Syntactic pattern recognition

Mateusz Bułat • 20.04.2023

Definition

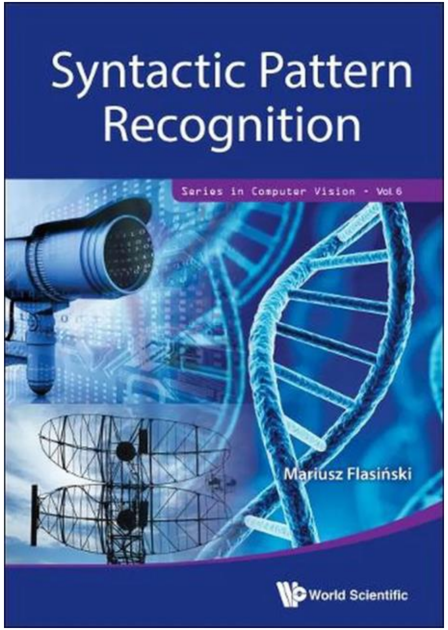
Syntactic pattern recognition is representing pattern as a structure of the form of string, tree or graph and a set of structures as a formal language.

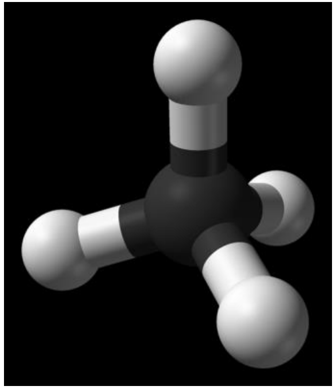
A generation of such a language is made with a formal grammar.

Syntactic Pattern Recognition

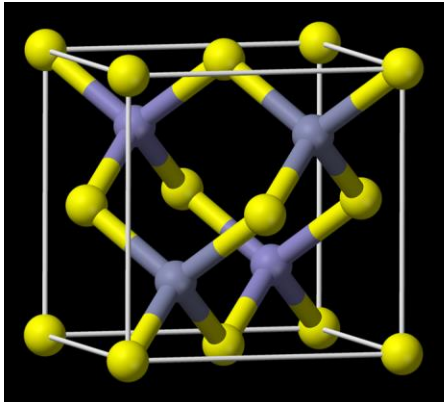
Determining when to use SPR:

1. Objects are distinguishable basing on their structure
2. Structural patterns can be reused
3. Hierarchy-oriented multilevel recognition is possible
4. We require structure-based interpretation





CH4 molecule

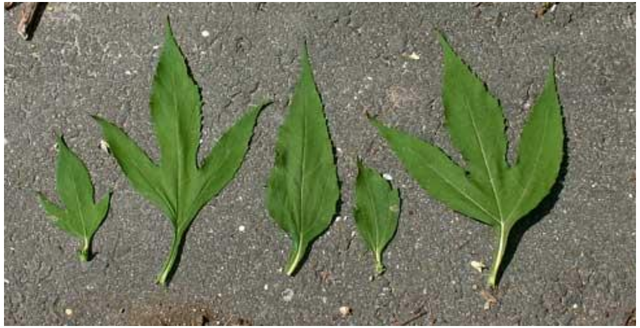


Zinc Sulphide (Sphalerite)

Structure - property of natural objects and artifacts

Examples:

- atoms
- molecules
- all biological and biochemical compounds
- plants
- animals
- artifacts (cars, houses, computational devices)
- languages

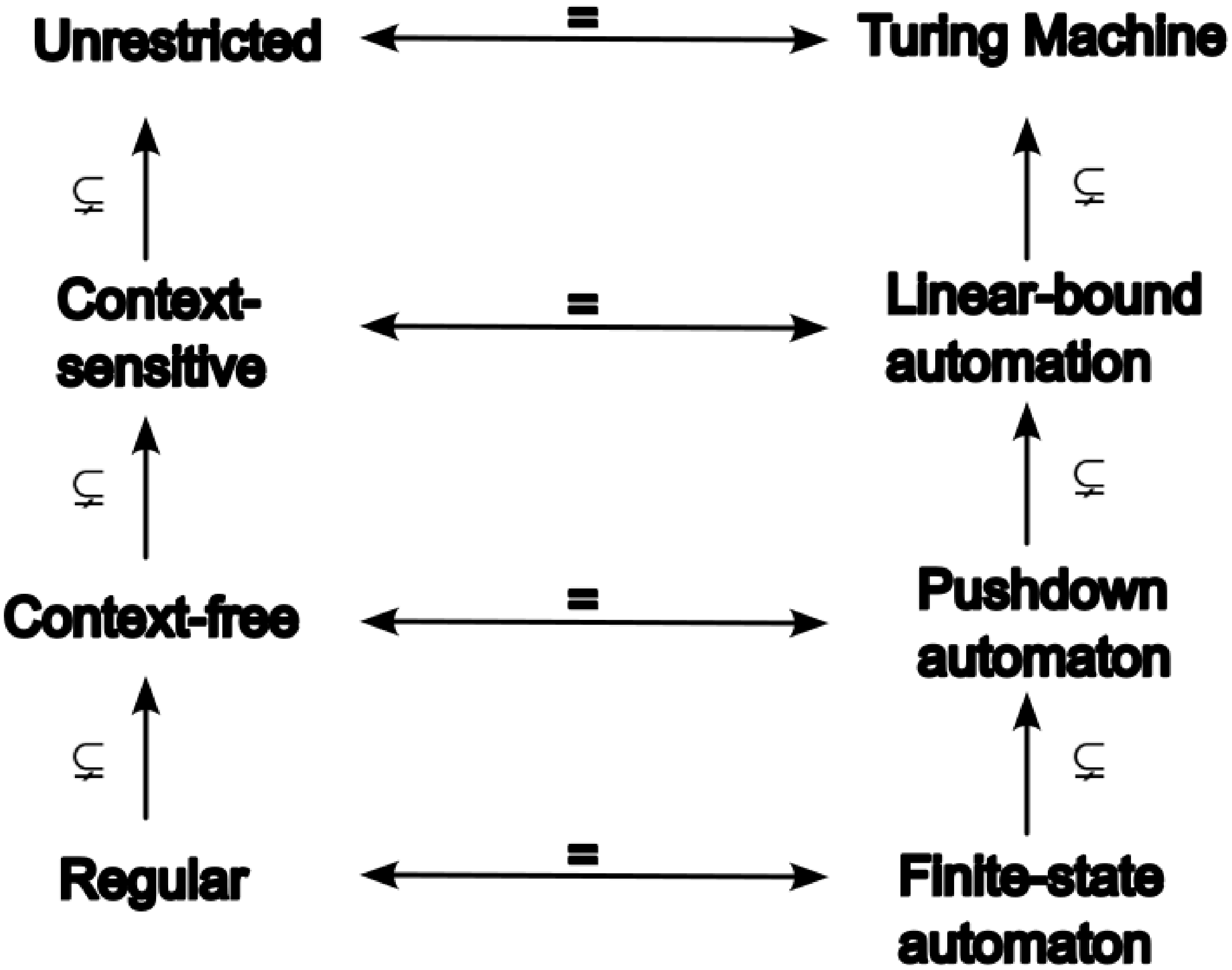


Ragweed leaves

Formal grammars

What do they do and how do they work

- they express how to form “words” or strings belonging to a formal language
- they are a set of production rules
- therefore they can generate a language from mentioned rules
- it’s possible to form an automata that can recognize if given word belongs to the formal language


Unrestricted
=
Turing Machine
⊊
⊊
Context-sensitive
=
Linear-bound automation
⊊
⊊
Context-free
=
Pushdown automaton
⊊
⊊
Regular
=
Finite-state automaton

Common challenges of SPR research

- effects are strongly correlated with proper primitive selection;
- inferring a formal grammar is not an easy task for real life examples;
- having both good computational efficiency and huge discriminative power is not easy to achieve

This is why we “enhance” certain formal grammars — to keep good computational efficiency but increase discriminative power.

Context-free languages

$$Context-free\ grammar\ G\ is\ quadruple\ such\ as:\ G = \left(\sum_N, \sum_T, P, S\right)$$

$$\sum_N are\ nonterminal\ symbols,\ \sum_T are\ terminal\ symbols$$

$$\sum = \sum_N \cup \sum_T$$

$$P\ is\ a\ set\ of\ productions\ of\ the\ form:\ A \rightarrow \gamma$$

$$where\ A \in \sum_N,\ \gamma \in \sum^*$$

$$S\ is\ a\ start\ symbol,\ S \in \sum_N$$

$$We\ assume\ that\ \sum_N \cap \sum_T = \varnothing$$

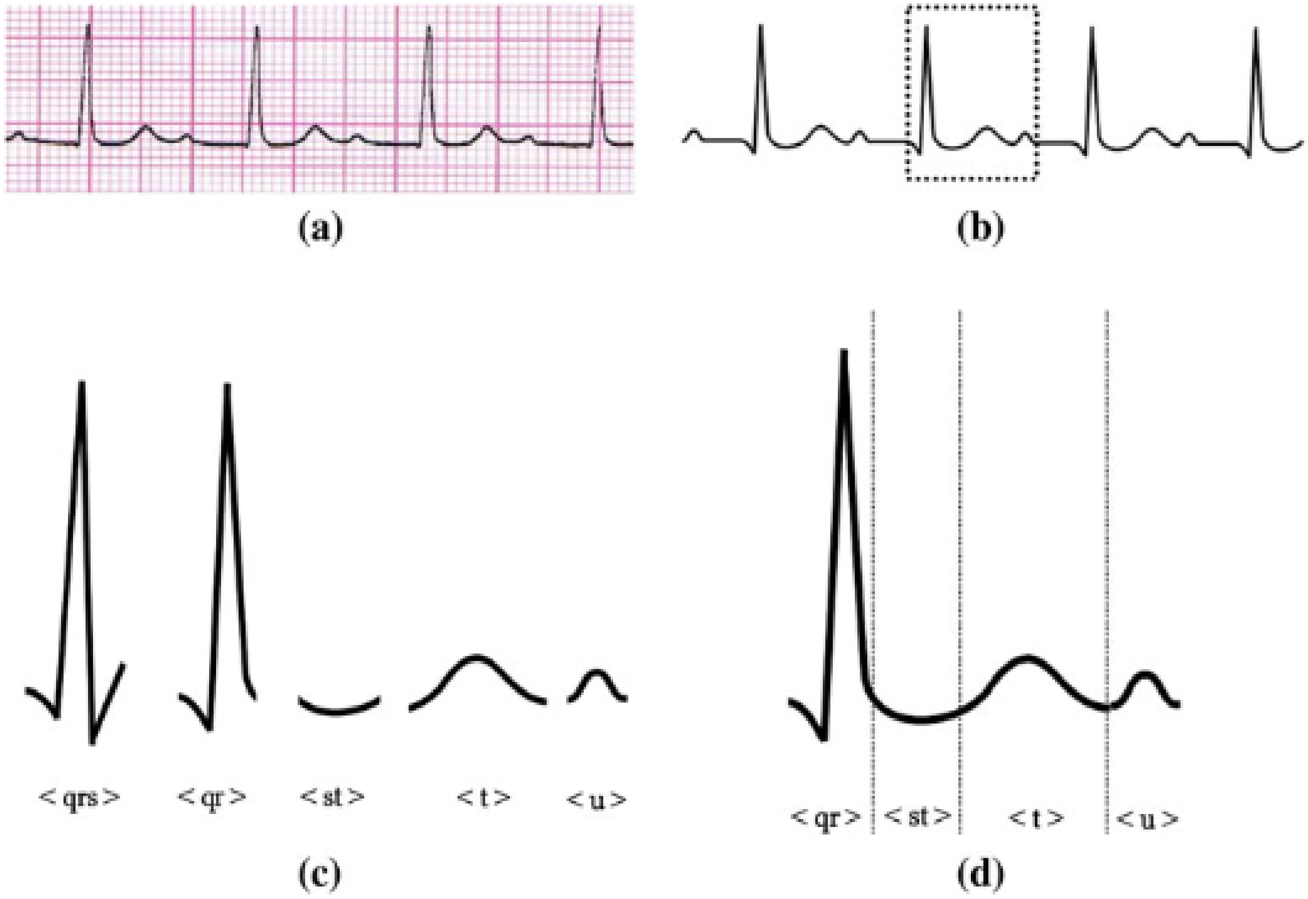
Easy example of context-free grammar

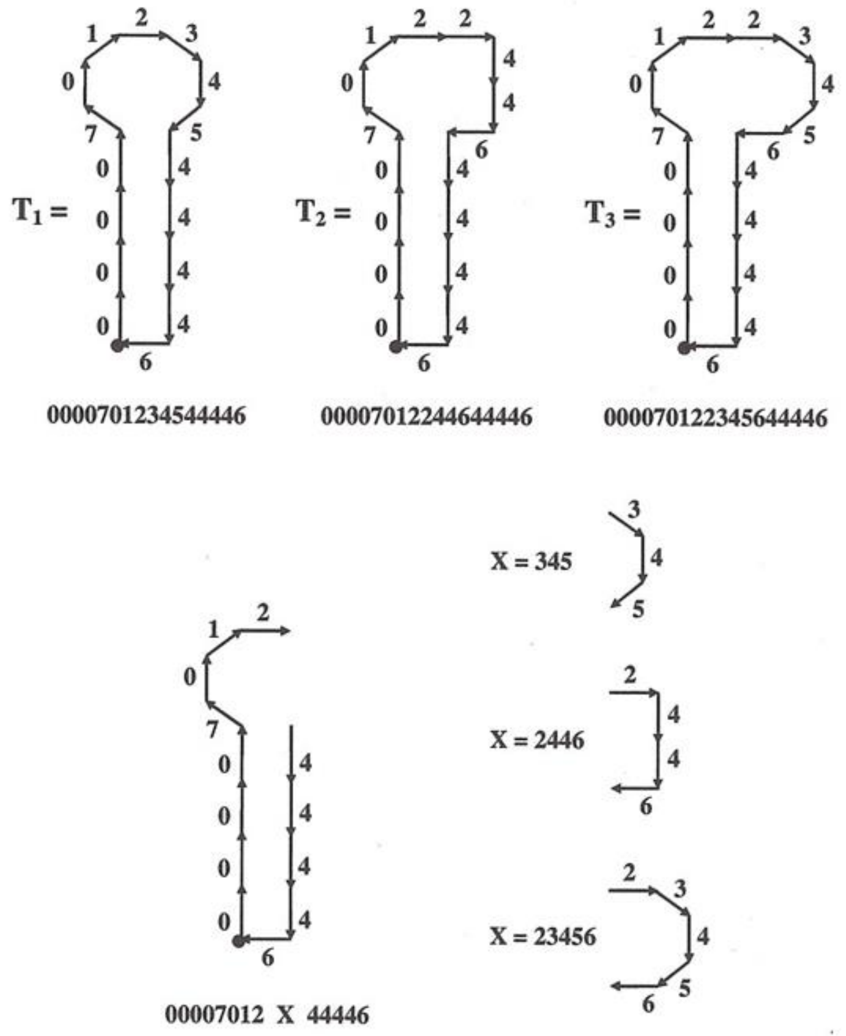
Reminder: non-terminal symbols (such as S) don’t exist in final words. Terminal symbols are building blocks for formal language strings.

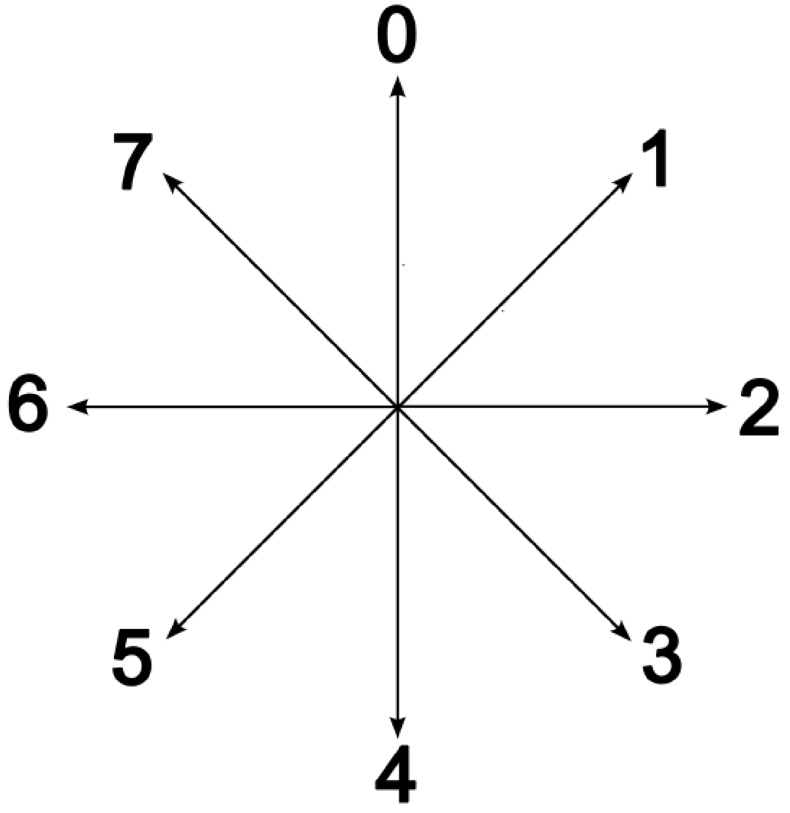
Terminals correspond to primitives identified in real world structures.

$$Context-free\ grammar:$$
$$S \rightarrow aSb,$$
$$S \rightarrow \varepsilon,$$
$$It's\ easy\ to\ see\ those\ rules\ generate\ \left\{ a^n b^n :\ n\ \geqslant\ 0 \right\}\ language$$
$$example\ strings\ are:\ ab,\ aabb,\ aaaabbbb$$

Fig. 2 Phases of image processing in a syntactic pattern recognition system (ECG): **a** an input image, **b** an image after an image processing phase, **c** examples of ECG primitives (elementary patterns), **d** a structural representation after a segmentation




T1 =
T2 =
T3 =
0000701234544446
00007012244644446
00007012234564446
X = 345
X = 2446
X = 23456
00007012 X 44446


0
1
2
3
4
5
6
7

List of popular use-cases

- Scene analysis
- Picture and diagram analysis
- Feature recognition for computer aided design and manufacturing
- Analysis of visual events and activities
- Structure analysis in chemistry
- Optical character recognition
- Structure analysis for process monitoring and control
- Structure analysis in bioinformatics and medicine

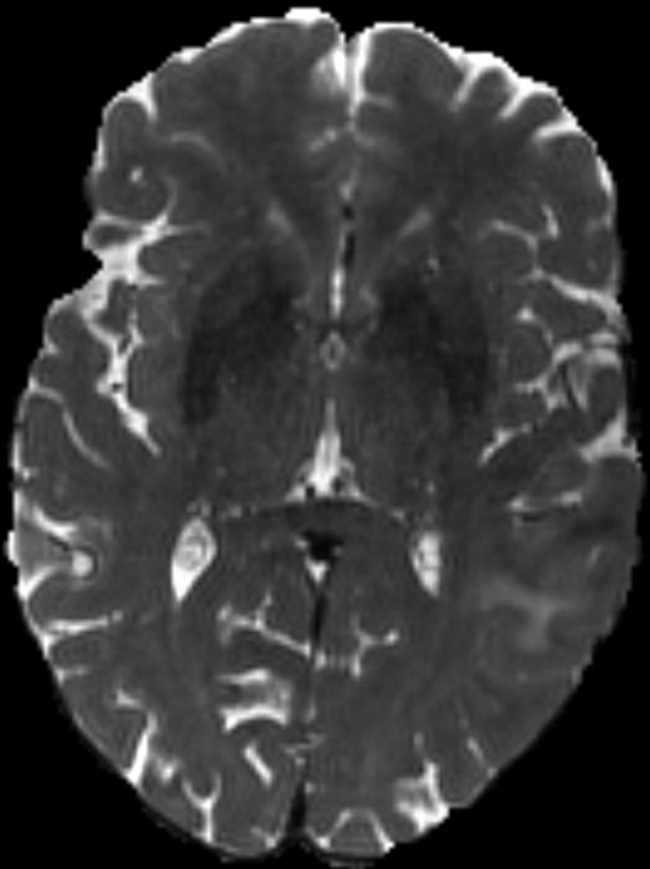
Current research

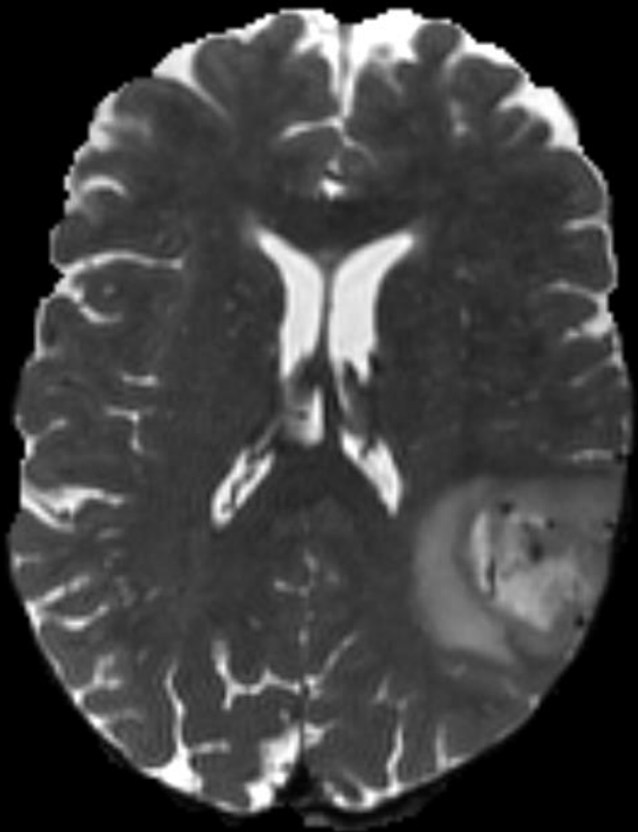
Medical image analysis

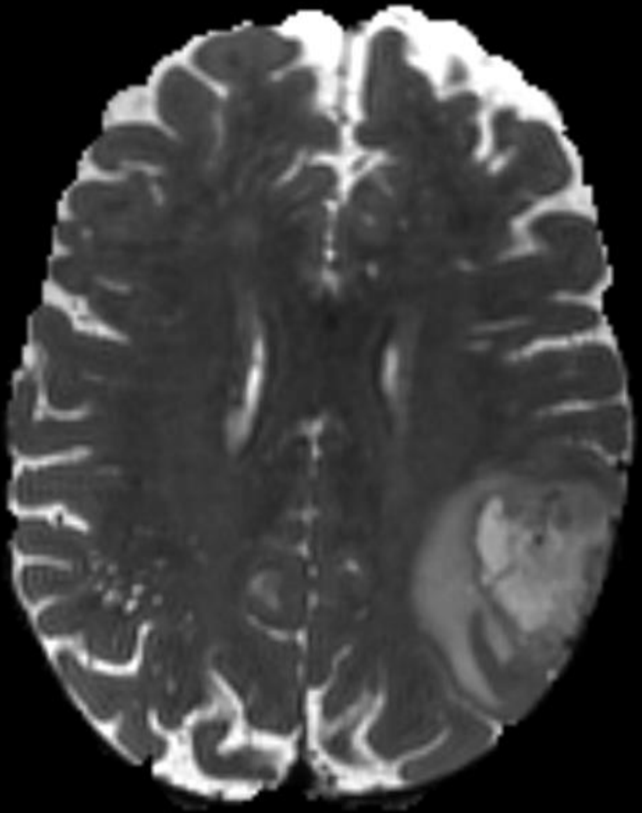
Data we have

It consists of series of MR and CT images. CT images have the damaged parts contoured out. It allows us to extract information specific to damaged regions.

Goal is to accurately point out damaged parts of the brain and contour them.





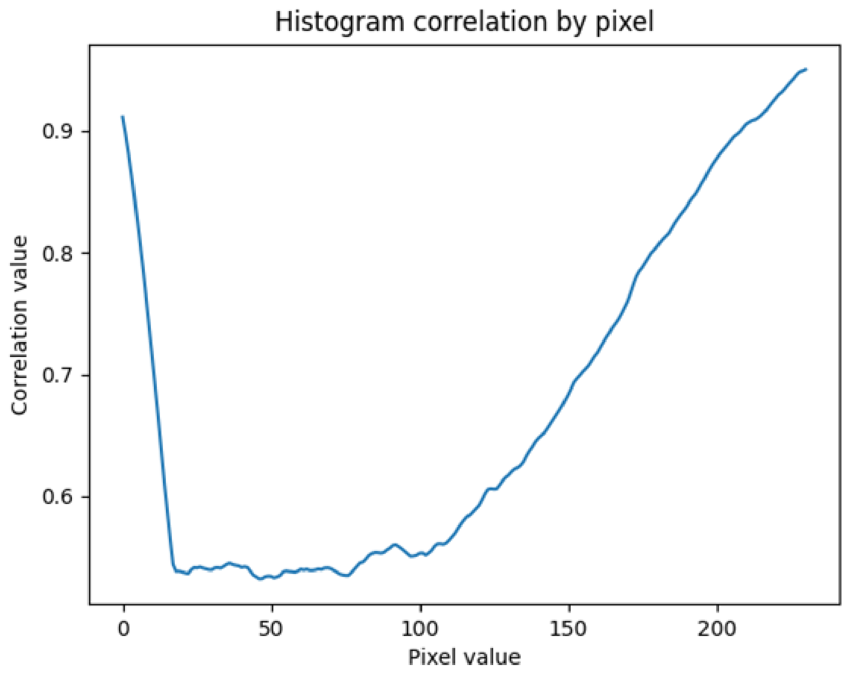


How to detect damaged regions

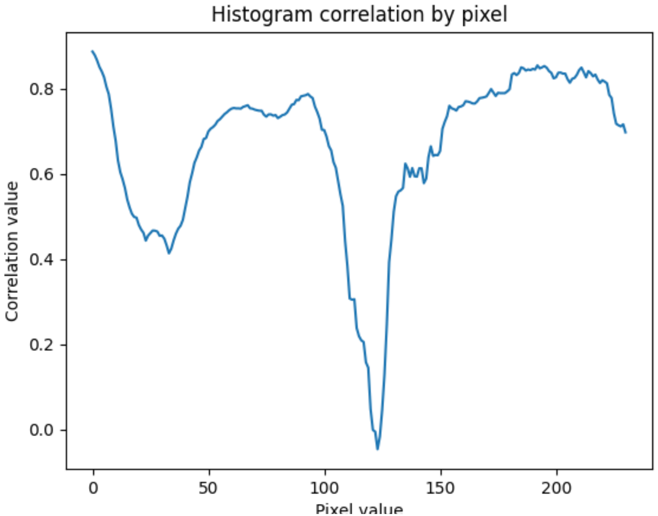
It can be done both by intensity detection and edge detection

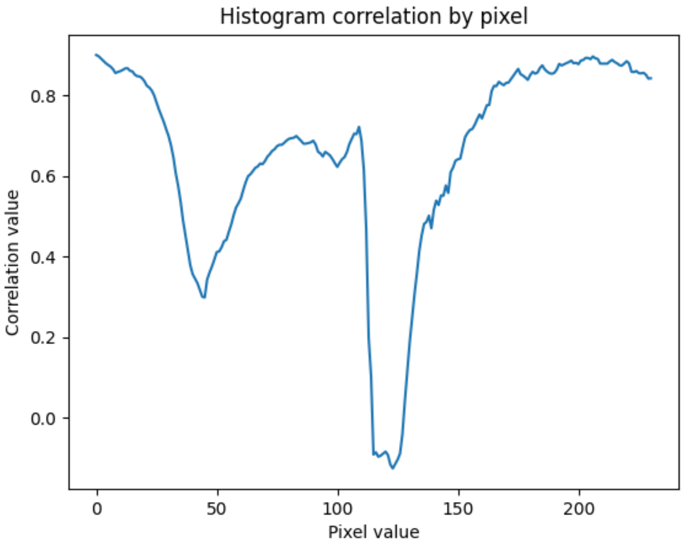
Currently we work on intensity detection that later will be supplied with contour detection. Reason for that order is easy — cancer growth has irregular shape patterns so intensity detection will allow to narrow out the detection plane.

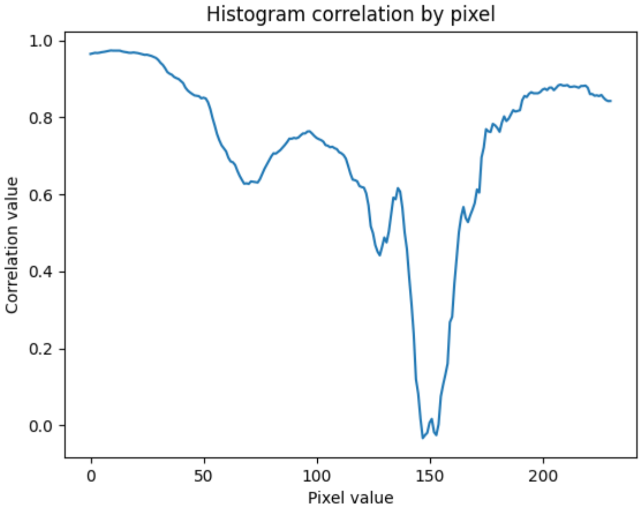
By comparing histogram correlation values we can detect non-symmetric regions.

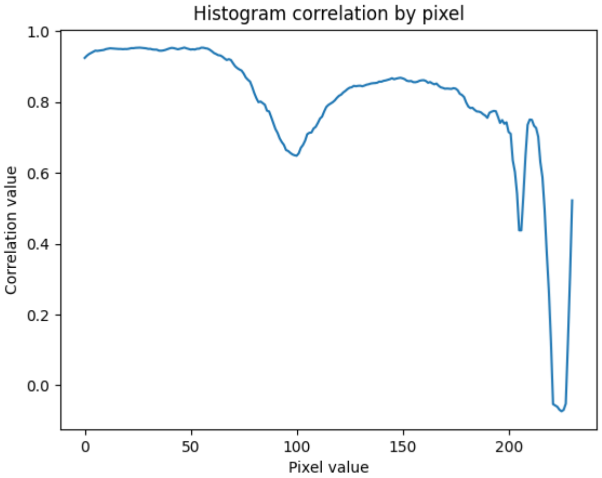


Vertical non-symmetric regions can be noticed by inspecting multiple row correlation plots









Applying syntactic pattern recognition

After having specified regions that require attention we will try to extract contours from those regions.

Syntactic pattern recognition is particularly useful for shape recognition. Having created an “language” of ill region contours inside the brain we should be able to point out which contours found are corresponding to oligodendroglioma.

Challenges

- oligodendroglioma contours may be different for each “level” of the MR scan — they take different shape near the ear level and different towards top of the head
- setting the optimal tiling (megapixel) size when analysing brightness intensity of an image
- setting the optimal threshold for intensity measurement
- accurate description of primitives

What more we can do

Important part of medics job is to contour out anatomically relevant structures. In radiotherapy some part of the radiation hits healthy brain structures, so having them contoured beforehand helps inform specialists and reduce possible damage to those structures.

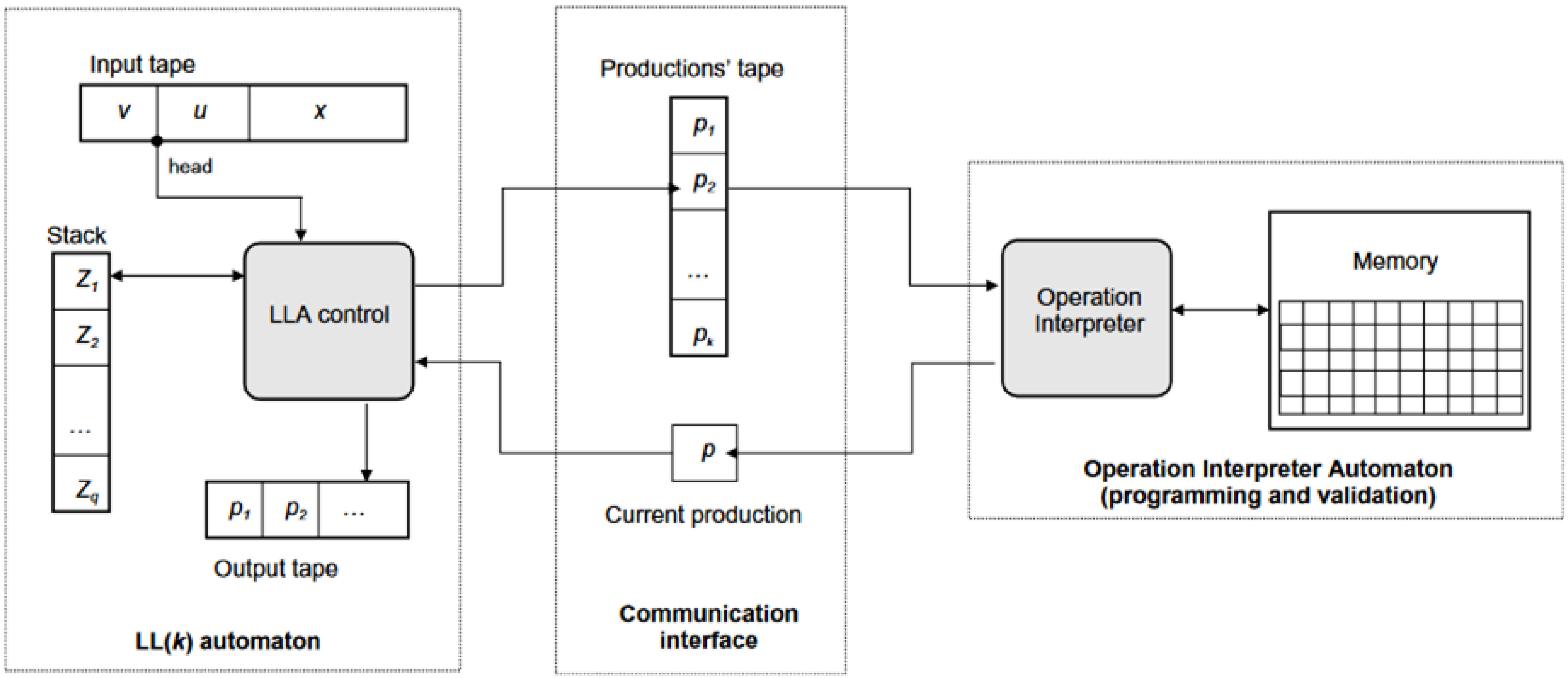
Using contour detection we can help specialists with the job.

The main tool - GDPLL(k)

GDPLL(k) language

$$G = (V, \Sigma, O, P, S, M)$$
$$V \text{ is a finite, nonempty alphabet;}$$
$$\Sigma \text{ is a finite, nonempty set of terminal symbols; } \Sigma \subset V$$
$$\text{nonterminal symbols are defined as } N = V \setminus \Sigma$$
$$O \text{ is a set of basic operations on the values stored in the memory}$$
$$S \in N \text{ is the starting symbol; } M \text{ is the memory;}$$
$$P \text{ is a finite set of productions of the form:}$$
$$p_i = (\mu_i, L_i, R_i, A_i), \text{ where}$$
$$\mu_i : M \to \{TRUE, FALSE\} \text{ is the predicate of applicability of the production } p_i$$
$$\text{defined with the use of operations performed over } M;$$
$$L_i \in N \text{ and } R_i \in V* \text{ are left} - \text{ and right} - \text{handed sides of } p_i \text{ respectively;}$$
$$A_i \text{ is the sequence of operations over } M, \text{ which should be performed if}$$
$$\text{the production is to be applied}$$

GDPLL(k) parsing automaton

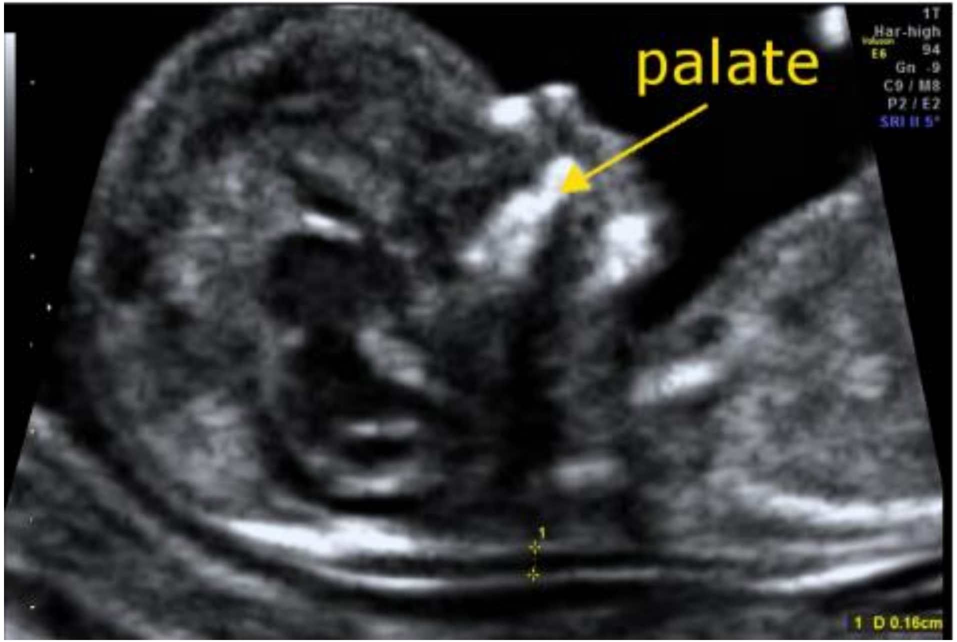


Using GDPLL(k) to solve problems

- one can create grammar corresponding to structures observed in real world — then it's easy to discern elements not belonging to those real world structures;
  It's done with an automaton corresponding to a found grammar,
- with sample good enough it's possible to infer a grammar automatically of course it needs to be done only once,
- parsing automaton has good computational complexity of O(n)

What already was accomplished using GDPLL(k) in medical field

It was possible to get great results in diagnostics of fetal palates



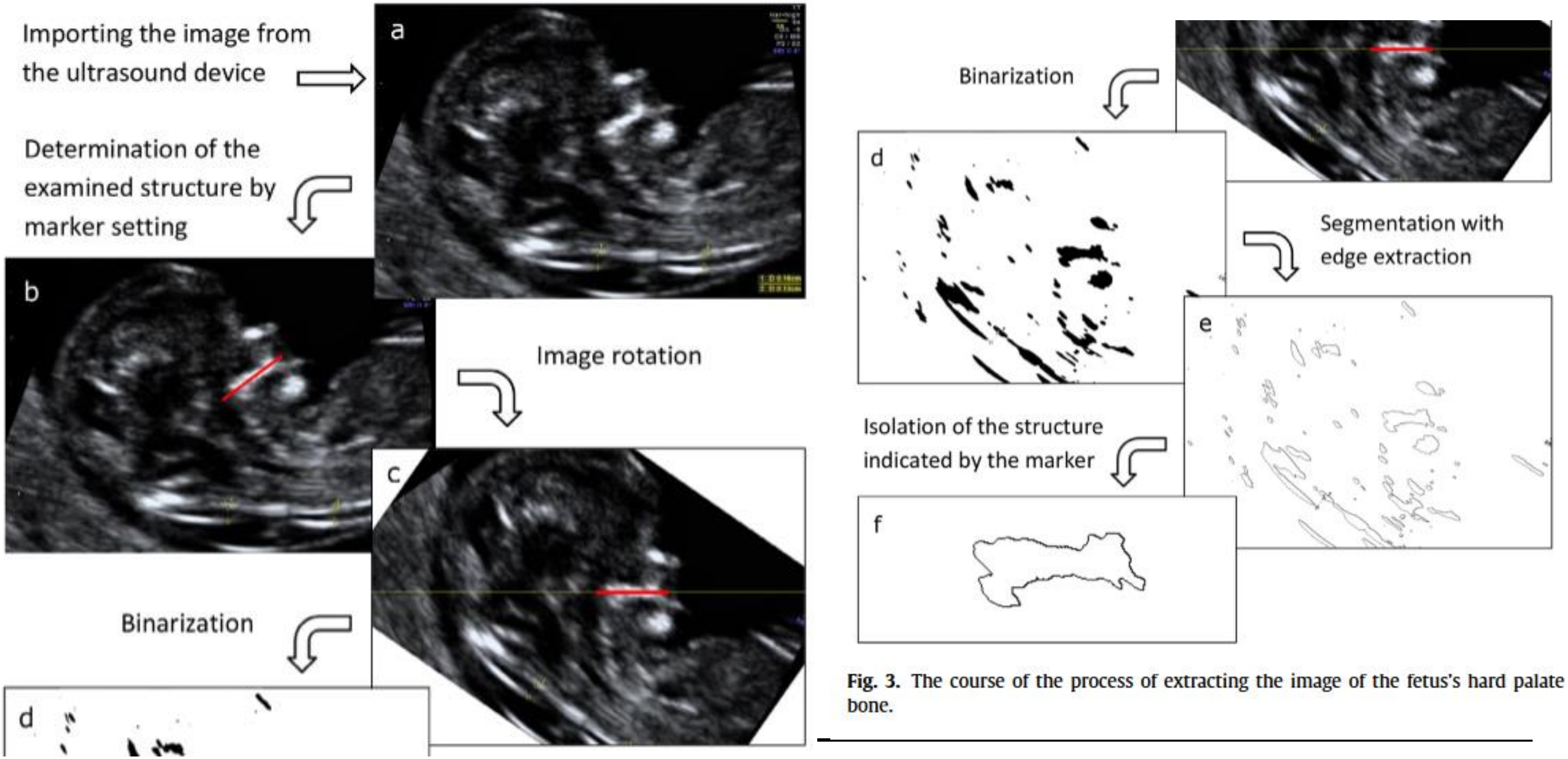


Fig. 3. The course of the process of extracting the image of the fetus's hard palate bone.

Shapes visible after applying different binarization thresholds (200, 176, 152, 128).

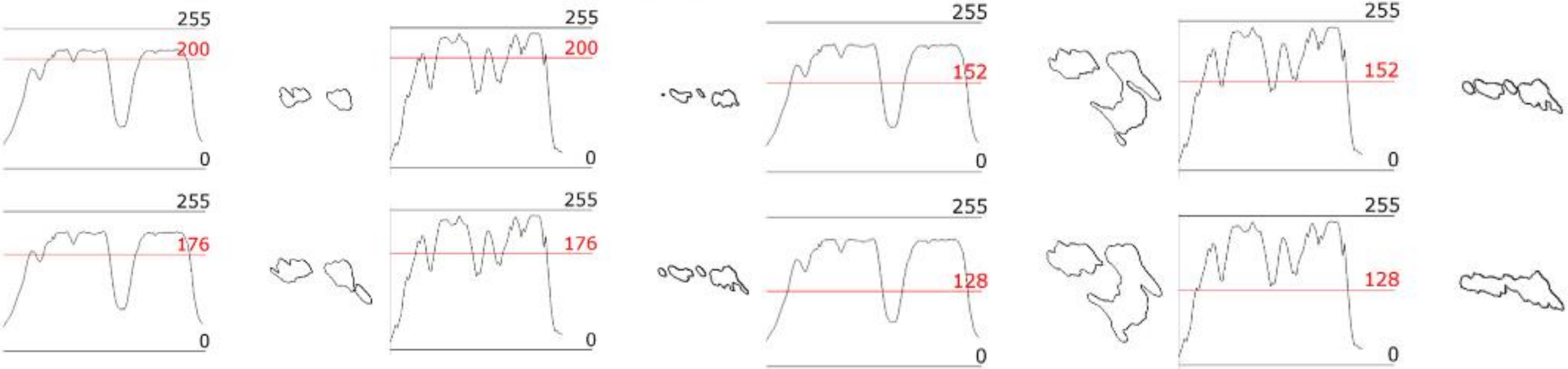


Fig. 5. Histogram and sequence of images in the fetus with cleft palate (a) and normal anatomy of the palate bone (b).

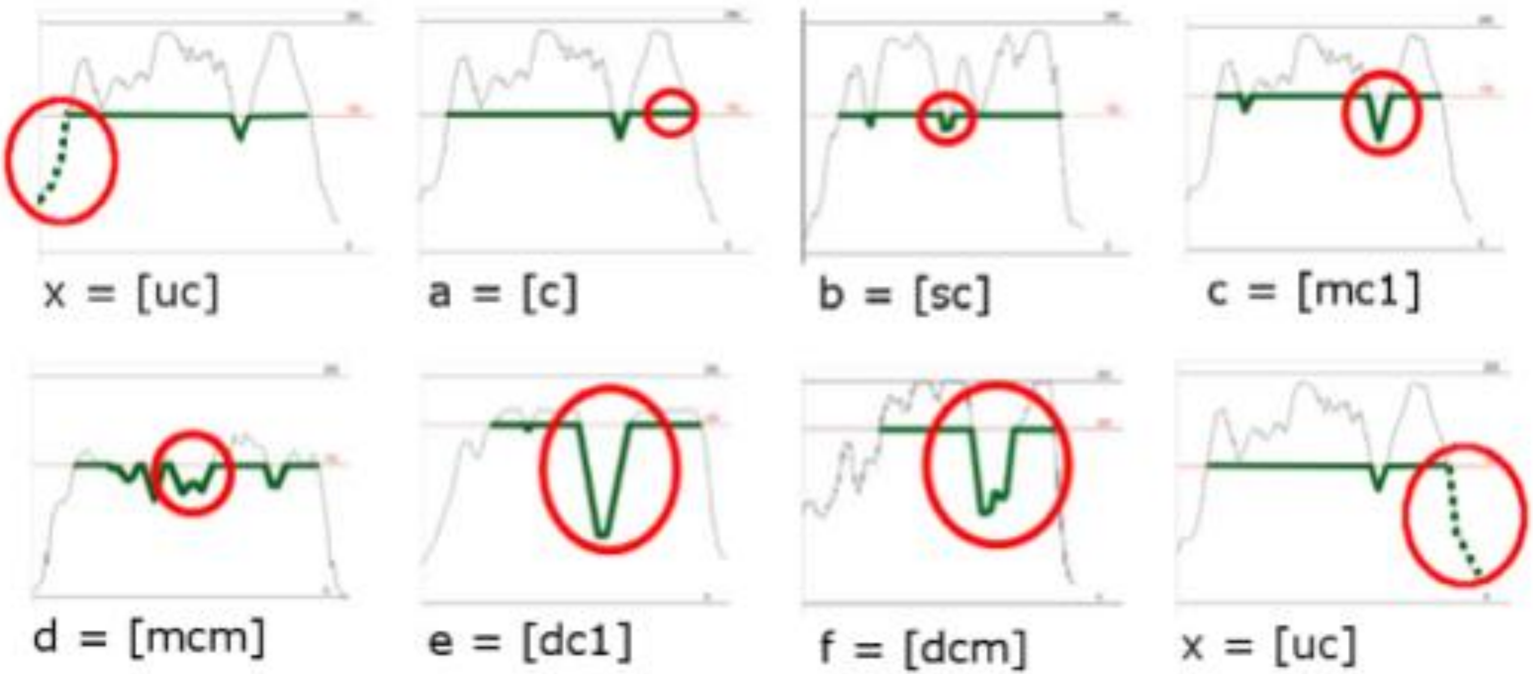


Fig. 4. Primitives for syntactic analysis of histogram defined as: a - continuous structure, b - small cavity of structure, c - medium cavity with one minimum, d - medium cavity with many local minima, e - deep cavity with one minimum, f - deep cavity with many local minima, x - unrecognized cavity.

—

Here you can see example patterns for both healthy and pathological cases.

Table 1
Examples of strings for physiological and pathological cases.

	Physiological cases		
Threshold	Case A	Case B	Case C
200	$x^2aca^3caca^4x$	$x^2afab^2a^4ba^4x$	$x^2a^5ca^2x^3$
176	$x^2aca^3caba^4x$	x^2afa^8x	$x^2a^5ca^2bx^2$
152	$x^2a^2ba^3ba^5x$	$x^2a^2da^8x$	$x^2a^5ba^3x^2$
128	$x^2a^{12}x$	$x^2a^3b^2a^8x$	$x^2a^{11}x^2$

	Pathological cases		
Threshold	Case D	Case E	Case F
200	$x^3a^5ea^3x$	$x^4a^4fa^2x$	a^5ca^3x
176	$x^3ba^5ea^3x$	$x^2ca^4da^2x$	a^5ca^3x
152	$x^2a^5ca^3x$	$x^2aca^4da^2x$	a^5ba^4x
128	$x^2a^5ca^3x$	$x^2a^5b^2a^2x$	a^5ba^4x

Summing up

1. GDPLL(k) are grammars with good descriptive power
2. They work really well in real-world setting analysis since real world consists mainly of structures
3. Medical field supplies good patterns for analysis
4. GDPLL(k) does not need big amounts of data to work and is explainable

Bibliography

Slides and information about fetal palate diagnostics is from:
Jurek, Janusz, Wojciech Wójtowicz, and Anna Wójtowicz. "Syntactic pattern recognition-based diagnostics of fetal palates." *Pattern Recognition Letters* 133 (2020): 144-150.

Informations and syntactic usage example pictures are from:
Flasiński, Mariusz. *Syntactic pattern recognition*. 2019.

Flasiński, Mariusz, and Janusz Jurek. "Fundamental methodological issues of syntactic pattern recognition." *Pattern analysis and applications* 17 (2014): 465-480.

Jurek, Janusz. "Recent developments of the syntactic pattern recognition model based on quasi-context sensitive languages." *Pattern Recognition Letters* 26.7 (2005): 1011-1018.

MRI scans are from:
Calabrese, E., Villanueva-Meyer, J., Rudie, J., Rauschecker, A., Baid, U., Bakas, S., Cha, S., Mongan, J., & Hess, C. (2022). **The University of California San Francisco Preoperative Diffuse Glioma MRI (UCSF-PDGM) (Version 4) [Dataset].** The Cancer Imaging Archive. DOI: 10.7937/tcia.bdgf-8v37