Elektrotehniški vestnik 77(2-3): 149–154, 2010
Electrotechnical Review, Ljubljana, Slovenija

Računalniško branje padavinskih grafov

Gašper Derganc, Peter Peer

Univerza v Ljubljani, Fakulteta za računalništvo in informatiko, Tržaška 25, 1000 Ljubljana, Slovenija
E-pošta: gasper.derganc@gmail.com, peter.peer@fri.uni-lj.si

Povzetek. V članku je predstavljena metoda za avtomatično detekcijo in digitalizacijo krivulje padavin v padavinskih grafih s papirnatih trakov, ki se uporabljajo v avtomatskih merilnih postajah. Metoda sestoji iz več korakov. Na digitalni sliki grafa padavin se krivulja padavin loči od ozadja. S sodelovanjem metod drsečega povprečja in sledenja roba krivulje se krivulja detektira in edinstveno določi – vsakemu stolpcu slike ustreza natanko ena točka. Ta detektirana krivulja je vhod v proces izdelave natančnega časovnega zaporedja padavin. Poleg postopkov analize slik se metoda opira tudi na mehanske značilnosti merilnega instrumenta. Natančno časovno zaporedje padavin je potrebno za nadaljnje analize padavinskih dogodkov, kot so klasifikacija, analiza ekstremnih dogodkov, kalibracija modelov odtoka površinskih voda, napovedovanje meteoroloških pojavov in pri številnih raziskovalnih projektih. Algoritem je bil preizkušen na 58 slikah pluviografskih trakov. Primerjava med rezultati, pridobljenimi z opisanim algoritmom, ter uradnimi podatki z Agencije Republike Slovenije za okolje je pokazala, da algoritem večinoma zelo natančno določi potek krivulje in s tem natančno časovno zaporedje padavin. Tako bi bil zelo primeren, kot jedro sistema, za digitalizacijo padavinskih podatkov, ki bi prek grafičnega vmesnika omogočal branje z optičnim čitalnikom, ogled rezultatov avtomatične detekcije in morebitne popravke.

Ključne besede: računalniški vid, digitalizacija, pluviograf, meteorologija, padavine

Automatic pluviograph strip chart reading

Extended abstract. An algorithm aimed at automatic detection and digitalization of the rainfall signal recorded by the float based rain gauges on paper strip charts (Fig. 1) is presented. The algorithm consists of several steps that gradually lead to the desired goal. The rainfall signal is extracted from the digital image of the strip chart. By using the moving average method (Fig. 3) and curve edge following method (Fig. 2) the rainfall curve is detected and uniquely determined. In each image column there is one single point representing the rainfall curve plotline. From the curve plotline a high-resolution rainfall time series is obtained. Besides image analysis techniques in the design of the algorithm, the mechanical features of the recording instrument were taken into consideration. The availability of high resolution rainfall time series is required in many applications, including rainfall classification, analysis of extreme rainfall events, calibration of raifall-runoff models, weather prediction models and many research projects. The algorithm was tested on 58 pluviograph strip chart images. A comparison between the data obtained with the proposed algorithm and the official data from the Environmental Agency of the Republic of Slovenia shows that the algorithm usually accurately detects the rainfall curve and consequently an accurate rainfall time series is obtained (Tab. 2). Since it is not always 100 % reliable, it should be used as a component of a system that would enable inspection of the detected curve and when required, it should also enable interactive changing of the parts needing correction.

Key words: computer vision, digitalisation, pluviograph, meteorology, rainfall

Prejet 21. oktober, 2009
Odobren 5. januar, 2010

1 Uvod

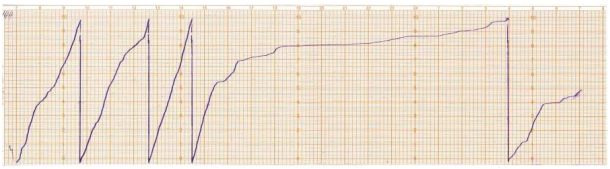
Količino padavin merimo kot višino vodnega stolpca, ki se akumulira na določeni horizontalni površini v določenem časovnem intervalu. Navadno je izražena v milimetrih, kjer $1\ \mathrm{mm}$ padavin ustreza $1\ \mathrm{kg/m^2}$ oziroma, povedano drugače, če zlijemo $1\ \mathrm{kg}$ vode po površini enega kvadratnega metra, bo višina vodne plasti enaka enemu milimetru. Padavine merimo s pomočjo ročnih (pluviometri) ali avtomatskih (pluviografi) instrumentov. Pluviometri v nasprotju z pluviografi ne podajajo spreminjanja količine padavin v času [1]. Rezultat meritev s pluviografom je diskretna funkcija intenzitete padavin v odvisnosti od časa.

Intenziteta padavin je predstavljena kot količina padavin P na enoto časa t:

$$\mathrm{intenziteta} = \frac{dP}{dt} \approx \frac{\Delta P}{\Delta t} \qquad (1)$$

Danes se podatki o padavinah ponavadi zbirajo z avtomatskimi digitalnimi pluviografi, katerih podatki v digitalni obliki so takoj na voljo za nadaljnjo obdelavo.

Pred uvedbo digitalnih pluviografov so pluviografi zapisovali rezultate na papirnate trakove (slika 1) velikosti 422×113 mm. V Sloveniji trenutno deluje 38 merilnih mest s pluviografi s plovcem, ki tako zapisujejo svoje meritve. Ti trakovi na osi x prikazujejo čas (24 ur – od



Slika 1. Trak pluviografa
Figure 1. Pluviograph strip chart

7:00 do 7:00) ter na osi y podatek o polnosti posode (od 0 do $10\ \mathrm{mm}$). Ko se posoda napolni, se izprazni, kar je na traku vidno kot hiter padec krivulje do vrednosti 0 po osi y. Podatki s teh postaj so koristni za preverjanje podatkov z digitalnih merilnih postaj ter natančnejšo analizo padavin iz preteklih obdobij, ko se avtomatske digitalne merilne naprave še niso uporabljale. Koristni pa so tudi zaradi svoje robustnosti, saj so digitalni pluviografi nagnjeni k napakam prav pri ekstremnih vremenskih pojavih, ki so za meteorologe in hidrologe najzanimivejši.

Tudi podatke na grafih je treba digitalizirati. Ta postopek se izvaja s pomočjo digitalizatorske table ter zahteva veliko zbranosti in natančnosti ter je časovno zelo zahteven. Na pluviografskem traku, položenem na tablo, se označijo robovi območja zanimanja ter določi potek krivulje z označevanjem točk krivulje. Iz pridobljenega zaporedja točk, katerih vmesne vrednosti so določene z linearno interpolacijo, se s pomočjo programa izračuna intenziteta padavin. Za takšno obdelavo mesečnih podatkov s posamezne merilne postaje strokovnjak potrebuje od 10 minut do 1 ure, odvisno od količine padavin, zabeležene na trakovih.

Pridobitev natančnejših padavinskih časovnih zaporedij visoke resolucije s starejših pluviografskih trakov bi omogočilo boljši vpogled v preteklo padavinsko dogajanje. Trenutni postopek digitalizacije je zelo zamuden in monoton. V tem članku opisani postopek bi ga lahko občutno pospešil ter morda tudi povečal natančnost tako pridobljenih podatkov. Za digitalizacijo pluviografskih trakov bi potrebovali le branje z optičnim čitalnikom in preprosto popravljanje napak.

Pri implementaciji postopka sta bili uporabljeni dve neodvisni množici slik trakov pluviografov. Implementacija algoritma je bila izvedena s pomočjo učne množice. Učna množica vsebuje osem trakov z merilne postaje Kal nad Kanalom. Za testiranje je bilo uporabljenih 58 slik trakov z merilne postaje Podkraj, ki so naključno izbrane izmed slik trakov iz leta 2006. Za analizo rezultatov so bili na voljo podatki o dnevnih, urnih, polurnih in 5-minutnih intenzitetah padavin z merilne postaje Podkraj, kot so jih zabeležili na Agenciji Republike Slovenije za okolje.

2 Sorodno delo

S podobno tematiko so se ukvarjali na univerzi Cagliari v Italiji [2]. Njihov sistem je namenjen digitalizaciji padavinskih podatkov s pluviografov z zlivajočima se posodicama.

Bistveni koraki postopka so: predprocesiranje, segmentacija, avtomatična detekcija signala in interaktivno postprocesiranje.

Vhod v postopek je digitalna slika papirnatega traku z ločljivostjo 300 DPI. Predprocesiranje se izvaja nad digitalno sliko traku pluviografa in poskrbi, da je krivuljo mogoče predstaviti v kartezičnem koordinatnem sistemu, kjer točke slike z enako vrednostjo abscise ustrezajo istemu časovnemu trenutku.

Korak segmentacije vsebuje upragovanje komponente R barvnega prostora RGB vhodne slike. Nato pa se izvede še nehierarhično rojenje (angl. nonhierarchical cluster analysis) v prostoru barv HSV, katerega rezultat sta dva razreda. Prvi vsebuje slikovne elemente krivulje, drugi pa slikovne elemente, ki so posledica pisanja s svinčnikom in niso zanimivi za nadaljnjo obdelavo.

V koraku avtomatične detekcije signala se iz segmentirane slike enolično določi potek krivulje. Korak je sestavljen iz petih postopkov. To so: robustna detekcija krivulje, zavračanje madežev, omejitev na monotona zaporedja, popravki in prilagoditve detektirane krivulje, iskanje točk, kjer se smer gibanja pisala obrne.

Zaradi grobe podobnosti problemov je podobna tudi osnovna zgadba postopkov. Razlike so posledica različnih vhodnih trakov in različnega načina pisanja pluviografov na trakove.

Bistvene razlike med našimi trakovi in trakovi s pluviografa z zlivajočima se posodicama so:
• Prekinjena krivulja – pisalo se vertikalno pomika po diskretnih intervalih, ki so določeni s prostornino posamezne posodice.
• Os y pri trakovih s pluviografov z zlivajočima posodicama ne pomeni polnosti posode (ni praznjenja). Ko pisalo doseže rob, le spremeni smer pomikov.
• Opazna ukrivljenost skale in krivulje, ki je posledica krožnega gibanja pisala.

V tem članku opisani algoritem v koraku segmentacije pretvori sliko v prostor barv CIELAB ter nad posameznimi komponentami slike določa točke slike, ki pripadajo krivulji s postopkom rasti regij. Algoritem v [2] izvaja upragovanje v prostoru barv RGB, nato pa še nehierarhično rojenje v prostoru barv HSV. Uporabljena metoda robustne detekcije v [2] je podobna metodi dvostopenjskega drsečega povprečenja, ki se v našem algoritmu poleg metode sledenja roba krivulje uporablja v koraku detekcije. Sledenje roba krivulje pri trakovih s pluviografov z zlivajočima se posodicama (zaradi prekinitev krivulje) ne bi bilo primerno.

3 Algoritem za avtomatsko branje padavinskih grafov

Osnovna ideja algoritma je naslednja:

1. Loči krivuljo od ozadja – določi, katere točke slike pripadajo krivulji (Segmentacija).
2. Določi natančen potek krivulje – zaporedje točk, ki pokrivajo celotno dolžino traku z natanko eno točko na stolpec (Detekcija krivulje).
3. Iz koordinat točk izračunaj količino vode v zbiralni posodi in iz razlik teh količin sosednjih točk določi intenziteto padavin.

Vhod v algoritem je slika v formatu JPG, PNG ali BMP. Barvna paleta slike je RGB z barvno globino 24 bitov na slikovni element. V učni in testni množici slik je uporabljena ločljivost 300 DPI, ki je nekakšen kompromis med natančnostjo in časom izvajanja algoritma. Ločljivost ni nujno fiksna, saj se ji algoritem lahko prilagodi z nastavitvijo parametrov. Pomembo je le, da ta ni prenizka, saj se z zmanjševanjem ločljivosti manjša tudi količina informacije slike.

Za pravilno delovanje mora vhodna slika izpolnjevati določene pogoje:
• Slika je poravnana. Vse slikovne točke stolpca i ustrezajo istemu časovnemu intervalu. Poleg rotirane slike je vzrok za neporavnanost lahko tudi nenatančno vstavljen ali po vstavitvi premaknjen trak.
• Na grafu je le ena krivulja (pri dneh brez dežja se namreč včasih uporabi isti trak, kar se pokaže v več krivuljah).
• Uporabljeno je modro črnilo pisala.

3.1 Segmentacija

Namen segmentacije je čim bolje ločiti zapisano krivuljo od ozadja. Kot rezultat dobimo binarno sliko, kjer imajo slikovni elementi, prepoznani kot del krivulje, vrednost 1, preostali pa vrednost 0. Zaradi značilnosti ozadja in barve črnila je primerna uporaba nelinearnega prostora barv CIELAB [3, 7], saj sta barvi ozadja in črnila v tem prostoru laže ločljivi – bolj oddaljeni.

Izbira slikovnih elementov, ki pripadajo krivulji, je izvedena s postopkom upragovanja. Upragovanje je implementirano kot rast regij (angl. region growing) [6]. Kot seme so izbrani vsi slikovni elementi, ki presegajo visok prag, katerega naj bi ga dosegali le slikovni elementi krivulje. Kot kriterij za nadaljnjo rast se uporablja spodnji prag in standardno odstopanje soseščine trenutnega slikovnega elementa. Tako se v nasprotju z navadnim upragovanjem, ki izbere vse slikovne elemente, ki so po vrednosti med dvema pragoma, upošteva tudi sosednost. Zmanjša se možnost napačne detekcije, zaradi uporabe standardnega odstopanja pa upošteva tudi lokalne značilnosti krivulje, katere intenziteta lahko variira. Tako se poveča tudi natančnost. Mogoča pa je tudi uporaba z zamenjanima vlogama pragov – regija se začne pri vrednosti manjši od praga za začetek in izbira nadaljnje slikovne elemente z vrednostmi, večjimi od drugega praga.

Enačbi, uporabljeni pri upragovanju, sta:
Povprečje vrednosti regije:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \tag{2}$$

kjer je x_i sivinska vrednost i-tega slikovnega elementa regije, ki vsebuje n slikovnih elementov.
Standardno odstopanje:

$$\sigma = \sqrt{\frac{\sum_{j=1}^{n}(x_j - \bar{x})^2}{n-1}}, \tag{3}$$

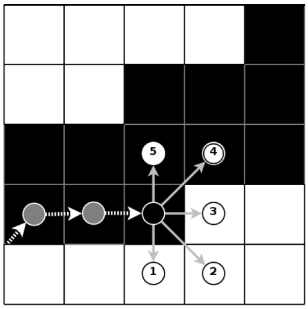
kjer je x_j sivinska vrednost j-tega slikovnega elementa in n število vseh slikovnih elementov $K \times K$ velike okolice trenutnega semena.

Omenjeni postopek se uporabi na sivinskih slikah komponent a^* in b^* barvnega prostora CIELAB. Ko govorimo o sivinskih slikah imamo seveda v mislih slike, kjer so vrednosti slikovnih elementov razporejene med 0 in 255. Na sivinski sliki komponente a^* se rast regije začne pri slikovnem elementu, ki ima vrednost večjo od 160, spodnjo mejo pa določa vrednost 150. Pri komponenti b^* se rast regije začne pri slikovnem elementu z vrednostjo, manjšo od 110. Spodnja meja pa je določena z vrednostjo 126. Za velikost okolice je izbrana vrednost $K = 3$. Vrednosti so določene kot posledica empiričnih izkušenj, pridobljenih s pomočjo učne množice slik.

3.2 Detekcija krivulje

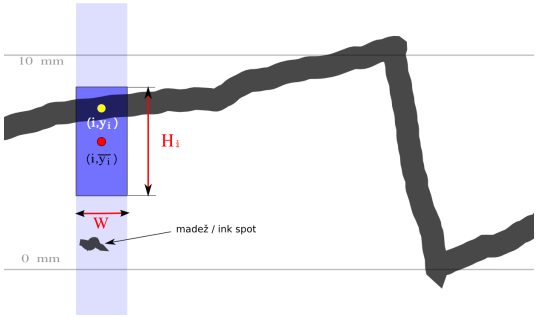
Rezultat segmentacije je binarna matrika $\mathbf{A}$, kjer je vrednost elementa a_{ij} ($i \in [1, M]$ in $j \in [1, N]$ pri velikosti slike $M \times N$ slikovnih elementov) enaka 1, če gre za slikovni element na krivulji in 0, če je to slikovni element ozadja. Zaradi debeline zaznane krivulje in morebitnih madežev je v stolpcu j ponavadi več kot en slikovni element z vrednostjo 1. Da lahko natančno izračunamo intenziteto padavin, pa potrebujemo natančno določeno krivuljo – le en slikovni element na stolpec slike. To dosežemo s sodelovanjem dveh metod, ki se izmenjujeta.

Lokalno sledenje roba krivulje hitro in natančno določi potek krivulje tam, kjer je ta povezana. Mogoče so napake, če je krivulja razmazana. Naslednji slikovni element se izbira v zaporedju, prikazanem na sliki 2. S takšnim zaporedjem izbire korakov zagotovimo, da so rezultirajoče točke vedno na spodnjem robu krivulje. Uporabljeno je sestopanje: če ne more naprej, sestopi za največ 10 slikovnih elementov. Če kljub sestopanju ni napredk,a pa se postopek ustavi ter delo prepusti koraku povprečenja.



Slika 2. Simbolični prikaz sledenja roba krivulje
Figure 2. Edge following

Globalno dvostopenjsko drseče povprečenje računa



Slika 3. Dvostopenjsko povprečenje
Figure 3. Two-phase averaging

povprečno vrednost ordinate prek več sosednjih stolpcev (slika 3). Je veliko manj dovzetno za nepravilnosti, vendar tudi manj natančno določa krivuljo. Za vsak stolpec i v prvem koraku izračunamo začetno vrednost $\bar{y}_i$ kot drseče povprečje (angl. moving average) y vrednosti detektiranih slikovnih elementov prek več sosednjih stolpcev:

$$\bar{y}_i = \frac{1}{\bar{N}_i} \sum_{j=1}^{M} \sum_{k=i-\lfloor \frac{W}{2} \rfloor}^{i+\lfloor \frac{W}{2} \rfloor} a_{jk} \cdot j, \tag{4}$$

kjer je

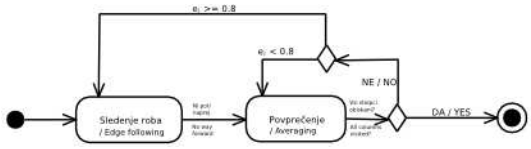
$$\bar{N}_i = \sum_{j=1}^{M} \sum_{k=i-\lfloor \frac{W}{2} \rfloor}^{i+\lfloor \frac{W}{2} \rfloor} a_{jk}. \tag{5}$$

$\bar{N}_i$ je število vseh detektiranih slikovnih elementov v oknu lihe širine W in višine M s središčem v i-tem stolpcu. V našem primeru smo uporabili $W = 7$, kar je približno enako debelini zapisane krivulje. Če je število $\bar{N}_i = 0$, je vrednost $\bar{y}_i$ označena kot neznana. Kljub vsemu lahko madeži povprečje odpeljejo daleč od prave vrednosti krivulje. Zato v drugem koraku iz prvega koraka pridobljeno vrednost uporabimo le kot začetni približek, ki določa središče novega, manjšega okna enake širine. Namen drugega koraka je čim tesneje zaobjeti detektirane slikovne elemente okna. Končni približek je nato določen kot:

$$y_i = \frac{1}{N_i} \sum_{j=\bar{y}_i-\lfloor \frac{H_i}{2} \rfloor}^{\bar{y}_i+\lfloor \frac{H_i}{2} \rfloor} \sum_{k=i-\lfloor \frac{W}{2} \rfloor}^{i+\lfloor \frac{W}{2} \rfloor} a_{jk} \cdot j, \tag{6}$$

kjer je N_i število vseh detektiranih slikovnih elementov v oknu s središčem v točki $(i, \bar{y}_i)$ širine W ter višine H_i. Višina H_i je odvisna od števila detektiranih slikovnih elementov v začetnem oknu in se izračuna po enačbi (7).

$$H_i = \max(2 \cdot W, \frac{\bar{N}_i}{W}) \tag{7}$$



Slika 4. Pomen koeficienta zaupanja
Figure 4. Curve extraction depending on the evidence factor

Ko imamo določen približek, ki naj bi določal krivuljo v stolpcu i, lahko iz razmerja $e_i = N_i/\bar{N}_i$ sklepamo, koliko lahko temu približku zaupamo. Na sliki 4 je prikazan vpliv v stolpcu i izračunanega koeficienta zaupanja e_i na preklapljanje med metodama povprečenja in metodo sledenja roba. Če je koeficient zaupanja v stolpcu i premajhen ($< 0{,}6$), določimo vrednost detektirane krvulje v tem stolpcu kot neznano. Z uporabo dveh korakov in koeficienta zaupanja pri zmerni količini šuma (madežev) lahko natančno določimo lego krivulje. Pri veliki količini madežev pa kljub vsemu lahko pride do napačne detekcije.

S kombinacijo prej opisanih metod ne pridobimo nujno zaporedja z natanko enim slikovnim elementom na stolpec slike. Pridobljeno zaporedje bi moralo ustrezati določenim omejitvam (odsekoma naraščajoče zaporedje), kar pa pri tako dobljenih rezultatih ni nujno res. Do tega nas pripeljejo še trije koraki, ki odpravijo morebitne presledke, nastale v koraku povprečenja, in večje število slikovnih elementov v posameznem stolpcu, ki so lahko posledica sledenja roba krivulje. V teh korakih obdelujemo le prej pridobljeno zaporedje točk in ne same slike traku. Ti koraki so:

- Zavračanje točk z madežev. – Odkrijemo jih kot velike skoke zaporedja v kratkem časovnem intervalu.
- Omejitev na naraščajoče funkcije. – Potek krivulje bi moral biti določen z odsekoma naraščajočim zaporedjem, kjer so odseki ločeni v točkah inverzij. Na vsakem odseku določimo optimalni potek krivulje z uporabo algoritma za iskanje najdaljšega

naraščajočega podzaporedja [4]. Pred tem poiščemo točke inverzij kot strme skoke zaporedja z zgornjega dela slike traku do spodnjega.

- Natanko ena točka na stolpec slike. – V stolpcih, kjer je v zaporedju več elementov, pridobimo le enega tako, da mu določimo vrednost ordinate, kot povprečje vrednosti ordinat teh elementov. Vrzeli (stolpce z nedoločenimi/neznanimi vrednostimi) zapolnimo z linearno interpolacijo.

Rezultat detekcije krivulje je odsekoma naraščajoče zaporedje točk (koordinat slikovnih elementov) z natanko eno točko na stolpec območja zanimanja slike.

3.3 Izračun intenzitete padavin

Dobljeno odsekoma naraščajoče zaporedje (dolžine n) pretvorimo v kumulativno zaporedje razlik ordinatnih vrednosti med točkami. Razlika ordinatnih vrednosti dveh zaporednih elementov zaporedja pomeni količino padavin v času, določenem s širino slikovnega elementa.

Za pretvorbo v slikovnih elementih podane intenzitete padavin v nam želeno mero [mm / čas] pa potrebujemo le še podatek o razmerju med velikostjo slikovnega elementa in milimetri in časom, označenim na skali papirja. Najlaže to storimo tako, da poiščemo območje zanimanja – območje, v mejah katerega naj bi se nahajala krivulja. To programsko preprosto določimo tako, da na sliki poiščemo mejne vrednosti, označene na skali papirja. S poznavanjem mer in položaja območja zanimanja lahko izračunamo intenziteto padavin za poljuben časovni interval.

Razmerje $\Omega_1 = 10\ \mathrm{mm}/$višina območja zanimanja (razmerje med količino padavin v posodi ob praznjenju posode in številom slikovnih elementov v stolpcu območja zanimanja) določa padavinsko ločljivost – najmanjšo intenziteto padavin, ki jo lahko zabeležimo. Razmerje $\Omega_2 = 1440\,\mathrm{min}/$širina področja zanimanja določa časovno ločljivost – najmanjši mogoči časovni interval. Konstanta 1440 je število minut v 24 urah.

4 Rezultati

4.1 Kvalitativna ocena

Pri kvalitativni oceni po ogledu detektirane krivulje rezultatu pripišemo natančnost/pravilnost glede na vnaprej določene razrede. Na sliki, ki nam jo prikaže program, primerjamo potek krivulje z detektiranim potekom, ki je prikazan na isti sliki, in poiščemo stolpec, v katerem je vertikalno odstopanje največje – to odstopanje nato določi razred napake. Uporabljeno merilo so mm, ki na traku označujejo polnost posode pluviografa ter omogočajo preprosto odčitavanje napake. To merilo nima neposredne povezave z intenziteto padavin, manjka mu namreč časovna komponenta (interval). Razredi:

- **Razred 0**: Ni opaznih napak ali odstopanj od poteka krivulje ali pa so ta odstopanja manjša od $0{,}2\ \mathrm{mm}$ (želimo čim več elementov v tem razredu).
- **Razred 1**: Prišlo je do manjših odstopanj od krivulje. Največje opaženo odstopanje od pravilne lege krivulje je večje od $0{,}2\ \mathrm{mm}$ in manjše od $0{,}5\ \mathrm{mm}$.
- **Razred 2**: Prišlo je do večjih odstopanj od krivulje. Največje opaženo odstopanje od pravilne lege krivulje je večje od $0{,}5\ \mathrm{mm}$ in manjše od $10\ \mathrm{mm}$.
- **Razred 3**: Prišlo je do napak, ki vodijo v napačen izračun celodnevne količine padavin za več kot $10\ \mathrm{mm}$ (npr. napačno detektirana inverzija).

Razred 0	Razred 1	Razred 2	Razred 3
44	6	3	5

Tabela 1. Rezultati kvalitativnega testiranja
Table 1. Qualitative testing results

Iz tabele 1 je razvidno, da v $75{,}8$ % primerov (razred 0) algoritem sam zelo natančno določi potek krivulje in posledično izračuna natančno časovno zaporedje intenzitet padavin.

4.2 Kvantitativna ocena

Pri kvantitativni oceni smo vrednosti, pridobljene z našim algoritmom, primerjali z vrednostmi, ki jih hrani Agencija Republike Slovenije za okolje. Vrednosti Agencije Republike Slovenije za okolje so pridobljene s pomočjo digitalizatorske table po postopku, opisanem v uvodu.

Pri kvantitativni oceni napake nismo upoštevali primerov, ki pripadajo razredu 3 (niso reprezentativni, saj bi ti primeri nujno potrebovali ročno popravljanje) ter primerov slik pluviografskih trakov, katerih podatki za primerjavo niso primerni (dnevi z več kot 5 cm zapadlega snega in dnevi, ko so zabeležili le topljenje snega). Pri snegu je postopek merjenja količine padavin kompleksnejši – treba je upoštevati tako količino padavin, izmerjeno na traku, kot tudi zapadli sneg, ki pa se ne stopi takoj. Tako podatki, zabeleženi na trakovih, ne odražajo pravega časovnega poteka padavin in se lahko kvantitativno močno razlikujejo od dejansko zabeleženih. Pri podatkih z urno, polurno in 5-minutno ločljivostjo so vrednosti nekaterih intervalov manjkale, kar je dodatno zmanjšalo število primerjav. Napako smo ocenjevali z merami, ki se tipično uporabljajo za ocenjevanje natančnosti/primernosti regresijske krivulje. Srednjo absolutno napako (MAE) in relativno srednjo absolutno napako (RMAE) [5] uporabimo zato, ker lahko tudi na ta

problem gledamo kot na nekakšno iskanje funkcije, ki se čim bolje prilega podanim točkam.

Srednja absolutna napaka (MAE) [5]:

$$\mathrm{MAE} = \frac{1}{I}\sum_{i=1}^{I} |f(i) - \hat{f}(i)| \; [mm] \tag{8}$$

Relativna srednja absolutna napaka (RMAE) [5]:

$$\mathrm{RMAE} = \frac{N \cdot \mathrm{MAE}}{\sum_{i=1}^{N} |f(i) - \bar{f}|} \tag{9}$$

V enačbah I pomeni število intervalov, ki smo jih primerjali, $f(i)$ je vrednost, zabeležena na Agenciji Republike Slovenije za okolje, $\hat{f}(i)$ pa izračunano vrednost (za i-ti interval). Povprečno vrednost $\bar{f}(i)$, uporabljeno v enačbi (9), pa smo izračunali po enačbi $\bar{f} = \frac{1}{N}\sum_{i=1}^{N} f(i)$.

Mera MAE torej pokaže, za koliko mm se izračunani podatki povprečno razlikujejo od uradnih podatkov – poda povprečoa absolutno razliko med primerjanimi podatki. Mera RMAE pa prikazuje napako relativno – glede na dejanski mogoči razpon vrednosti funkcije $f(i)$. Vrednost RMAE je v tem primeru vedno med 0 in 1, kjer 0 pomeni popolno ujemanje podatkov.

Časovna ločljivost	MAE [mm]	RMAE	I
dan	0,3844	0,0321	45
ura	0,1350	0,0960	394
pol ure	0,1055	0,1379	788
5 minut	0,0563	0,3958	4728

Tabela 2. Rezultati kvantitativnega testiranja
Table 2. Quantitative testing results

Rezultati so predstavljeni v tabeli 2). Iz tabele je razvidno, da je srednja absolutna napaka (MAE) razmeroma majhna in se pričakovano zmanjšuje s krajšanjem časovnega intervala. Pri analizi dobljenih rezultatov je dobro vedeti, da se ti lahko razlikujejo tudi zaradi človeškega faktorja – trak je lahko zamenjan pozneje kot ob 7:00 ali pa je nenatančno vstavljen.

Testiranje se je izvajalo na osebnem računalniku z dvojedrnim procesorjem, na katerem teče operacijski sistem Linux. Za obdelavo testne množice (58 slik) je postopek potreboval 5 minut in 6 sekund. Za povprečno obdelavo posamezne slike traku je torej porabil približno 5,28 sekunde. Za celotno obdelavo pa je treba temu času prišteti še čas branja z optičnim čitalnikom in morebiten čas interaktivnega procesiranja.

Rezultat testiranja na celodnevni ($\mathrm{MAE} = 0{,}3844$ mm) ter urni ($\mathrm{MAE} = 0{,}1350$ mm) časovni ločljivosti je zelo dober. Pri višji časovni ločljivosti so odstopanja precej velika, kar pa je pričakovano, saj se načina pridobitve podatkov zelo razlikujeta. Medtem ko algoritem določi točko krivulje v vsakem stolpcu slikovnih elementov slike, so točke pri digitalizaciji z digitalizatorsko tablo izbrane manj na gosto. Tako je potek krivulje, ki ga izračuna algoritem, pri pravilni detekciji krivulje natančneje določen. Čeprav v tem poglavju govorimo o napakah, ta izraz ni povsem na mestu. Gre bolj za primerjavo dveh metod, ki sta obe izpostavljeni lastnim napakam.

5 Sklep

Glede na rezultate lahko ocenimo algoritem kot dober. Po kvalitativni oceni v 75,8 % primerov dobimo rezultat, ki je primeren za takojšnje shranjevanje. Rezultati pa kažejo tudi, da v vseh primerih ne more zadovoljivo določiti intenzitet padavin, to je zaradi raznolikosti vseh mogočih dejavnikov zelo težko dosegljivo. Primeren je kot jedro sistema, ki prikaže detektiran potek krivulje in ob pravilni detekciji omogoča potrditev ter shranjevanje rezultatov, v nasprotnem primeru pa tudi ročno popravljanje. Sistem bi lahko vključeval tudi korak branja z optičnim čitalnikom. Tako bi lahko poskrbeli, da so nastavitve optičnega čitalnika vedno enake, in avtomatično poravnali sliko. Z uporabo takšnega orodja bi se postopek digitalizacije precej olajšal in pospešil.

Podoben pristop bi bilo z manjšimi spremembami algoritma mogoče uporabiti tudi za digitalizacijo higrografov in termografov.

6 Literatura

[1] Mitja Brilly, Mojca Šraj, *Osnove hidrologije* - 1. izd, Ljubljana, Fakulteta za gradbeništvo in geodezijo, 2005.

[2] Roberto Deidda, Giuseppe Mascaro, Enrico Piga, Giorgio Querzoli, "An automatic system for rainfall signal recognition from tipping bucket gage strip charts", *Journal of Hydrology*, 333, str. 400-412, 2007.

[3] David A. Forsyth, Jean Ponce, *Computer Vision - A Modern Approach*, Prentice Hall, 2002.

[4] Dan Gusfield, *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*, New York, Cambridge University Press, poglavje 12.5, 1997.

[5] Igor Kononenko, *Strojno učenje*, Ljubljana, Fakulteta za računalništvo in informatiko, str. 51-52, 1997.

[6] Ashidi N. Mat-Isa, Yusof M. Mashor, Hayati N. Othman, "Seeded Region Growing Features Extraction Algorithm; Its Potential Use in Improving Screening for Cervical Cancer", *International Journal of The Computer, the Internet and Management*, 13(1), str. 61-70, 2005.

[7] Wikipedia CIELAB: http://en.wikipedia.org/wiki/Lab_color_space (1.9.2009).

Gašper Derganc je diplomiral leta 2009 na univerzitetnem študiju Fakultete za računalništvo in informatiko Univerze v Ljubljani.

Peter Peer je docent na Fakulteti za računalništvo in informatiko Univerze v Ljubljani.