ქართული ნაბეჭდი სიმბოლოების სტილიზების ერთი ალგორითმის შესახებ

[1]ოთარ ვერულავა, [1]თოდუა თეა, [1]გვიჩიანი თამარ, [2]ჟვანია თალიკო
[1]საქართველოს ტექნიკური უნივერსიტეტი, ხელოვნური ინტელექტის დეპარტამენტი №76
[2]საქართველოს ტექნიკური უნივერსიტეტი, ეკონომიკური ინფორმატიკის დეპარტამენტი №96

ანოტაცია:
ნაშრომში განხილულია ქართული ნაბეჭდი სიმბოლოების ამოცნობის პრობლემის გადაწყვეტის მიზნით სიმბოლოების წინასწარი კომპიუტერული დამუშავების ერთი ალგორითმი. ეს ალგორითმი საშუალებას იძლევა მივიღოთ სიმბოლოს ძირითადი მონახაზი – ჩონჩხი და ამ გზით მას მოვაცილოთ ზედმეტი ინფორმაცია, დავტოვოთ სიმბოლოსათვის დამახასიათებელი ის ელემენტები, რომელიც აუცილებელია მის ამოსაცნობად. შემუშავებული ალგორითმის განხორციელებით მიიღება სტანდარტიზებული სიმბოლოები ნებისმიერი ქართული შრიფტისათვის.

საკვანძო სიტყვები: *ამოცნობა, პრეპარირება, დაგლუვება, სტილიზება, დავიწროება*

ქართული ნაბეჭდი სიმბოლოების ამოცნობისადმი მიძღვნილ სხვადასხვა ნაშრომებში განსაკუთრებული ყურადღება ეთმობა სიმბოლოების წინასწარი კომპიუტერული დამუშავების საკითხებს.

საწყისი გამოსახულება მოცემულია მატრიცის სახით, რომლის ნებისმიერი ელემენტი ნულის ან ერთის ტოლია:

$$X = \left\| x_{ij} \right\|, \ \forall x_{rs} \in \{0,1\}, \ i = \overline{0,I}; \ j = \overline{0,J}.$$

საწყისი გამოსახულების დამუშავების პირველ ეტაპზე ხორციელდება დაგლუვების პროცედურა, რაც საშუალებას იძლევა შემცირდეს გამოსახულების გრადაციები რასტრის ნაპირებზე და ბეჭდვისა და სკანირების ტექნოლოგიური პროცესებით გამოწვეული წყვეტები.

000000000111110000000000000	000000000111111100000000000
000000000111110000000000000	000000000111111100000000000
000000000111111100000000000	000000000111111100000000000
000000000111111100000000000	000000000111111100000000000
000000111111111111000000000	000000111111111111000000000
000000111111111111000000000	000000111111111111000000000
000000111111111111111000000	000000111111111111111000000
000000000111111111111000000	000000000111111111111000000
000000000111111111111110000	000000000111111111111110000
000000000001111111111110000	000000000001111111111110000
000000000001111111111111100	000000000001111111111111111
000000000000011111111111100	000000000000011111111111111
000000000000011111111111100	000000000000011111111111111
000000000000000111111111100	000000000000000111111111111
000000000000000111111111100	000000000000000111111111111
000000000000000000111111111	000000000000000000111111111
000000000000000000111111111	000000000000000000111111111
111111100000000000111111111	111111100000000000111111111
111111100000000000111111111	111111100000000000111111111
111111100000000000111111111	111111100000000000111111111
111111100000000000111111111	111111100000000000111111111
111111100000000000111111111	111111100000000000111111111
111111100000000000111111111	111111100000000000111111111
111111111000000000111111111	111111111000000000111111111
111111111000000000111111100	111111111000000000111111111
111111111000000111111111100	111111111000000111111111111
001111111000000111111111100	111111111000000111111111111
001111111111011111111111100	111111111111011111111111111
000011111111011111111110000	111111111111011111111111111
000011111111011111111110000	111111111111011111111111111
000000111111111111111000000	111111111111111111111111111
000000111111111111111000000	111111111111111111111111111

სურ.1. საწყისი და დაგლუვებული გამოსახულება

ქართული ნაბეჭდი სიმბოლოების სტრუქტურული ანალიზის უკეთ განხორციელების მიზნით შეიქმნა ე.წ. გადასვლების მატრიცა. X მატრიცის ყოველი სტრიქონისათვის გადასვლა ეწოდება ცვლილებას ბინარული ელემენტების თანმიმდევრობაში.

ცხადია, რომ ცვლილების ფაქტი, ანუ გადასვლა არის მაშინ, როდესაც ნულოვანი ელემენტის შემდეგ ერთის ტოლი ელემენტია ან, პირიქით.

ექსპერიმენტულად დადგინდა გადასვლების მაქსიმალური შესაძლო რაოდენობა. იგი განხორციელდა სიმბოლო "ლ"-ს შემთხვევაში და მან შეადგინა ცხრა გადასვლა.

$X = \left\| x_{ij} \right\|$ მატრიციდან, სადაც $\forall x_{ij} \in \{0,1\}$, $i = \overline{0;I}$, $j = \overline{0;J}$, $I \in N^{+}$, $J \in N^{+}$,

გადასვლებისLმატრიცის მიღების პროცედურა აღიწერება შემდეგნაირად:

საწყისი მატრიცის i-ური სტრიქონი $X_i = x_{i1,} x_{i2,} x_{i3} ..., x_{ij}, ..., x_{iJ}$

$X_i \supset \left(XR_i^1, XR_i^2 ..., XR_i^{k_i} ..., XR_i^{K_i} \right)$, სადაც $\forall XR_i^{k_i} = x_i^{k_i}[h_{k_i}]; h_{k_i} = \overline{j_{k_i;} J_{k_i}}$; $K \in N^{+}$

h_{k_i} წარმოადგენს $XR_i^{k_i}$ ქვესიმრავლეში ელემენტების თანმიმდევრობას.

გადასვლების რაოდენობა i-ური სტრიქონისათვის $K_i = \mathrm{Card}\left\{ XR_i^{k_i} \right\}$

$$\text{თუ } \forall x_i^{k_i}[h_{k_i}] = 0 \Rightarrow GAD_i^{k_i} = 0$$

$$\text{თუ } \forall x_i^{k_i}[h_{k_i}] = 1 \Rightarrow GAD_i^{k_i} = 1$$

$GD = \max\{K_i\}$; $GD \le GANR$; სადაც GD გამოთვლებით მიღებული გადასვლების მაქსიმალური რაოდენობაა, $GANR$ - გადასვლების მატრიცის სვეტების განზომილებაა.

თუ $GAD_i^{K_i} = 0 \Rightarrow GAD_i^{k_i+f_i} = 0$;

თუ $GAD_i^{K_i} = 1 \Rightarrow GAD_i^{k_i+f_i} = 1$;

$K_i + f_i = GANR; f \in N^+$; საბოლოოდ, მიიღება მატრიცა $GAD = \left\| GAD_i^{k_i+f_i} \right\|$, რომლის განზომილებაა $Ix(K_i + f_i)$.

გადასვლების მატრიცის შედგენის პროცესი მოიცავს სიმბოლოს სტრუქტურაზე ინფორმაციის მიღებას. კერძოდ, ნებისმიერი i -ური სტრიქონისათვის გამოითვლება ერთებისა და ნულების უწყვეტი თანმიმდევრობების რაოდენობა და შესაბამისად, ერთებისა და ნულების რიცხვი აღნიშნულ თანმიმდევრობებში. მიღებული გადასვლების მატრიცის მიხედვით ხდება ახალი მატრიცის შედგენა, რომელიც სიმბოლოს უფრო დეტალიზებულ აღწერას იძლევა. მასში ასახულია მატრიცის თითოეული სტრიქონისათვის ნულებისა და ერთების უწყვეტ თანმიმდევრობებში ნულებისა და ერთების რიცხვითი მნიშვნელობების გამოთვლის შედეგები. გარდა ნულებისა და ერთებისა, სიმბოლოს სტრუქტურის აღსაწერად შემოდის ორიანიც. ორიანი მოკლე ხაზს აღნიშნავს, ნულები, ცხადია, სიმბოლოს სტრუქტურის არარსებობის მაჩვენებელია. ამ მატრიცისთვის დამახასიათებელი ზემოთ აღნიშნული თვისებების გამო მას ვუწოდებთ ხაზის სიგრძეთა მაჩვენებელ მატრიცას (შემდგომში მას შემოკლებით ვუწოდებთ „LL“ მატრიცას).

0100000000	0	2	0	0	0
0100000000	0	2	0	0	0
0100000000	0	2	0	0	0
0100000000	0	0	2	0	0
0100000000	0	2	0	0	0
0100000000	0	2	0	0	0
0100000000	0	0	2	0	0
0100000000	0	0	0	2	0
0100000000	0	0	0	2	0
0100000000	0	0	0	2	0
0111111111	0	0	0	1	1
0111111111	0	0	0	1	1
0111111111	0	0	0	1	1
0111111111	0	0	0	1	1
0111111111	0	0	0	1	1
0111111111	0	0	0	0	2
0111111111	0	0	0	0	2
1011111111	2	0	0	0	2
1011111111	2	0	0	0	2
1011111111	2	0	0	0	2
1011111111	2	0	0	0	2
1011111111	2	0	0	0	2
1011111111	2	0	0	0	2
1011111111	2	0	0	0	2
1011111111	2	0	0	0	2
1011111111	2	0	0	0	2
1011111111	2	0	0	0	2
1011111111	2	0	0	0	2
1011111111	2	0	0	0	2
1111111111	1	1	1	1	1
1111111111	1	1	1	1	1

სურ.2. გადასვლების მატრიცა და LL მატრიცა (სიმბოლო “ა”)

ქვემოთ ჩამოყალიბებულია გადასვლების მატრიციდან „LL“ მატრიცის მიღების ზოგადი პრინციპები:

როდესაც გადასვლების მატრიცაში გვაქვს 01 ან 10 გრძელი ხაზი (გრძელია ხაზი, რომელშიც ერთების უწყვეტი თანმიმდევრობა რასტრის სიგანის 4/5-ზე მეტია), „LL“ მატრიცაში რასტრის მთელ სიგანეზე ჩაიწერება ერთიანები. ეს სიტუაცია ანალოგიურია გადასვლების მატრიცაში ერთების მხოლოდ ერთი უწყვეტი თანმიმდევრობის არსებობის შემთხვევისა.

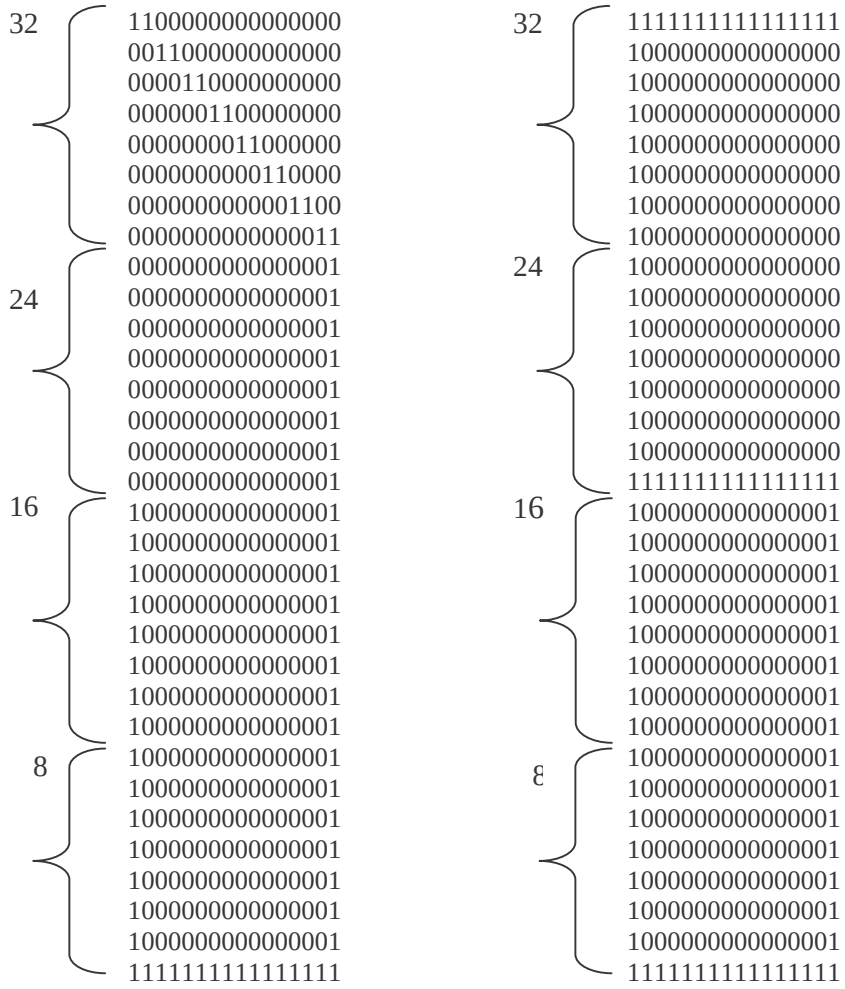
10 გადასვლის არსებობისას, თუ ერთიანების რაოდენობა რასტრის სიგანის 50-60%-ია, „LL“ მატრიცის i-ურ სტრიქონში ჩაიწერება 11000, რაც კარგად აჩვენებს მოცემული მატრიცის i-ური სტრიქონისათვის ხაზის სიგრძესა და მის მდებარეობას. იმ შემთხვევაში, როდესაც ერთების უწყვეტი თანმიმდევრობა რასტრის სიგანის 60-70%-ია, მაშინ, ცხადია, ერთების უწყვეტ თანმიმდევრობაში არსებული ერთების მეტი რაოდენობა აისახება „LL“ მატრიცის i-ურ სტრიქონში ერთი ერთიანის დამატებით: 11100.

განსაკუთრებული ყურადღება ექცევა 010 გადასვლის აღწერის გადატანას „LL“ მატრიცაში. ვინაიდან 010 გადასვლის შემთხვევაში, უმეტესწილად ადგილი აქვს დახრილი ხაზების არსებობას (სიმბოლოები “ა”, “რ”, “ლ”, “დ”, “ღ”). პრობლემა მდგომარეობს იმაში, რომ „LL“ მატრიცაში შესაბამის ადგილზე ჩაიწეროს გამოსახულების სტრუქტურა. ამისათვის, გათვალისწინებულ უნდა იქნას მოცემული სტრიქონისათვის ერთისა და ნულების უწყვეტ თანმიმდევრობებში ერთებისა და ნულების რაოდენობა. იმისგან დამოკიდებულებით, რამდენად მეტია ან რამდენად ნაკლებია ნულების პირველი უწყვეტი თანმიმდევრობა ნულების მეორე უწყვეტ თანმიმდევრობაზე და რასტრის სიგანის რა ნაწილს შეადგენს ერთების რაოდენობა, გამოსახულების i-ური სტრიქონისათვის „LL“ მატრიცაში მიიღება შემდეგი სახის ჩანაწერები: 00020, 02000, 01110, 00022. იმ შემთხვევაში, როდესაც ნულების პირველი უწყვეტი თანმიმდევრობა ტოლია ან მცირედ განსხვავდება ნულების მეორე უწყვეტი თანმიმდევრობისგან, გამოსახულების i-ური სტრიქონისათვის „LL“ მატრიცაში მიიღება: 00200.

მსგავსი მიდგომებია შემუშავებული 101 და უფრო რთული 1010, 0101, 01010, 10101, 101010, 0101010 გადასვლებისთვის.

LL მატრიციდან მიღებულ იქნა სიმბოლოს ჩონჩხი, რომელიც თავისუფალია ყოველგვარი ზედმეტი ელემენტებისგან (სურ.3). ქართული ანბანის ზოგი სიმბოლოსათვის აუცილებელი გახდა სიმბოლოს სრუქტურაში არსებული დახრის განსაზღვრა, რომელიც მათ განასხვავებდა ანბანის სხვა სიმბოლოებისგან. ამისათვის LL მატრიცის თითოეული სტრიქონისათვის განისაზღვრა ორისა და ორებისგან განსხვავებული ელემენტების რაოდენობა. დახრილი ხაზის ძიება ხდება რასტრის ზედა და ქვედა ნაწილში, გარკვეული, ექსპერიმენტულად განსაზღვრული ჩარჩოს ფარგლებში, ვინაიდან ქართული ნაბეჭდი სიმბოლოებისათვის დახრა სწორედ ამ არეებშია შესაძლებელი. თუ ორ მეზობელ i და i+1 სტრიქონში ორების რაოდენობა ერთის ტოლია და ამასთან $X_{ij}=2$ და $X_{i+1,j+1}=2$ და ასეთი სიტუაცია საძიებელი ჩარჩოს შიგნით ერთხელ მაინც იქმნება, ამ შემთხვევაში სიმბოლოში ფიქსირდება დახრა.

სიმბოლოს ჩონჩხის გამოსახაზად შერჩეულ იქნა მატრიცა განზომილებით 32X16. ის ვერტიკალზე პირობითად გაყოფილია ოთხ ნაწილად (8; 16; 24; 32)



სურ.3. საბოლოო შედეგები სიმბოლო “ა”-სა და “წ”-ს მაგალითზე

მაგალითისათვის განვიხილოთ სიმბოლოს გამოხაზვის ზოგადი პრინციპები რასტრის მარცხენა და მარჯვენა ნაპირების მაგალითზე. ნაპირებზე 01 გადასვლის არსებობისას თუ ვერტიკალზე ნულებისა და ერთების რაოდენობა ტოლია, მაშინ i=16-დან მატრიცის ბოლო სტრიქონამდე j=0 სვეტში ჩაიწერება ერთები. თუ LL მატრიცის მარცხენა ან მარჯვენა ნაპირზე არსებობს გამოსახულების სტრუქტურა (ორები ან ერთები), რომელიც იწყება 16-სა და 24-ს შორის არიდან, გრძელდება რასტრის ბოლოს სტრიქონამდე და ამასთანავე, გამოსახულების სტრუქტურები უფრო ახლოსაა 16-თან, მაშინ ერთები ჩაიწერება j=0 ან j=15 სვეტში, ისევე, როგორც ზემოთ მოყვანილ შემთხვევაში, i=16-დან მატრიცის ბოლო სტრიქონამდე. იმ შემთხვევაში, თუ გამოსახულების სტრუქტურები უფრო ახლოსაა 24-თან, მაშინ სიმბოლოს მარცხენა ან მარჯვენა ნაპირზე იქმნება ვერტიკალური ხაზი i=8 სტრიქონიდან მატრიცის ბოლო სტრიქონამდე, j=0 სვეტში.

ზემოთ აღწერილი პროცედურების გამოყენებით მიიღება სიმბოლოს ძირითადი მონახაზი – მისი ჩონჩხი. ამ გზით მას მოცილებულია ზედმეტი ინფორმაცია, დატოვებულია სიმბოლოსათვის დამახასიათებელი ის ძირითადი ელემენტები, რომელიც აუცილებელია მის ამოსაცნობად. შემუშავებული ალგორითმის განხორციელებით მიიღება სტანდარტიზებული სიმბოლოები ნებისმიერი ქართული შრიფტისათვის.

ლიტერატურა:

1. თ. თოდუა. ქართული ნაბეჭდი სიმბოლოების წინასწარი კომპიუტერული დამუშავება.. //ქართული ელექტრონული სამეცნიერო ჟურნალები: კომპიუტერული მეცნიერებები და ტელეკომუნიკაციები. 2004|№1(3), http://gesj.internet-academy.org.ge/gesj_articles/1045.pdf;
2. თ. თოდუა, მ. ჩხაიძე. გამოსახულებათა ამოცნობა გადასვლების მატრიცის მეთოდით. "მეცნიერება და ტექნოლოგიები" №10-12,თბილისი, 2002 წ.
3. Arcelli C. Pattern thinning by contour tracing, Comput. Graphics Image Processing, vol. 17, 1981, pp.130-144.

Article received: 2006-11-15