbols contains also a special symbol which denotes an interval between proper musical symbols. That's why we consider that rectangles into which the note line is subdivided are snug against one another.

Rectangles of the first type are destined for musical symbols that are made of more simple, elementary ones. These are chords represented by a vertical sequence of notes. The number of different chords is huge in comparison with the number of simple musical symbols considered in the previous paragraph.

The third stage lies in generating images of chords inside rectangles of the first type that have been created at the second stage. Chord is regarded as a complex object that can be made in accordance with certain rules from such elementary musical symbols as note heads (of two types: black and white), stems, flags designating duration of chords and so on.

Described process of note score's images production illustrates the way the set E of ideal images as the base of formal task (1) is defined.

If the problem were to recognize only ideal images, we could speak about recovering of such sequence of operations which would produce an image that should be recognized. Really, knowing this sequence it's not difficult to find a sequence of chords as a set of tones sounding simultaneously and duration of every chord. But because of unavoidable imperfection of real images their recognition requires solution of optimization task (1). It means it's necessary to find such sequence of operations, that results in creation of an image that minimally differs from the input one.

Sequential nature of image production makes creation of such recognition algorithm natural, which sequentially reproduces stages of image production in the same order, as an image was produced. It means, that at the beginning at the base of reasonable heuristic considerations placement of separate note lines is determined. Then position of separate musical symbols including chords at every obtained line is determined. Finally, every chord is analyzed with a purpose to obtain tones it consist of and their duration. This is well-known approach (see, for example, [1, 3, 4]) but it has week sides. They lie in fact that every stage of recognition ends with a final decision. This decision can be wrong but can not be corrected at the next stages. Moreover, wrong decision at any stage inevitably results in mistakes at the next stages.

An algorithm based on exact solution of the task (1) has no such drawbacks. In spite of the fact that size of the set E increases exponentially with the length of musical composition the task (1) can be solved exactly without review of all images in this set. Such solution of the task (1) is based on modern methods of structural recognition (see [5]). Due to sequential nature of ideal images generation they can be viewed as sentences in certain formal languages determined by constructions like context-free grammars by Chomsky [6]. But specific character of our applied task lies in the fact that musical scores do not form a sequence arranged in one direction. Elementary musical symbols are situated one relative to another as in horizontal, as in vertical direction. Structural analysis of such complex formations calls for using more general constructions than context-free languages and grammars by Chomsky. These constructions are so-called two-dimensional context-free grammars and languages, defined and investigated in [5]. Also an algorithm for task (1) solution in the case, when E is two-dimensional context-free language is described there. Investigation of the specific character of the music scores images as the objects for machine analysis gives a possibility to build such an algorithm for their recognition that is much more effective than a general one.

This effective algorithm have been used for music scores recognition. At the next two sections we describe this algorithm together with the grammatical constructions it is based on.

3 Formal description of the structure of music score image

Let I and J be fixed natural numbers. For the rectangle subset of two-dimensional integer grid

$$T(I,J) = \{(i,j) \mid i = 0 \ldots I-1,\ j = 0 \ldots J-1\}.$$

we will give the name **the field of view**. Elements of the field of view are called pixels. Parameters I and J will be called height and width of the field of view respectively.

The set $U = \{0, 1\}$ is called **the set of signals.** In our case value "zero" of the signal corresponds to a white pixel and value "one" – to a black pixel.

Function of the form $x : T \to U$ will be called later **an image**. The set of all possible images is denoted as U^T. Names height and width of image x denote height and width of the field of view image x is defined. These values are denoted later as $h(x)$ and $l(x)$ respectively.

We are interested in images of two types. First type includes images of the musical score page as a whole, second one – images of certain musical symbols. Images of the second type will be called **templates** or **etalon images** taking into account that in process of recognition they determine etalon view of musical symbols. Etalon images are defined at fields of view of smaller height and width than the height and width of the field of view of the whole musical score page.

Rectangular subset of the field of view

$$\Pi_{h,l}^{(i_1,j_1)} = \{(i,j) \mid i_1 \le i < i_1 + h,\ j_1 \le j < j_1 + l,\\ i_1, j_1, h, l \ge 0,\ i_1 + h \le I,\ j_1 + l \le J\}$$

is called **a fragment of the field of view.** Value of the parameter h is called height and l – width of the fragment $\Pi_{h,l}$.

Restriction of the image x on the fragment Π of the field of view will be called **image fragment** and will be denoted $x(\Pi)$. Names height and width of image fragment $x(\Pi)$ denote height and width of the fragment of the field of view image x is defined. These values are denoted later as $h(x(\Pi))$ and $l(x(\Pi))$ respectively.

Two-dimensional context free grammar G is a five-tuple $<V, K, P, w, \varepsilon>$, where V is the set of terminals, K – the set of nonterminals (nonterminal symbols), P – the set of derivation rules, w – penalty function, ε – an axiom.

Terminals (elements of the set V) are etalon images (templates) of symbols and elements of musical scores. We do not call them "terminal symbols" in purpose, as it is usually be doing at the theory of formal languages, because they are not symbols in usual meaning of this word. But in all other respects they are analogous to terminal symbols.

Function w is defined at pares "template - image fragment of the same size" and possesses the value at the set of real numbers. Value $w(v, x(\Pi))$ determines difference between template $v \in V$ and image fragment $x(\Pi)$. We will assume also that if they coincide, then $w(v, x(\Pi)) = 0$.

Nonterminals (elements of the set K) are used inside processes of image analysis and recognition for image fragments naming.

The set of derivation rules P contains **rules of horizontal and vertical concatenation** and **substitution rules**. Rules of horizontal and vertical concatenation have the form:

$$A \mapsto B|C,\ A \mapsto \frac{B}{C},\ A, B, C \in K.$$

Vertical ($|$) and horizontal ($-$) dashes are used here to denote the direction of concatenation but not for regular expressions recording, as in theory of formal languages.

Substitution rules have the form:

$$A \mapsto b,\ A \in K,\ b \in V.$$

Comma is used for short notation. For example, notation

$$A \mapsto B, C|D$$

defines two rules

$$A \mapsto B|D \quad \text{and} \quad A \mapsto C|D$$

at once.

For a given image x application of the rule $A \mapsto b$ to its fragment $x(\Pi)$ demands coinciding of fragment $x(\Pi)$ and template b sizes (heights and widths) and means assignment to fragment $x(\Pi)$ label A and penalty $w(b, x(\Pi))$ on this condition.

Application of the rule $A \mapsto B|C$ to the fragment $x(\Pi)$ demands fulfilment of such condition: fragment $x(\Pi)$ can be subdivided in horizontal direction to two fragments $x(\Pi_1)$ and $x(\Pi_2)$ when the left fragment

$x(\Pi_1)$ has already label B and the right one $x(\Pi_2)$ has label C. If this condition is fulfilled, application of the rule means assignment to fragment $x(\Pi)$ certain penalty and label A. Value of penalty is a sum of two summands. The first summand is equal to fragment's $x(\Pi_1)$ (having label B) penalty, and the second one is equal to the penalty of fragment $x(\Pi_2)$ marked by label C.

Application of other rules is quite analogous: it is necessary to replace word "horizontal" by the word "vertical" and words "left/right" by "upper/down" in previous paragraph.

An image x belongs to the language of the grammar G, if such sequence of grammar rules' usage exists that results in assigning of label ε to the whole image (as to its trivial fragment). This sequence of rules have not contain all rules of the set P and some rules can occur in it more than once.

Mentioned sequence is called **derivation** of the image x in grammar G. The process of this sequence's use is called derivation process.

In image x derivation process at the stage of assigning the label ε to the image x (as to its trivial fragment) the penalty is assigned in accordance with rules defined before. Penalty value is equal to difference between x and ideal image, determined by the sequence of rules in derivation of x. **The best derivation** is such derivation that minimizes resulting penalty.

Let's return to the task (1) in the light of introduced definitions. The set E of ideal images consists of such images, which derivation penalty in grammar G is equal to zero. Function w determines penalty for any image derivation so it determines also value of function f. Task (1) of searching for the most similar image from the set E takes a form:

Task 3.1 *Input image x and grammar G are given. It's necessary to find the best derivation of x in G.*

Obtained as a result of task solution sequence of derivation rules determines the most similar to x image $e^* \in E$.

An example of the grammar, which determines certain, extremely simple set of music score pages is presented at fig. 1. This example is presented only as illustration of the main idea of such grammar construction.

Set V (not presented at fig. 1) consists of etalon images of music score's symbols and elements and images of empty, "white" rectangles. Set V contains all such etalon images which are present in notation of the set P rules.

Set of rules P contains rules of three groups. The first group determines the structure of music page as a whole as the sequence of music lines, the second group determines the structure of separate music line as a sequence of music symbols and finally the third group – structure of the chord as a sequence of whole notes and empty spaces between them.

Set K contains all nonterminals used in rules of the set P.

Function w (also not presented at fig. 1) is defined in such a way, that the value $w(v, x(\Pi))$ is equal to a number of different pixels in template v and fragment $x(\Pi)$.

As it has been already said, there is a general algorithm for the task 3.1 solution for any two-dimensional context-free grammar. This algorithm can be used for our music score images grammar also. Algorithm needs $O(I^2J^2(I+J))$ time for the input image derivation process. That time is quite tangible for a user. But some specific features of our "music" grammar give us a possibility to construct an algorithm demanding much less $O(IJ(I+J))$ time. Algorithm we are speaking about can be applied not only to our music score images grammar but to wide class of context-free grammars. This class, named later a class of **fixed nonterminals' size grammars** and algorithm of image derivation in grammars of this class are described in the next section.

4 Definition of fixed nonterminals' size grammars

Lemma 4.1 *(Sufficient condition for nonterminal's fixed height) Let $G=<V,K,P,w,\varepsilon>$ – two-dimensional context-free grammar. If nonterminal $A \in K$ satisfies conditions 1 and 2 of this lemma then it can be assigned to fragments of only one, fixed height in the process of derivation of any input image in grammar G.*

1. *Substitution rules, containing at their left side nonterminal A contain terminals of equal heights at their right side:*

$$\Big(\big((A \mapsto a_i, a_i \in V) \in P\big) \& \big((A \mapsto a_j, a_j \in V) \in P\big)\Big) \Rightarrow \Big(h(a_i) = h(a_j)\Big)$$