ACTA UNIVERSITATIS AGRICULTURAE ET SILVICULTURAE MENDELIANAE BRUNENSIS

Ročník LVIII 42 Číslo 6, 2010

EXPERIMENTÁLNÍ STANOVENÍ PARAMETRŮ VYBRANÝCH PRVKŮ DOKUMENTŮ Z RASTROVÝCH GRAFICKÝCH ZDROJŮ

J. Rybička, D. Kelnarová, P. Talandová

Došlo: 31. srpna 2010

Abstract

RYBIČKA, J., KELNAROVÁ, D., TALANDOVÁ, P.: *Experimental determination of chosen document elements parameters from raster graphics sources.* Acta univ. agric. et silvic. Mendel. Brun., 2010, LVIII, No. 6, pp. 421–432

Visual appearance of documents and their formal quality is considered to be as important as the content quality. Formal and typographical quality of documents can be evaluated by an automated system that processes raster images of documents. A document is described by a formal model that treats a page as an object and also as a set of elements, whereas page elements include text and graphic object. All elements are described by their parameters depending on elements' type. For future evaluation, mainly text objects are important. This paper describes the experimental determination of chosen document elements parameters from raster images. Techniques for image processing are used, where an image is represented as a matrix of dots and parameter values are extracted. Algorithms for parameter extraction from raster images were designed and were aimed mainly at typographical parameters like indentation, alignment, font size or spacing. Algorithms were tested on a set of 100 images of paragraphs or pages and provide very good results. Extracted parameters can be directly used for typographical quality evaluation.

raster image, recognition, document, paragraph, typography, text objects parameters

Formální stránka úpravy dokumentů je důležitá, a je proto vhodné provádět její kontrolu. Při zpracování dokumentů je kladen důraz nejen na obsahovou, ale i formální stránku dokumentu, která by rovněž měla podléhat kontrole a hodnocení kvality. Vzhledem k množství dokumentů a časté potřebě kontroly je vhodné tuto činnost automatizovat. Pro hodnocení byl zaveden hodnoticí systém, systém kritérií a model, který je na dokument aplikován a podle nějž se dokument hodnotí (Talandová, 2009; Talandová a Rybička, 2009).

Pro vlastní realizaci hodnocení je nezbytné vhodně zvolit metodu pro získání informací o prvcích dokumentu a jejich parametrech, které definuje model. Při hodnocení dokumentů lze vycházet z jejich struktury nebo z vizuální stránky. Strukturní popis dokumentu usnadňuje proces identifikace prvků a získání jejich parametrů. Vizuální podoba je však pro většinu dokumentů lépe dostupná a při posuzování vzhledu a při hodnocení z typografického hlediska je podstatně vhodnější. Jako vstupní formát dokumentů byl zvolen rastrový obraz.

Cílem tohoto článku tedy je navrhnout vhodnou metodu, která dokáže nalézt prvky stránky vstupního dokumentu, získat hodnoty jejich parametrů a klasifikovat prvky na základě zjištěných hodnot. Součástí metody je i experimentální stanovení hodnot parametrů u textových prvků dokumentu. Výsledkem aplikace této metody na vstupní dokument je reprezentace informací o jednotlivých prvcích dokumentu tak, aby mohla být uplatněna při hodnocení z typografického hlediska.

PŘEHLED LITERATURY A SOUČASNÉHO STAVU

Pro oblast rozpoznávání obrazu se využívají metody a techniky, které se liší náročností, hloubkou analýzy, charakterem zpracování dat, popř. i zaměřením na konkrétní aplikační oblast. Zpracování rastrového obrazu je dlouhodobě používaná technika s ustálenou posloupností kroků (snímání/digitalizace, předzpracování, segmentace, klasifikace a rozpoznávání objektů). Přehled těchto kroků a používaných metod uvádí Kelnarová (2010).

Metody uzpůsobené práci s dokumenty jsou součástí systémů pro optické rozpoznávání znaků (Eikvil, 1993). Proces optického rozpoznávání znaků zahrnuje snímání, digitalizaci a uložení dat, předzpracování, segmentaci, extrakci příznaků, klasifikaci a rozpoznávání a post-processing. Z hlediska rozpoznávání jsou významné fáze segmentace, extrakce a post-processingu. Segmentace v případě dokumentů zahrnuje rozdělení stránky na textové a netextové, případně i nepotištěné oblasti. U oblastí se zjišťuje barevné uspořádání, které se statisticky hodnotí. Výsledkem jsou informace o tom, zda oblast obsahuje text, nebo grafický objekt.

Beitzel a kol. (2003) se věnují přehledu metod, které se používají pro vylepšení efektivnosti získávání informací procesem OCR. Poukazují na řadu metod, které se zaměřují na způsob získávání informací z OCR a opravu chyb. Přehled metod pro segmentaci obrazu uvádějí také Španěl a Beran (2006). Zaměřují se na techniky detekce hran, metody založené na regionech, statistické, hybridní a znalostní metody. Přehled je však orientován do oblasti medicínských dat, nikoli zpracování textu. Odhad vhodných hodnot parametrů pro techniky binarizace dokumentů řeší Badekas a Papamarkos (2009).

Moderní přístupy k segmentaci a analýze dokumentů jsou častěji zaměřeny na oblast webových stránek. Cai a kol. (2003) zkoumají vizuální strukturu webové stránky, využívají matematického popisu rozdělení stránky na objekty a zavádějí popis hierarchické struktury stránky a heuristická pravidla pro segmentaci. Kunc a Burget (2008) se zabývají vizuální segmentací webových stránek a následnou klasifikací detekovaných oblastí na základě jejich vizuálních vlastností. Segmentace je založena na analýze stránky zdola nahoru společně s analýzou vizuálně významných prvků. Klasifikace využívá informace o vzájemné poloze oblastí a jejich vizuálních vlastnostech. Cao a kol. (2010) navrhují metodu segmentace webových stránek s užitím technik zmenšování obrazu a rozdělováním na části, nezávislou na kódu HTML. Výchozím formátem je webová stránka uložená jako rastrový obraz, pro předzpracování se používá detekce hran. Obraz je rozdělen na části, které se zmenšují, a oba kroky se opakují. Metoda je vhodná pro zjišťování podobnosti rozložení stránky (používá se při detekci phishingových stránek).

MATERIÁL A METODY

Návrh metody pro identifikaci a klasifikaci prvků dokumentu využívá metody a techniky z oblasti zpracování a rozpoznávání obrazu (segmentace, popis a klasifikace objektů) včetně metod používaných pro optické rozpoznávání znaků. Dále se vychází z typografických zvyklostí a ze souvisejícího formálního modelu dokumentu.

Komponenty navrhované metody

Navrhovaná metoda je tvořena pěti samostatnými komponentami, které lze řešit nezávisle na sobě. Je určeno rozhraní mezi komponentami a komponenty tvoří tento lineární řetězec (Kelnarová, 2010), viz též obr. 1:

1. *Definice prvků, jejich typů a parametrů* – odpovídá formální definici dokumentu (Talandová, 2009; Kelnarová, 2010). Model popisuje dokument na úrovni stránky a jejích prvků, přičemž pro stránku i každý typ prvku jsou v souladu s typografickými pravidly definovány parametry. Parametry se používají pro popis a hodnocení typografické kvality těchto prvků a tím i hodnocení kvality dokumentu.
2. *Segmentace stránky* – rozdělení stránky na prvky, které jsou v tomto kroku již atomické.
3. *Předklasifikace* – představuje mezikrok, ve kterém se oblasti získané v předchozí segmentaci rozdělí na textové a netextové (grafické).
4. *Získání charakteristik prvků a klasifikace* – představuje klíčovou etapu. V první části se provede výběr a jsou získány vhodné parametry prvků. Ve druhé části se na základě zjištěných údajů provede klasifikace prvků – zařazení do odpovídající třídy.
5. *Získání požadovaných parametrů prvků* – na základě typu prvku dané třídy se získají hodnoty jeho parametrů.

Definice prvků, jejich typů a parametrů

Obecně lze na stránce rozeznat následující tři typy objektů. Lze předpokládat, že všechny objekty lze převést na obdélníkové oblasti.

- *Textové prvky* (*T*) jsou objekty obsahující text, který je obvykle uspořádán do řádků. Mohou obsahovat znaky, číslice, interpunkci, speciální symboly a drobné grafické prvky, jako jsou odrážky u seznamů či odkazovací symboly u poznámek pod čarou.
- *Grafické objekty* (*G*) zahrnují obrázky, tabulky, kresby, fotografie, grafy, linky apod. Mohou obsahovat i malé úseky textu, avšak ty jsou v daném kontextu považovány za součást grafického objektu.
- *Bílá místa* mohou být rovněž považována za prvky stránky. Zde se však jedná pouze o význačná bílá místa, jako jsou okraje stránky nebo mezislovní či meziřádkové mezery.

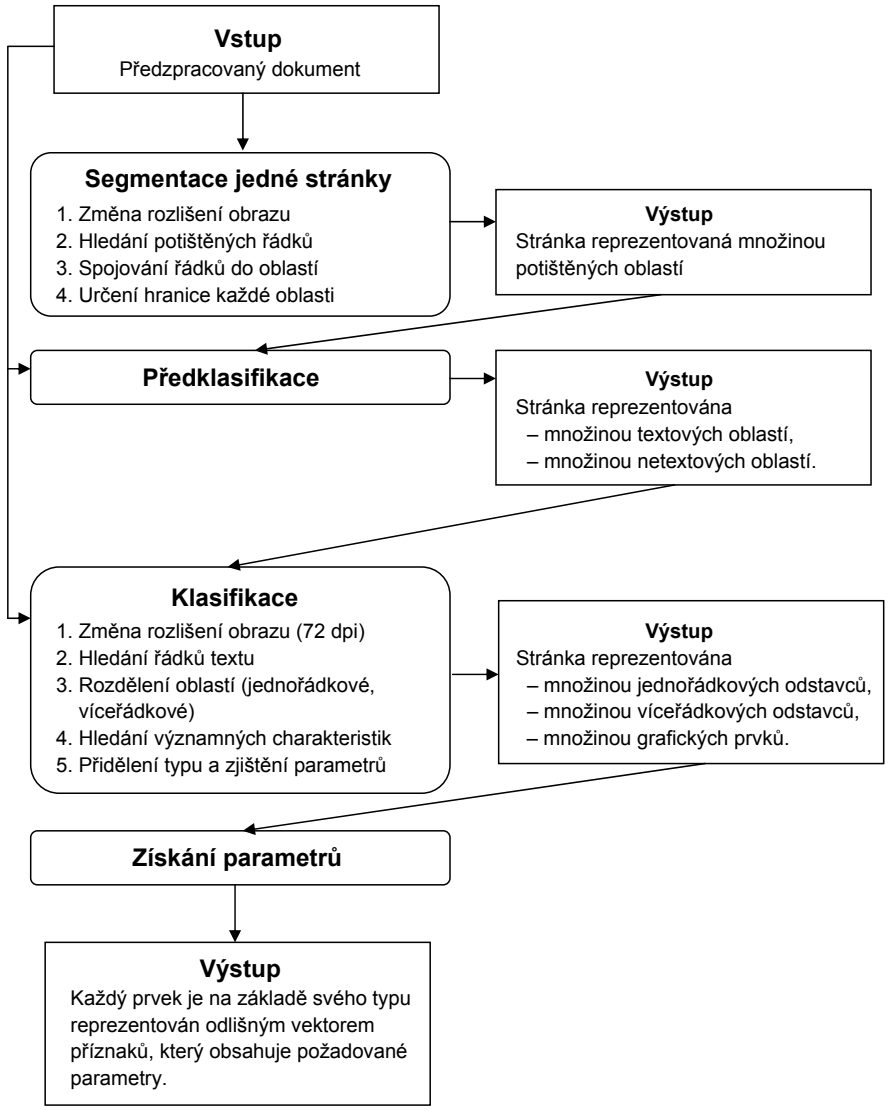
Textové prvky jsou v rámci navržené metody reprezentovány odstavci. Odstavce lze rozdělit na jednořádkové (F) a víceřádkové (E).

Parametry jednořádkových odstavců

Následující výčet uvádí parametry a jejich možné hodnoty, označení vychází ze zmíněné práce Talandové (2009).

- *Umístění* (φ) – určuje polohu prvku na stránce. Jeho obdélníková hranice je definována čtyřmi body A, B, C a D.

  První bod $A = [x, y]$ určuje polohu levého horního rohu prvku, bod $B = [x', y]$ označuje pravý horní roh, bod $C = [x', y']$ určuje pravý dolní roh a bod $D = [x, y']$ určuje levý dolní roh. Pro stanovení umístění pak stačí pouze čtyři souřadnice x, x', y, y', které jsou udávány v milimetrech.
- *Výška* (α) a *šířka* (β) – odvozeny ze souřadnic bodů definujících výše popsanou hranici prvku. Přípustnými hodnotami jsou všechna kladná čísla z množiny desetinných čísel. Tyto dva parametry jsou udávány v milimetrech (mm).
- *Plocha* (γ) – lze ji odvodit z výšky a šířky odstavce. Její jednotkou je čtverečný milimetr (mm^2).
- *Stupeň šedi* (δ) – vyjadřuje poměr černých a bílých pixelů, které tvoří plochu prvku, udává se v procentech. Množina přípustných hodnot zahrnuje kladná reálná čísla. Pro odstavce byla empiricky zjištěna průměrná hodnota 25 % (Talandová, 2009).
- *Zarovnání* (ω) – definuje horizontální umístění odstavce vzhledem k sazebnímu obrazci. Může nabývat hodnot: *na prapor vlevo*, *na prapor vpravo*, *na střed*, *do bloku* nebo *nedefinováno*. Hodnota *nedefinováno* vyjadřuje možnost, že umístění odstavce neodpovídá žádné z předešlých možností.



1: *Schéma postupu získávání parametrů z rastrových zdrojů – komponenty navržené metody* (*Kelnarová, 2010*)

1: *Parameters obtaining from raster resources procedure scheme – components of the proposed method*

- *Odsazení* (ϕ) – pokud jsou odstavce odděleny odsazením, hodnotou tohoto parametru bude kladné reálné číslo udávající velikost odsazení například v milimetrech.

Parametry víceřádkových odstavců

Výčet uvádí parametry víceřádkových odstavců, představující rozšíření parametrů oproti jednořádkovým odstavcům.

- *Zarovnání* (ω) – zde se liší množina přípustných hodnot tohoto parametru. Zarovnání víceřádkových odstavců může nabývat hodnot: *na prapor vlevo*, *na prapor vpravo*, *na střed* nebo *do bloku*.
- *Počet řádků* (η) – udává se jako celé kladné číslo.
- *Zarážka* (ζ) – může být indikátorem oddělení odstavců. Pokud by byla dodržena typografická pravidla, budou odstavce odděleny buď zarážkou, nebo odsazením. Hodnotou zarážky může být 0 (zarážka neexistuje) nebo kladné reálné číslo, které vyjadřuje její délku v milimetrech.
- *Délka východového řádku* (ρ) – vyjadřuje délku posledního řádku v odstavci v milimetrech.
- *Mezislovní mezery* (ψ) – tento parametr souvisí zejména s čitelností textu a vyjadřuje skutečnost, že mezi slovy v textu byly nalezeny mezery, jejichž velikost lze označit jako příliš velkou. Pokud je hodnota parametru 0, v odstavci nebyly nalezeny tyto nevhodné mezery. V opačném případě bude parametr nabývat hodnoty 1 a bude navíc odkazovat na další parametr, který se týká pozice nalezených mezer.
  - *Umístění mezery* (ψ_φ) – v případě, že v odstavci byly identifikovány nevhodné mezery, bude tento parametr obsahovat souřadnice jejich pozice na stránce, tj. souřadnice x, x', y, y'.

Typy grafických objektů a jejich parametry

Zkoumání a identifikace typů grafických objektů není významné pro další zpracování. Všechny grafické objekty patří do jedné třídy (G). Pro vytvoření modelu dokumentu je však nezbytné sledovat také u grafických prvků určité parametry analogické textovým prvkům (umístění, výška, šířka, plocha a stupeň šedi).

Parametry stránky

Stránka (S) může být z hlediska víceúrovňové struktury dokumentu považována za jeho prvek. Navíc je ze stejného hlediska považována za objekt, který je nadřazen jednotlivým prvkům stránky. Nejen z těchto důvodů se u stránky sledují následující parametry:

- *Výška* (α) a *šířka* (β) stránky – udává se v milimetrech a může nabývat hodnot z množiny kladných reálných čísel. Levý horní roh stránky reprezentuje počátek souřadnicového systému.
- *Plocha* (S_γ) *stránky* – udávána ve čtverečných milimetrech a odvozuje se z výšky a šířky stránky.
- *Rozměry okrajů* (M) – je obvyklé, že každá stránka má čtyři okraje – horní (M_t), levý (M_l), dolní (M_b) a pravý (M_r). Rozměry jsou udávány v milimetrech, přičemž u horního a dolního okraje se sleduje výška a u levého a pravého okraje šířka. Přípustné hodnoty jsou opět kladná reálná čísla.
- *Počet prvků na stránce* (r) – může nabývat hodnot z množiny celých kladných čísel včetně 0 (prázdná stránka).
- *Stupeň šedi* (δ) – je ovlivněn počtem a velikostí všech prvků na stránce. Vyjadřuje poměr černých a bílých pixelů na celé stránce udávaný v procentech.

Segmentace stránky

Vstup této fáze tvoří jednotlivé stránky digitalizovaného dokumentu v rastrovém formátu (například PNG nebo BMP).

Předpokládá se, že jsou předem známy informace o *rozlišení obrazu*, o *barevné hloubce* a *formátu stránky* (A4, A5, B5 atd.). Zdrojový obraz je již předzpracován (zbaven šumu, správně orientován).

Pro další zpracování je provedena úprava rozlišení na experimentálně stanovenou hodnotu 10 dpi (zaniknou drobné nepřesnosti v ohraničení objektů) a převod na monochromatický obraz. V tomto upraveném obrazu se pak hledají oblasti, určují jejich hranice a klasifikují typy (Kelnarová, 2010).

Výstupem celého procesu segmentace je množina potištěných oblastí. Oblast je pak reprezentována souřadnicemi čtyř bodů tvořících její hranici, přičemž tyto souřadnice jsou vyjádřeny v milimetrech.

Objekty na stránce tvoří množinu S:

$$S = \{O_1, \ldots, O_r\} \tag{1}$$

Každý objekt je určen čtyřmi body

$$O_j = (A_j, B_j, C_j, D_j) \tag{2}$$

pro $j = 1, \ldots, r$, kde r je počet oblastí na stránce, $A_j = [x_j, y_j]$, $B_j = [x'_j, y_j]$, $C_j = [x'_j, y'_j]$, $D_j = [x_j, y'_j]$, a dále platí

$$O_i \cap O_j = \emptyset \; \forall \; i, j \in \{1, \ldots, r\}, i = j. \tag{3}$$

Následným krokem je předklasifikace rozdělující některou zvolenou technikou OCR množinu objektů na dvě disjunktní podmnožiny textových objektů a grafických objektů:

$$S = \{T_1, \ldots, T_s\} \cup \{G_1, \ldots, G_t\}. \tag{4}$$

Přitom platí:

$$T = \{F_1, \ldots F_p\} \cup \{E_1, \ldots E_q\}) \wedge (F \cap E) = \emptyset \tag{5}$$

a

$$s = p + q. \tag{6}$$

Dále platí:

$$(T_i \cap T_j = \emptyset \; \forall \; \mathrm{i}, \mathrm{j} \in \{1, \ldots, s\}, i \neq j) \wedge (G_i \cap G_j = \emptyset \; \forall \; i, j \in \{1, \ldots, t\}, i \neq j) \tag{7}$$

a

$$r = s + t. \quad (8)$$

Tato klasifikace je prováděna v původním obrazu s původním rozlišením a v barevné hloubce 1 bit (monochromatický obraz).

Získání významných charakteristik prvků a klasifikace

V dalším zpracování se budeme zabývat pouze podmnožinou textových prvků – odstavců. Vstupními daty jsou textové objekty a jejich souřadnice, podmnožinu grafických objektů nebudeme dále uvažovat. Lze tedy psát:

$$S' = \{T_1, \ldots, T_s\}, \quad (9)$$

kde

$$T_j = (A_j, B_j, C_j, D_j), j \in \{1, \ldots, s\}, \quad (10)$$

přičemž $A_j = [x_j, y_j]$, $B_j = [x'_j, y_j]$, $C_j = [x'_j, y'_j]$, $D_j = [x_j, y'_j]$.

Vstupní obraz je nutné předzpracovat podobně jako při zahájení segmentace stránky, tentokrát však jinou úpravou rozlišení – 72 dpi. V tomto rozlišení odpovídá pixel jednomu typografickému bodu anglo-amerického typografického systému. Vzhledem ke skutečnosti, že převážná většina softwaru pro zpracování textů pochází ideově z americké oblasti, vnitřně tedy používá tento měrný systém. Stránku lze v této souvislosti chápat jako čtvercovou mřížku o velikosti strany elementárního čtverce 1 typografický bod (1 pt). Souřadnice lze pak snadno vyjádřit jako celočíselné indexy této mřížky.

Po úpravě rozlišení se dále provede převod do monochromatické barevné hloubky.

Pro další kroky klasifikace je nezbytné určit rozměry okrajů a sazebního obrazce. Vychází se z rozměru zdrojového obrazu – výška (α) a šířka (β).

Na základě hranic textových a netextových oblastí se nejdříve získají souřadnice jednotlivých okrajů stránky, z nichž se v dalším kroku odvodí souřadnice sazebního obrazce. Velikost stránky, kterou ze vstupu získáme v milimetrech, lze převést rovněž na typografické body (pt) přepočtem

$$t = \frac{\alpha}{0{,}353}, \quad (11)$$

$$c = \frac{\beta}{0{,}353}, \quad (12)$$

přičemž platí, že 1 pt = 0,353 mm. Pro snadnější vyjádření souřadnic dále označíme zápisem $x[A]$ souřadnici x bodu A, resp. $y[A]$ souřadnici y bodu A.

Obdélník vymezující horní okraj M_t je dán souřadnicemi

$$A_t = [1, 1], B_t = [c, 1], C_t = [c, y_t - 1], D_t = [1, y_t - 1], \quad (13)$$

kde y_t je minimální souřadnice y všech objektů stránky, tedy

$$y_t = \min y[Aj], j = 1, \ldots, r. \quad (14)$$

Analogicky získáme dolní okraj M_b se souřadnicemi

$$A_b = [1, y'_b + 1], B_b = [c, y'_b + 1], C_b = [c, t], D_b = [1, t], \quad (15)$$

kde $y'_b = \max y[C_j], j = 1, \ldots, r$, dále levý okraj M_l se souřadnicemi

$$A_l = [1, 1], B_l = [x_l - 1, 1], C_l = [x_l - 1, t], D_l = [1, t], \quad (16)$$

kde $x_l = \min x[A_j], j = 1, \ldots, r$ a pravý okraj M_r se souřadnicemi

$$A_r = [x + 1, 1], B_r = [c, 1], C_r = [c, t], D_r = [x + 1, t], \quad (17)$$

kde $x_r = \max x[B_j], j = 1, \ldots, r$.

Sazební obrazec je pak dán souřadnicemi

$$A = [x_l, y_t], B = [x', y_t], C = [x', y'_b], D = [x_l, y'_b]. \quad (18)$$

V dalším postupu navrhuje Kelnarová (2010) klasifikaci objektů pomocí neuronových sítí. Domníváme se však, že tuto klasifikaci můžeme doplnit experimentálními postupy a tak vhodně stanovit statistické charakteristiky, z nichž provedeme segmentaci prvků a stanovení vybraných vlastností.

Experimentální materiál

Pro účely analýzy bylo vytvořeno třicet grafických souborů ve formátu BMP, v rozlišení 72 dpi a barevné hloubce 1 bit. Formát BMP byl zvolen jako nejvhodnější z hlediska jednoduché struktury nekomprimovaného souboru, která umožňuje jeho snadné zpracování. Tyto grafické soubory byly získány z vysázeného textu ruční extrakcí jednotlivých odstavců nebo celých stránek.

Text byl vysázen písmem Times nebo Computer Modern stupně 10 pt s řádkováním 12 pt, jeden ze vzorků byl vysazen písmem Computer Modern Sans Serif. Z výsledné sazby ve formátu PDF byly jednotlivé stránky převedeny programem GSView 4.8 do obrazů ve formátu BMP, rozlišení 72 dpi a barevné hloubce 1 bit, následně zpracovány programem Corel Photopaint X3 – byly vyříznuty jednotlivé odstavce nebo stránková zrcadla s textem. Tímto krokem byla simulována stránková segmentace a klasifikace textových objektů podle Kelnarové (2010).

Deset souborů obsahuje celou stránku (skupinu odstavců) zarovnanou do bloku, dvacet souborů obsahuje celou stránku zarovnanou na prapor.

Takto získané soubory byly zpracovány vlastním programem, který převede rastrový obraz na textovou podobu, kde každý znak představuje jeden pixel původního obrazu, znakem tečka je indikován bílý pixel, znakem „o“ je indikován černý pixel. Výsledný tvar je možné vizuálně kontrolovat na shodu s původním obrazem a slouží také pro snazší orientaci v systému vzorků.

Program rovněž vyčíslí počty černých pixelů v jednotlivých řádcích a sloupcích celého obrazu a tyto hodnoty upraví pro další zpracování.

Označíme P_h počet černých pixelů v řádku obrazu (v horizontálním směru) a P_v počet černých pixelů ve sloupci (ve vertikálním směru).

Pro vizualizaci zaplnění jednotlivých řádků nebo sloupců černými pixely byly hodnoty zpracovány v tabulkovém procesoru Excel verze 2003 do podoby odpovídajících grafů.

VÝSLEDKY

Na základě zpracovaných hodnot budou řešeny další kroky zpracování prvků a získávání jejich parametrů – jedná se o tyto prvky, na nichž bude možné metodu demonstrovat:

1. segmentace řádků odstavce,
2. zjištění řádkování textu,
3. segmentace odstavců,
4. zjištění odstavcové zarážky,
5. zjištění zarovnání odstavce,
6. zjištění stupně šedi.

Segmentace řádků odstavce, zjištění řádkování textu

Segmentaci řádků odstavce lze provést na základě četnosti černých pixelů v horizontálním směru (P_h). Data pro tuto analýzu jsou přímo dostupná z podpůrného programového vybavení, které dodává hodnotu P_h pro každý řádek vstupního rastrového obrazu.

Předpokládáme-li sazbu podle typografických pravidel, tedy s nenulovým prokladem, lze usoudit, že obraz bude obsahovat pixelové řádky s $P_h = 0$. V tomto ideálním případě je možné segmentaci provést na základě pozic těchto prázdných pixelových řádků, které představují meziřádkové mezery a oddělují tak řádky odstavce. Z toho je možné určit počet řádků v odstavci (η).

V praktických případech však mohou nastat i jiné situace – překrytí elementů dvou po sobě jdoucích řádků. Toto překrytí způsobuje výskyt nenulové hodnoty četnosti pixelů, je tedy potřebné určit alespoň minimální hodnotu u.

Z experimentálních dat vyplývá, že jsou-li překrývajícími elementy dolní dotahy horního řádku a verzálkové akcenty dolního řádku, pohybuje se hodnota u kolem jednotek pixelů (1–2 % z $\max(P_h)$). Plně tedy dostačuje nalézt lokální minimum počtu pixelů v jednotlivých řádcích a nalezneme řádkovou hranici, pro kterou platí, že $P_h \leq u$.

Tento argument je prezentován grafy na obr. 2, v němž jsou shrnuty absolutní počty pixelů náhodně vybraných stran vysázených v řádkování 12 pt písmem stupně 10 pt. Levý graf je výsledkem zpracování stránky vysázené bezserifovým písmem, pravý graf obsahuje zpracované hodnoty identického textu sázeného serifovým písmem téhož stupně. Z grafů je vidět, že z hlediska řádkové segmentace nevykazují podstatné rozdíly.

Lze si ovšem představit těsnou sazbu, kdy se dolní dotahy horního řádku dotýkají horních dotahů (verzálek, akcentů) dolního řádku. Zde vzhledem k existenci serifů může dojít k situaci, kdy rozhraní řádků vykazuje dokonce lokální extrém. Metoda v tomto případě řádkovou segmentaci neprovede správně. Pro úspěšnou segmentaci musíme využít informaci o řádkování z jiných zdrojů nebo z jiného místa analýzy dokumentu.

Segmentace odstavců, určení odstavcové zarážky

Jednotlivé textové objekty stránky jsou segmentovány na základě zpracování monochromatického rastrového obrazu v rozlišení 10 dpi. Z toho vyplývá, že bude-li rozpoznán objekt a následnou analýzou OCR bude zjištěno, že se jedná o textový objekt, může obsahovat buď jeden odstavec, jsou-li odstavce odděleny vertikálním odsazením ϕ, nebo více odstavců, je-li provedena sazba bez vertikálních meziodstavcových mezer. Z toho důvodu je pro další analýzu nutné zjistit, kolik odstavců obsahuje příslušný textový objekt.

V segmentaci je potřebné vycházet z typografických metod vizuálního oddělení odstavců – odstavcovou zarážkou (ζ), případně rozměrem východového řádku (ρ). Z toho vyplývá, že v objektu potřebujeme znát pozice jednotlivých řádků, které byly získány z předchozího kroku. Dále je vhodné znát stupeň písma σ (v experimentálních datech 10 pt).

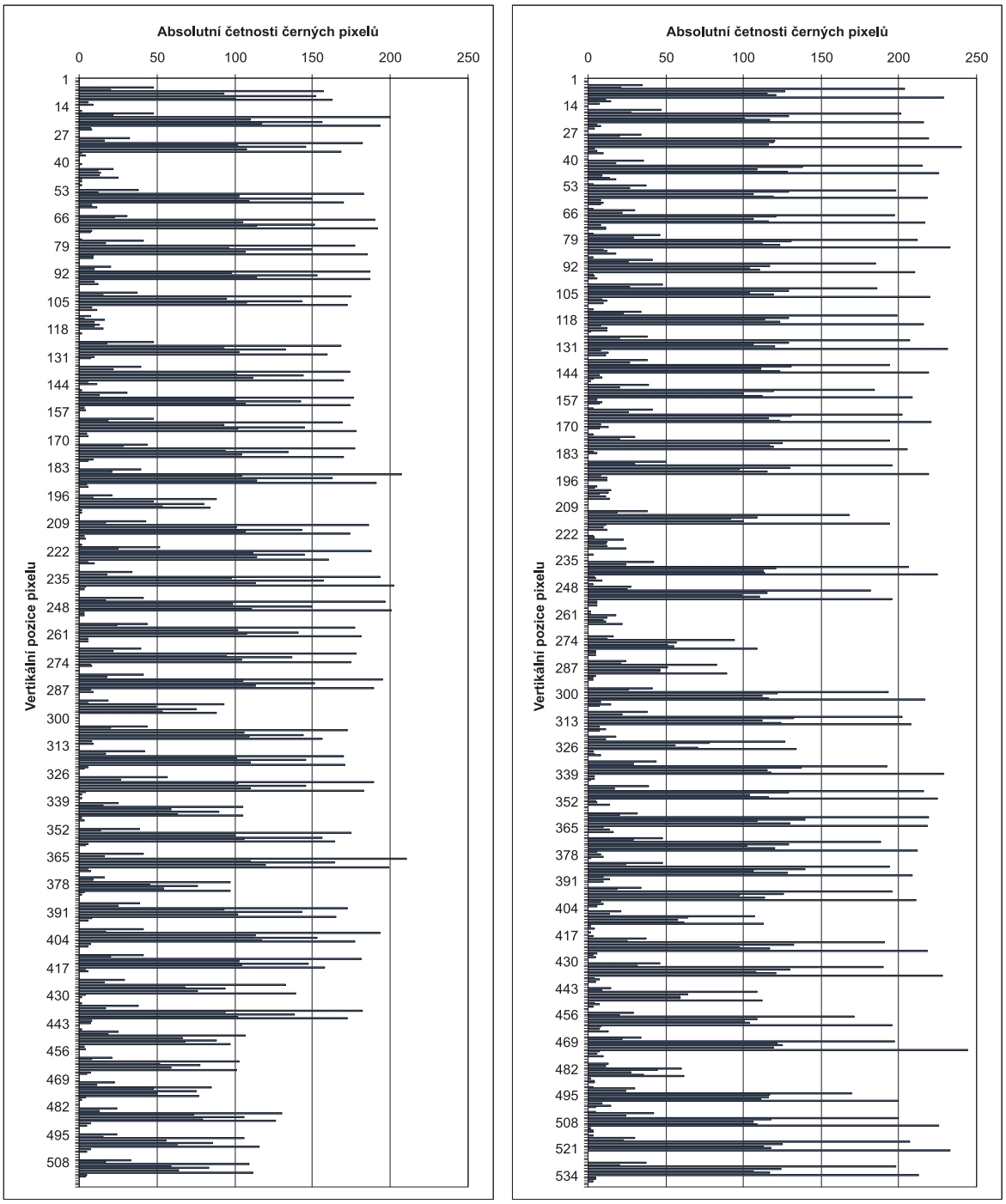
U každého řádku je potřebné určit velikost bílého místa na začátku a na konci. Podle typografických pravidel by měla být zarážka u kratších řádků o velikosti stupně písma ($\zeta = \sigma$), velikost bílého místa východového řádku nemusí podléhat žádným standardním rozměrům, východový řádek tedy může mít plnou šíři. V případě, že odstavce nejsou vizuálně odlišeny zarážkou, musí však východový řádek mít minimálně dva čtverčíky bílého místa na konci ($\beta - \rho \geq 2\sigma$). Na základě těchto pravidel lze tedy určit řádky, jejichž bílé místo na začátku nebo na konci činí nejméně jeden čtverčík.

Čtverčík je relativní jednotka a určení její velikosti je spojeno se stupněm písma σ. Máme-li však segmentovány řádky, pak jejich výška odpovídá přibližně velikosti čtverčíku. Znamená to, že zarážkový řádek zjistíme podle bílého místa odpovídajícího počtem pixelů přibližně výšce řádku. Obdobně lze detekovat dostatečné bílé místo na konci řádku.

Tyto metody detekce zarážkového a východového řádku nelze aplikovat na případy, kdy odstavce nejsou zarovnány do bloku.

Zjištění zarovnání odstavce

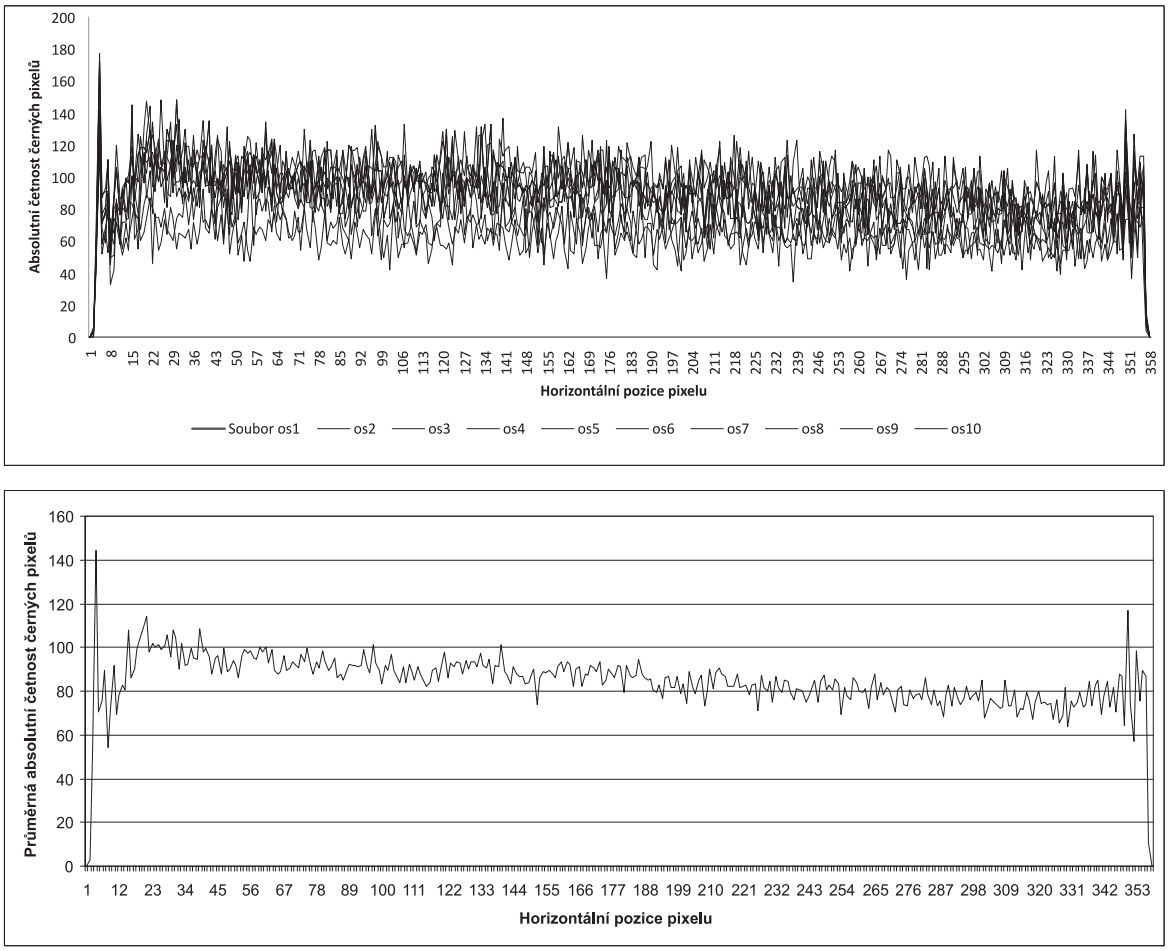
Pro zjištění zarovnání odstavce (ω) bude potřebné zpracovat počty černých pixelů ve vertikálním směru (P_v). Experimentální data ilustruje graf na obr. 3 pro odstavce zarovnané do bloku a graf na obr. 4 pro odstavce zarovnané na prapor (zarovnán je levý okraj). Data reprezentují celostránkové textové objekty. Průběh četností pixelů P_v na okrajích je u odstavců zarovnaných do bloku vyrovnaný,



2: *Grafy absolutních četností černých pixelů na vybrané stránce v horizontálním směru. Graf vlevo je výsledkem zpracování textu s bezserifovým písmem, graf vpravo je výsledkem zpracování identického textu sázeného serifovým písmem téhož stupně.*
2: *Graphs of the absolute frequency of black pixels on a chosen page in horizontal direction. Graph on the left is the result of sans-serif text processing, the graph on the right is the result of serif text processing, the text being the same*

zatímco při praporovém okraji klesá četnost plynule z ploché středové části k nule.

Velikost pásma, kde dochází k poklesu četností vertikálně počítaných pixelů, závisí na způsobu sazby a charakteru textu. Probíhá-li sazba podle typografických pravidel, je při sazbě na prapor vkládána neměnná mezislovní mezera ψ, slova nejsou dělena. Praporové pásmo tedy závisí na délce slov. Je-li v textu velké množství relativně dlouhých slov, zvětšuje se šířka tohoto pásma. V běžných případech lze považovat střední délku slova kolem pěti až šesti znaků, v tom případě je (opět za jinak standardní kresby písma a standardních dalších podmínek) praporové pásmo o velikosti přibližně tři čtverčíky. Známe-li tedy z analýzy řádkování přibližnou



3: *Ilustrativní graf četností černých pixelů vybraných stránek s odstavci zarovnanými do bloku* (*nahoře*) *a průběh průměrné četnosti černých pixelů vhodný pro analýzu*
3: *Illustrative graph of the frequency of black pixels of chosen pages with justified paragraphs* (*top*) *and the average frequency of black pixels suitable for analysis*

velikost čtverčíku σ, můžeme tuto informaci využít pro analýzu okrajových oblastí odstavců.

V souhrnném grafu absolutních četností získaných z dvaceti celostránkových textů je vyznačeno trojčtverčíkové praporové pásmo. Vizuálně lze usoudit, že hraniční čára tohoto pásma splňuje předpoklad odvozený v předcházejícím odstavci. Od této čáry směrem k pravému okraji začíná pokles četností černých pixelů.

Pro explicitní stanovení způsobu zarovnání na prapor můžeme použít integrál křivky průměrných četností černých pixelů v pravé části (například v pravé polovině) praporového pásma. Protože je však křivka f stanovena diskrétně, lze k numerické integraci využít následujícího vztahu:

$$I_{\mathrm{p}} = \frac{f(A) - f(B)}{2} + \sum_{i=2}^{N} f(x_i), \tag{19}$$

kde bodem A je dolní mez sledovaného pásma, tedy vzdálenost přibližně 1,5 čtverčíku od pravého okraje, bod B leží na pravém okraji a hodnota N odpovídá velikosti sledovaného pásma v počtech typografických bodů. Získanou hodnotu I_{p} porovnáme s integrálem stejně veliké oblasti v polovině řádku I_{c}. Pro průkazné zjištění typu zarovnání na prapor lze experimentálně ověřit, že platí

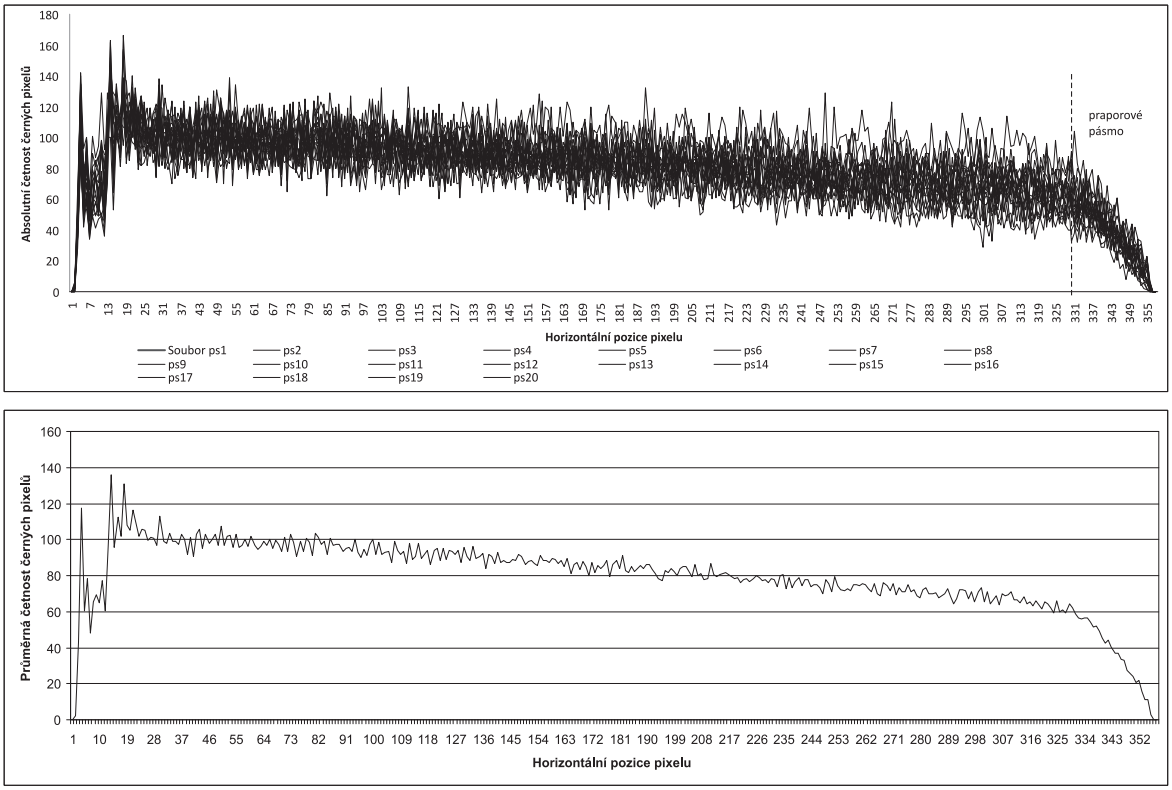
$$I_{\mathrm{p}} < \frac{1}{3} I_{\mathrm{c}}. \tag{20}$$

Zjištění stupně šedi

Jednou z charakteristik vyrovnané odstavcové sazby je i stupeň šedi (δ). Podpůrný program tuto hodnotu počítá jako poměr celkového počtu černých a bílých pixelů pro každý odstavec. Odchyluje-li se hodnota významně od běžných hodnot nebo od hodnot okolních odstavců, může jít buď o záměr (např. řádek nadpisu), nebo o chybu v sazbě. Tyto případy je třeba rozlišit analýzou dalších parametrů.

Výstupní informace

Výstupní informace vzniklá z procesu získávání parametrů z rastrových zdrojů a použitelná pro návazný systém hodnocení typografické kvality sazby zahrnuje prvky získané segmentací stránky, jejich parametry, parametry stránky a jejích okrajů.



4: *Ilustrativní graf četností černých pixelů vybraných stránek s odstavci sázenými na prapor (nahoře) a průběh průměrné četnosti černých pixelů vhodný pro analýzu*
4: *Illustrative graph of the frequency of black pixels of chosen pages with ragged-right paragraphs (top) and the average frequency of black pixels suitable for analysis*

$$S = F \cup E \cup G \cup M, \tag{21}$$

$$S = (\alpha_i, \beta_i, S_{\gamma_i}, M_i, \delta_i), \tag{22}$$

$$F_j = (\varphi_j, \alpha_j, \beta_j, \gamma_j, \delta_j, \omega_j, \phi_j), \tag{23}$$

$$E_k = (\varphi_k, \alpha_k, \beta_k, \gamma_k, \delta_k, \omega_k, \phi_k, \eta_k, \zeta_k, \rho_k, \psi_k, \psi_{\varphi_k}), \tag{24}$$

$$G_l = (\varphi_l, \alpha_l, \beta_l, \gamma_l, \delta_l). \tag{25}$$

DISKUSE A ZÁVĚR

Navržený způsob experimentálního stanovení parametrů vybraných prvků dokumentů z rastrových grafických zdrojů závisí na několika faktorech. Důležitá je kvalita vstupních obrazových dat a na ní závisející úspěšnost předzpracování, následné segmentace obrazu i fáze předklasifikace.

Při segmentaci bylo experimentálně zjištěno, že postačující hodnota rozlišení obrazu je 10 dpi. Při tomto rozlišení zaniknou drobné nepřesnosti, které by ztěžovaly zpracování, a obraz pak lze snadněji segmentovat. Při předklasifikaci lze na základě zjištěných rozměrů a pozic některou metodou OCR z původního zdroje rozlišit textové a grafické objekty. V tomto rozlišení se ztrácejí rozdíly mezi jednotlivými typy písma. Metodu tedy lze použít bez ohledu na zvolené písmo a lze ji aplikovat nejen na serifová a bezserifová latinková písma, ale i na nelatinková hlásková písma (např. cyrilice, azbuka).

Při následné fázi získávání parametrů a klasifikace je třeba opět změnit rozlišení, hodnota 10 dpi je příliš nízká. Zvolená hodnota 72 dpi je velmi výhodná při zpracování hodnot parametrů, kde pixel odpovídá jednomu typografickému bodu.

Další zpracování vychází z reprezentace obrazu odstavce formou matice černých a bílých pixelů. Tato podoba nejlépe odpovídá chápání odstavce, který je považován za obdélník s přesně definovanými možnými tvarově nepravidelnými částmi (zarážka, východový řádek, zarovnání na prapor). Zároveň zvolená podoba odráží způsob vnímání odstavce jako bílé plochy pokryté relativně rovnoměrně černými body. Hodnocení vycházející z rastrového obrazu tak má před strukturním popisem dokumentu zřejmou výhodu, ačkoli strukturní popis může obsahovat podrobnější informace, zejména o typech prvků a některých jejich parametrech. Ukazuje se rovněž, že získávání hodnot některých parametrů lze provést přímou analýzou četností černých pixelů, přičemž není nutné pro klasifikaci využívat neuronových sítí.

Získání charakteristik obrazu odstavce je založeno na zpracování matice pixelů. Součty pixelů po řádcích matice jsou uvedeny v grafu na obr. 2, kde lze výrazně pozorovat hodnoty reprezentující řádky textu i volné řádky. Odhalování chyb v sazbě

se zaměřuje na velké rozdíly v délce řádků (ukazují na nezarovnaný okraj), rovnoměrnost mezer mezi řádky (pravidelnost sazby) a na řádky příliš zaplněné pixely tam, kde by měla být pravidelná meziřádková mezera.

Součty pixelů ve sloupcích obrazu podávají přehled o tvaru odstavce, zejména o zarovnání odstavce, použití odstavcové zarážky a délce východových řádků odstavce. Tyto parametry opět ukazují na pravidelnost sazby a používají se při hodnocení typografické kvality. Zejména způsob zarovnání je velmi výrazným jevem, který je zobrazen v grafech na obr. 3 a 4. Průměrná četnost černých pixelů ve sloupci zůstává relativně stálá, výrazně klesá v případě jiného zarovnání než do bloku. Tímto způsobem je možné odhalit jinou formu zarovnání, popř. příliš nepravidelný okraj sazby. Součty pixelů ve sloupcích se uplatní i při analýze vertikálně strukturovaných dat, jako je sazba textů do více sloupců. Obdobným způsobem lze ze segmentovaných jednotlivých řádků změřit mezislovní mezery.

Popsané způsoby zjištění parametrů umožňují extrahovat z dokumentu parametry odstavců. Spolu s obecně používanými parametry, které lze zjistit pro všechny typy prvků (umístění, výška, šířka, plocha objektu), jsou tak dostupné všechny parametry potřebné pro hodnocení. Navazujícím krokem bude implementace systému pro stanovení typografické kvality dokumentu – porovnání získaných hodnot s typografickými pravidly. Součástí řešení bude i vhodná reprezentace, která uživateli označí a popíše problém vyskytující se v dokumentu a navrhne jeho řešení.

Informace získané při tomto způsobu zpracování dokumentu postačují k následnému hodnocení kvality, přesto se otevírá prostor pro další řešení. Parametrem, který z obrazu detekovat nelze, je například použité písmo. Lze sice odhadnout, zda se jedná o serifové nebo bezserifové písmo (určité možnosti jsou patrné z grafů na obr. 2), přesné určení konkrétního typu je však díky rozmanitosti písem značně komplikované a není prezentovanými metodami uspokojivě řešitelné.

Metoda je věnována textovým prvkům dokumentu, v budoucnu by měla být rozšířena o identifikaci tabulek a jiných strukturovaných prvků a o práci s barvami.

SOUHRN

Článek se zabývá experimentálním stanovením parametrů prvků dokumentů z rastrových obrazů. Formální kvalita dokumentu je považována za stejně důležitou jako jeho obsah, proto byl pro účely kvality analýzy dokumentů navržen formální model dokumentu, který je popsán v předcházejících pracích. Model popisuje stránku dokumentu jako množinu prvků různých typů, přičemž hlavní skupiny jsou textové a grafické objekty. Stránka dokumentu a všechny prvky jsou popsány množinou parametrů závisejících na typu prvku, přičemž nejdůležitější jsou typografické parametry textových objektů.

Hodnoty parametrů prvků jsou získávány z rastrových obrazů dokumentu, což je pro typografickou analýzu vhodnější, a dále mohou být použity i techniky zpracování obrazu. Obraz stránky je zpracován a segmentován na jednotlivé odstavce a jejich parametry jsou zpracovány v procesu analýzy obrazu odstavce.

Obraz je reprezentován jako matice černých a bílých pixelů, z nichž jsou počítány důležité odstavcové charakteristiky. Pro tento účel byla navržena sada algoritmů. Algoritmy jsou zaměřeny na získání parametrů z matice a vycházejí z typografických pravidel. Algoritmy byly testovány na množině obrazů celých stránek vysazeného textu. Přinášejí velmi dobré výsledky, typografické charakteristiky jsou zde evidentní. Obraz stránky proto může být analyzován bez pomoci typografa, a přesto mohou být získány požadované parametry, které lze přímo využít pro automatizované hodnocení typografické kvality.

rastrový obraz, rozpoznávání, dokument, odstavec, typografie, parametry textových objektů

SUMMARY

The paper is aimed at document elements parameters extraction from raster graphics sources. Documents' formal quality is considered to be as important as the content. For the purpose of document quality analysis, a document formal model was designed and described in previous papers. The model describes a document page as a set of elements of different types, the main groups being text objects and graphic objects. Document page and all elements are described by a set of parameters depending on elements' types, whereas the most important ones are typographical parameters of text objects.

Elements parameters values are obtained from raster document image which is more suitable for typographical analysis than structural description and image processing techniques can be used. The page image is processed and segmented into individual paragraphs and their parameters are obtained in a process of image analysis.

The image is represented as a matrix of black or white pixels, from which the important paragraph characteristics are computed. For this purpose, a set of algorithms was designed. Algorithms are aimed at parameters extraction from the matrices and are inspired by typographical rules. Algorithms were tested on a set of 30 images of pages and provide very good results, typographical characteristics are evident. A page image can be therefore analyzed without typographer's help and required parameters can be obtained. Extracted parameters can be directly used for typographical quality evaluation.

Článek vznikl v rámci výzkumného záměru MSM 6215648904/03/03/06.

LITERATURA

BADEKAS, E., PAPAMARKOS, N., 2009: Estimation Of Appropriate Parameter Values For Document Binarization Techniques [online]. [cit. 2010-08-26]. Dostupné z http://ecal.ee.duth.gr/uploadedfiles/Papaparkos/ Journals/206-3193.pdf.

BEITZEL, S., JENSEN, E., GROSSMAN, D., 2003: Retrieving OCR Text: A Survey of Current Approaches [online]. [cit. 2010-08-02]. Dostupné z http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.20.3743&rep= rep1&type=pdf.

CAI, D. a kol., 2003: VIPS: a Vision-based Page Segmentation Algorithm [online]. [cit. 2010-08-24]. Dostupné z http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.118.638.

CAO, J., MAO, B., LUO, J., 2010: A segmentation method for web page analysis using shrinking and dividing [online]. [cit. 2010-08-26]. Dostupné z http://pdfserve.informaworld.com/436593__920033458.pdf.

EIKVIL, L., 1993: OCR Optical Character recognition [online]. [cit. 2010-05-04]. Dostupné z http://citeseerx. ist.psu.edu/viewdoc/summary?doi=10.1.1.25.3684.

KELNAROVÁ, D., 2010: Rozpoznávání prvků dokumentů z grafických předloh. Diplomová práce. Brno: Mendelova univerzita v Brně, 101 s.

KUNC, M., BURGET, R., 2008: Klasifikace prvků dokumentu na základě vizuálních rysů [online]. [cit. 2001-8-25]. Dostupné z http://znalosti2008.fiit.stuba.sk/download/articles/znalosti2008-Kunc.pdf.

ŠPANĚL, M., BERAN, V., 2010: Obrazové segmentační techniky: Přehled existujících metod [online]. 2006-01-19 [cit. 2010-08-23]. Dostupné z http://www.fit.vutbr.cz/~spanel/segmentace/.en.

TALANDOVÁ, P., 2009: Automatizované hodnocení kvality dokumentů. Disertační práce. Brno: MZLU v Brně, 160 s.

TALANDOVÁ, P., RYBIČKA, J., 2009: Stanovení metod automatizovaného hodnocení formální kvality dokumentů. Acta Univ. agric. et silvic. Mendel. Brun., sv. LVII, č. 6, s. 305–313. ISSN 1211-8516.

Adresa

doc. Ing. Jiří Rybička, Dr., Ing. Dagmar Kelnarová, Ing. Petra Talandová, Ph.D., Ústav informatiky, Mendelova univerzita v Brně, Zemědělská 1, 613 00 Brno, Česká republika, e-mail: rybicka@mendelu.cz, xkelnaro@node.mendelu.cz, petra.talandova@pef.mendelu.cz