„Sylfaen" შრიფტის ქართული სიმბოლოების ანალიზი და ამოცნობა

ოთარ ვერულავა, ზურაბ წვერიკმაზაშვილი
საქართველოს ტექნიკური უნივერსიტეტი

რეზიუმე

ნაბეჭდი ტექსტის სკანირებით მიღებული გრაფიკული გამოსახულების გადასაყვანად ტექსტურ ფორმატში შერჩეულია პროგრამული ანალიზის მეთოდი. შერჩევის კრიტერიუმებია ნაბეჭდი ტექსტის პრეპარირების, კერძოდ, მასშტაბირების და სეგმენტაციის გამოყენება სახეთა აღწერების – ეტალონების ასაგებად. განხილულია პრეპარირების შედეგების მიხედვით სიმბოლოების ამოცნობის პროცესის ფორმირება „Sylfaen" შრიფტის ქართული სიმბოლოებისათვის უნიკოდ კოდირებით, რომელიც შეიცავს 74 სიმბოლოს: ქართული ანბანის სიმბოლოებს, არაბულ ციფრებს, სასვენ ნიშნებს, სხვადასხვა სახის ფრჩხილებს, ზოგიერთ მათემატიკურ სიმბოლოს. ზემოთ ჩამოთვლილი პროცედურებისთვის ფორმირებულია ალგორითმები და მათი შესაბამისი პროგრამული კოდები C++ ენაზე.

საკვანძო სიტყვები: ნაბეჭდი ტექსტი. ანალიზი. ამოცნობა. პრეპარირება.

1. შესავალი

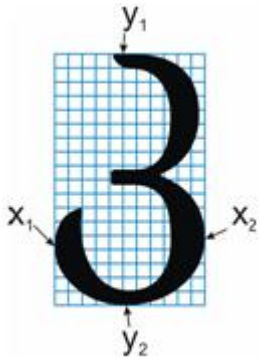
ქართული ტექსტის პროგრამული ანალიზის მეთოდები ფრიად მრავალფეროვანია მათი უმრავლესობა განხილულია ნაშრომებში [1-5]. პირობითად შესაძლებელია მათი დაჯგუფება ორ მიმართულებად. პირველს შეიძლება მივაკუთვნოთ ასომთავრული ანუ სართულების გარეშე ანალიზის მეთოდები, რომელშიც შედის ე.წ. „მეზობელი პიქსელების" მეთოდი, ხოლო მეორე მიმართულებას – ტექსტის ანალიზი სართულების გამოყოფით. კერძოდ, ასომთავრული ანალიზის შემთხვევაში იკარგება ქართული სიმბოლოების მნიშვნელოვანი ნიშანი, როგორიცაა მათი განლაგება სართულების მიხედვით, რაც იწვევს სირთულეებს ზოგიერთი სიმბოლოების გარჩევისას, მაგალითად: „პ" და „კ", „გ" და „ბ" და ა.შ. მეორე შემთხვევაში სირთულეებია ასომთავრულად დაბეჭდილი ტექსტის (სათაურები, ქვესათაურები და ა.შ) ამოცნობის პროცესში. აქედან გამომდინარე ცხადია, რომ საჭიროა ქართული ტექსტის ანალიზის ისეთი მეთოდის ფორმირება, სადაც შენარჩუნებული იქნება ორივე მეთოდის დადებითი მხარეები და ნიველირებული იქნება უარყოფითი.

2. ანალიზი

ნაშრომის მიზანია ანალიზის ისეთი მეთოდის შემუშავება, სადაც შესაძლებელი იქნება სიმბოლოების სართულებად განლაგების წარმოდგენა და სხვა ამოცნობის ნიშნების ისეთი პარამეტრების განსაზღვრა, რომელიც უფრო ეფექტურს და საიმედოს გახდის ამოცნობის პროცესს. ხორციელდება ამოსაცნობი ფურცლის „შავ-თეთრ" რეჟიმში სკანირების შედაგად მიღებული „bmp" ფორმატის მქონე გრაფიკული ფაილის ბინარულ მატრიცად გარდაქმნა, რომლის თითოეული ელემენტისთვის მნიშვნელობის მინიჭება ხდება გრაფიკული რასტრის შესაბამისი ელემენტის ანუ პიქსელის ფერის მიხედვით. თეთრი ფერის შემთხვევაში გვექნება – 0, ხოლო შავი ფერის შემთხვევაში – 1.

ფურცლის ანალიზი იწყება ზედა რიგის მარცხენა პიქსელიდან და თანმიმდევრულად გადის ყველა პიქსელის რიგის ბოლომდე და შემდეგ გადადის ქვედა რიგში. თანდმიმდევრული ანალიზის პროცესი წყდება მაშინ, როდესაც გვხვდება 1-ის ტოლი მატრიცის ელემენტი.

ამ მომენტიდან განხილვის პროცესი იცვლება და ვიწყებთ მოცემული პიქსელის 1-ის ტოლი მნიშვნელობის მქონე მეზობელი პიქსელების ძიებას. ამ პროცესში ხდება ყველა მეზობელი პიქსელის კოორდინატის დამახსოვრება და მათი მიკუთვნება სეგმენტირებულ სიმბოლოთა სიმრავლის ახალი ელემენტისთვის. ეს სიმრავლე წარმოდგენილია ვექტორის სახით. შემდეგ ხდება გამოყოფილი სიმბოლოების მინიმალური და მაქსიმალური წერტილების განსაზღვრა აბსცისათა X და ორდინატთა Y ღერძების მიხედვით (ნახ.1).



ნახ.1

$$x_1^* = \min\{x_i\};\ x_2^* = \max\{x_i\}; \qquad (1)$$

$$y_1^* = \min\{y_i\};\ y_2^* = \max\{y_i\}; \qquad (2)$$

სადაც: $\forall x_i$, $\forall y_i$ - კოორდინატების შესაბამისი პიქსელების მნიშვნელობები ერთის ტოლია;

$i = \overline{1:N}$, N - რასტრის განზომილება, ანუ პიქსელების საერთო რაოდენობაა;

x^*, y^* - შესაბამისი ღერძების უმცირესი და უდიდესი მნიშვნელობებია იმ სიმბოლოს მიხედვით, რომლის ანალიზიც განხორციელდა.

(1) და (2) გამოსახულებებით განსაზღვრული x^*, და y^* მნიშვნელობები საშუალებას გვაძლევს დავადგინოთ მოცემული სიმბოლოს განლაგება სართულების მიხედვით. თუ ამ მნიშვნელობებს შევადარებთ ტექსტის სხვა სიმბოლოების ანალოგიურ მნიშვნელობებს, შეიძლება ითქვას, რომ განხილული პარამეტრები ირიბად წარმოადგენს სიმბოლოების სართულებს. ამის გარდა, ეს მახასიათებლები საშუალებას გვაძლევს საჭიროების შემთხვევაში დავადგინოთ ტექსტში გამოყენებული შრიფტის კეგელის მნიშვნელობა.

3. მასშტაბირება

მასშტაბირების მიზანია მივიღოთ თანაბარგანზომილებიანი რეალიზაციები, რაც საშუალებას მოგვცემს ეფექტურად განვახორციელოთ ამოცნობის პროცესი, კერძოდ, უცნობი რეალიზაციისა და ეტალონური აღწერის შედარების პროცედურა, გადაწყვეტილების მიღება, მარტივდება პროგრამული მოდულის აგებაც.

მოცემული შრიფტისათვის მივიღეთ, რომ ნებისმიერი მნიშვნელობის კეგელის დაყვანა მოხდება 14 ზომის კეგელამდე, რისთვისაც გამოყენებული იქნება დაყვანის არსებული პროგრამული მოდულები, რომელთა მარტივი მოდიფიკაციით ვაღწევთ დასახულ მიზანს. ამ მეთოდის გამოყენებით ყველა ზომის შრიფტი დაგვყავს 14 კეგელის ზომაზე, რაც საშუალებას გვაძლევს 14 ზომის კეგელის სიმბოლოები გამოვიყენოთ მოცემული შრიფტის ეტალონურ აღწერებად. უნდა აღინიშნოს, რომ მასშტაბირების პროცესში შესაძლებელია ზოგიერთი სიმბოლოს დამახინჯება, რაც აუცილებლადაა გასათვალისწინებელი ამოცნობის პროცესში.

4. სეგმენტაცია – რეალიზაციის მიღება

მეზობელი პიქსელის ანალიზით მიღებული შედეგი წარმოადგენს ვექტორულად განლაგებულ სიმბოლოთა თანმიმდევრობას, რომელშიც შემდგომში გამოიყოფა რიგები. რიგების გამოყოფა ხდება შემდეგი წესის მიხედვით: თუ განსახილველი სიმბოლოს მინიმუმი, რომელიც განისაზღვრება 1-ელ ნახაზზე მოცემული y_2 წერტილის მიხედვით, მეტია წინა სიმბოლოებით განსაზღვრულ მაქსიმუმზე $\max\{y_2\}$, მაშინ იწყება ახალი რიგი.

შემდეგ პროცედურაში ხდება სიმბოლოების თანმიმდევრობის შეცვლა – დალაგება აბსცისათა ღერძის კოორდინატთა მნიშვნელობების ზრდის მიხედვით. ამ პროცედურის აუცილებლობა გამოწვეულია იმ ფაქტით, რომ მეზობელი პიქსელების პრინციპით ანალიზისას სიმბოლოები ლაგდება სიმაღლის მიხედვით და არა ბუნებრივი თანმიმდევრობით. სიტყვებისა და ცალკეული სიმბოლოების გამოყოფა ხდება ჰარის მიხედვით.


ნახ.2

აღწერილი პროცედურების შესრულების შემდეგ ვიღებთ სიმბოლოს აღწერას ბინარული მატრიცის სახით (ნახ.2), რომელიც შეესაბამება შრიფტის ბუნებრივ ზომებს, რაც შემდგომში დაგვყავს ეტალონური რეალიზაციების (14 ზომის კეგელი) ზომაზე.

5. ამოცნობა

ამოცნობის პროცესში ხდება უცნობი ან სასწავლო ნაკრების რეალიზაციების (დაყვანილი 14 კეგელის ზომაზე) და ეტალონური რეალიზაციების შედარება.

ავღნიშნოთ უცნობი რეალიზაციების შესაბამისი ბინარული მატრიცა RE -თი, რომლის ელემენტებია x_i, ხოლო ეტალონური მატრიცა ET -თი, რომლის ელემენტებია – e_i, სადაც $i = \overline{1;N}$. რადგან მატრიცები თანაბარზომიანია, ამიტომ შედარებისას ვიყენებთ სუპერპოზიციის პრინციპს. ვითვლით რეალიზაციის მატრიცის ერთის ტოლი მნიშვნელობის ელემენტებისა და შესაბამისი ეტალონური აღწერების ერთიანების დამთხვევათა რიცხვს, რომელსაც ავღნიშნავთ - φ -ით. ასევე ვითვლით არდამთხვეული პიქსელების რაოდენობას – ავღნიშავთ λ -თი, რის შემდეგაც გამოვთვლით სხვაობას ($\varphi - \lambda$). ამის შემდეგ ვითვლით ერთიანების საერთო რაოდენობას რეალიზაციაში – ავღნიშნავთ $\sum_i x_i$, ვანგარიშობთ სხვაობის პროცენტულ რაოდენობას რეალიზაციის ერთიანების საერთო რაოდენობის მიმართ შემდეგი გამოსახულებების მიხედვით:

$$S = \frac{\varphi - \lambda}{\sum_i x_i} \cdot 100 \qquad (3)$$

გადაწყვეტილება უცნობი ან საგამოცდო რეალიზაციის მიხედვით მიიღება S ფუნქციის მაქსიმუმით:

$$RE \in A_i \quad \text{თუ} \quad S_i = \max\ S_j, \quad j = \overline{1;I}\ \ i = \overline{i=1;I}, \tag{4}$$

სადაც $I = Card\{A\}$, სახეთა რაოდენობაა A სიმრავლეში.

6. კლასტერირება

მატრიცების დაყვანისა და სუპერპოზიციის მეთოდით შედარების ალგორითმები დიდ გამოთვლით რესურსებს მოითხოვს. ამოცნობის დროის შემცირების მიზნით საჭიროა ეტალონების კლასტერირება სახეების მიხედვით, რისთვისაც ვიყენებთ მათ მატრიცებს. დაყვანა ხდება X და Y ღერძების განზომილებების მიხედვით ცალ-ცალკე, რომლის მიზანია ნაკლები რაოდენობის დაყვანა - შედარების ოპერაციების შესრულება. კლასტერირებისთვის ტარდება შემდეგი პროცედურები: ვადგენთ რეალიზაციის მატრიცის R_y განზომილების რამდენი პროცენტით შეცვლა (გაზრდა, შემცირება) არის საჭირო, რომ გაუთანაბრდეს ეტალონის E_y განზომილებას.

$$L_1 = \frac{(E_y - R_y)}{R_y} \cdot 100 \tag{5}$$

სადაც: E_y, R_y - რეალიზაციის და ეტალონის განზომილებებია Y ღერძის მიხედვით.

L_1 - რეალიზაციის Y განზომილების ცვლის პროცენტი.

შემდეგ ხდება რეალიზაციის R_x განზომილების X ღერძის მიხედვით L_1 პროცენტით შეცვლა და შედეგის ეტალონთან შედარება X ღერძის მიხედვით. დასაშვებია ± 8 პიქსელის ცდომილება:

$$(E_x - C_1) \le R_x + \frac{R_x L_1}{100} \le (E_x + C_1) \tag{6}$$

სადაც: $C_1 = 8$; R_x და E_x - რეალიზაციის და ეტალონის X განზომილებებია;

C_1 - ცდომილების კოეფიციენტია.

იგივე პროცედურა ტარდება რეალიზაციის R_x განზომილების მიმართ.

$$L_2 = \frac{(E_x - R_x)}{R_x} \cdot 100$$

$$(E_y - C_2) \le R_y + \frac{R_y L_2}{100} \le (E_y + C_2) \tag{7}$$

სადაც: $C_2 = 6$.

ცდომილების კოეფიციენტები დადგენილია ექსპერიმენტულად. ყველა ის მატრიცა რომელიც დააკმაყოფილებს (6) გამოსახულებაში მოცემულ პირობას, მოთავსებული იქნება რეალიზაციისთვის შესადარებელ ეტალონთა - R_c სიმრავლეში, ხოლო (7) გამოსახულების პირობის დაკმაყოფილების შემთხვევაში ეტალონი ასევე მოთავსებული იქნება R_c-ში, თუ R_c უკვე არ შეიცავს ასეთ ეტალონს.

7. ექსპერიმენტული კვლევა

პროგრამა შესრულებულია პროგრამირების ენა C++ -ზე მულტიპლატფორმული მხარდაჭერით. გამოყენებულია კომპილატორი GNU g++ და პროგრამული ინტერფეისი QT.

ორ და მეტ ნაწილიანი სიმბოლოების ამოცნობას როგორიცაა: „ : “, „ ? “, „ % “ და ა.შ. ესაჭიროება სპეციალური დამუშავება, რომელთა ამოცნობა ხდება მათი შემადგენელი სეგმენტების X და Y ღერძების შესაბამისი კოორდინატების ადგილმდებარეობის დადგენით ერთმანეთის მიმართ.

წინამდებარე ალგორითმი არაეფექტურია ბეჭდვის შედეგად წარმოქმნილი ისეთი დეფექტების შემთხვევაში, როგორიცაა ერთმანეთზე გადაბმული და მცირე ზომის ფრაგმენტებად დანაწევრებული სიმბოლოები, რაც საჭიროებს სეგმენტაციისა და კლასტერირების დამატებითი ალგორითმების ინტეგრირებას.

ამოცნობისათვის გამოყენებული იქნა შემდეგი ეტალონები:

აბგდევზთიკლმნოპჟრსტუფქღყშჩცძწჭხჯ31234567890.,!? '"^*@#$& ()<>{}[]-+\ /|/o''

ამოცნობის შედეგები ერთი და იგივე 4 ფორმატის ფურცლისთვის სხვადასხვა გარჩევადობით:

გარჩევადობა 600dpi		გარჩევადობა 300dpi	
შრიფტი	უნიკოდი	შრიფტი	უნიკოდი
შრიფტის კეგელი	10	შრიფტის კეგელი	10
სიმბოლოების რაოდენობა	3422	სიმბოლოების რაოდენობა	3422
სეგმენტაციის დრო	2.63 წ.	სეგმენტაციის დრო	2.51 წ.
ამოცნობის დრო	45.69 წ.	ამოცნობის დრო	32.85 წ.
მთლიანი დრო	48.32 წ.	მთლიანი დრო	35.36 წ.
დაშვებული შეცდომები	0	დაშვებული შეცდომები	3
ამოცნობის პროცენტი	100 %	ამოცნობის პროცენტი	99.91 %

3. დასკვნა

შემუშავებულია ქართული ნაბეჭდი სიმბოლოების ანალიზის, პრეპარირებისა და ამოცნობის ალგორითმი შესაბამისი პროგრამული მოდულების აგებით. ექსპერიმენტულმა კვლევებმა ამოცნობის პროცესების განხორციელებით აჩვენა შემუშავებული მეთოდის ეფექტურობა, კერძოდ 600dpi-ზე სკანირებული სიმბოლოებისთვის უშეცდომო - (100%) ამოცნობა, ხოლო 300dpi-ზე – 99.91% საიმედობა.

ლიტერატურა:

1. ვერულავა ო., ხუროძე რ. ამომცნობი სისტემების თეორიის საფუძვლები. სტუ, თბ., 2001

2. ვერულავა ო., ირემაძე ი., თოდუა თ. სახეთა რეალიზაციების პრეპარირება მეზობელი პიქსელების მეთოდით. საერთაშ.სამეც.კონფ. „ინფორმაციული ტექნოლოგიები 2008”. სტუ, თბ., 2008

3. თოდუა თ. ქართული ნაბეჭდი სიმბოლოების წინასწარი კომპიუტერული დამუშავება. ელ-სამეც.ჟურნ. „კომპიუტერული მეცნიერებები და ტექნოლოგიები". http://gesj.internet-academy.org.ge №1, 2004

4. Khurodze R., Verulava O., Todua T. Some aspects of preparation of machine printed symbols. Proceedings of the 7th world multiconference on Systemics, Cybernetics and Informatics, Orlando, Florida. http://www.iiisci.org/sci2003/

5. Mohamed Cheriet, Nawwaf Kharma, Cheng-Lin Liu, Ching Y. Suen. Character Recognition Systems A guide for Students and Practitioners. A John Wiley & Sons, Inc. 2007

THE ANALYSIS AND RECOGNITION OF THE GEORGIAN SYMBOLS OF "SYLFAEN" FONT

Verulava Otar, Tsverikmazashvili Zurab
Georgiam Technical University

Summary

This work covers the problem of choosing the method of software analysis for converting scanned image into text format. Criterions for choosing a way of the analysis are preparation processes particularly: using scaling and segmentation for convenience of forming pattern descriptions and etalons. The method of recognition is generated for the Georgian symbols of a font "Sylfaen" with Unicode encoding, which includes 74 symbols: Georgian alphabet, Arabian figures, punctuation marks, brackets of different kind and some mathematical symbols. For the procedures set forth above are generated algorithms and corresponding source codes in programming language C++.

АНАЛИЗ И РАСПОЗНАВАНИЕ ГРУЗИНСКИХ СИМВОЛОВ ШРИФТА «SYLFAEN»

Верулава О., Цверикмазашвили З.

Резюме

Рассмотрена задача анализа печатного грузинского шрифта, при этом для перевода графического изображения в текстовый формат разработан метод программного анализа. Критерием выбора способа анализа являются процессы препарирования: масштабирования и сегментация для удобства составления эталонных описаний образов. Сформирован метод распознавания для грузинских символов шрифта «Sylfaen» с кодировкой Юникод. Количество распознаваемых образов составляет 74 символов: грузинский алфавит, арабские цифры, знаки препинания, разные виды скобок, некоторые математические символы. Для вышеперечисленных процедур сформированы алгоритмы и соответствующие программные модули на языке программирования C++.