Sem vložte zadání Vaší práce.

České vysoké učení technické v Praze
Fakulta informačních technologií
Katedra teoretické informatiky

Bakalářská práce

Identifikace oblastí společného významu v naskenovaném dokumentu

František Haifler

Vedoucí práce: doc. RNDr. Ing. Marcel Jiřina, Ph.D.

12. května 2015

Poděkování

Tímto bych chtěl poděkovat vedoucímu své práce za vstřícné chování a svědomité vedení mé práce, a také za mnohou pomoc, kterou mi při plnění práce poskytl. Dále také panu Ing. Jakubovi Novákovi za úvod do implementačního prostředí a jeho používání.

Prohlášení

Prohlašuji, že jsem předloženou práci vypracoval(a) samostatně a že jsem uvedl(a) veškeré použité informační zdroje v souladu s Metodickým pokynem o etické přípravě vysokoškolských závěrečných prací.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona, ve znění pozdějších předpisů. V souladu s ust. § 46 odst. 6 tohoto zákona tímto uděluji nevýhradní oprávnění (licenci) k užití této mojí práce, a to včetně všech počítačových programů, jež jsou její součástí či přílohou a veškeré jejich dokumentace (dále souhrnně jen „Dílo“), a to všem osobám, které si přejí Dílo užít. Tyto osoby jsou oprávněny Dílo užít jakýmkoli způsobem, který nesnižuje hodnotu Díla a za jakýmkoli účelem (včetně užití k výdělečným účelům). Toto oprávnění je časově, teritoriálně i množstevně neomezené.

V Praze dne 12. května 2015 .....

České vysoké učení technické v Praze
Fakulta informačních technologií
© 2015 František Haifler. Všechna práva vyhrazena.
Tato práce vznikla jako školní dílo na Českém vysokém učení technickém v Praze, Fakultě informačních technologií. Práce je chráněna právními předpisy a mezinárodními úmluvami o právu autorském a právech souvisejících s právem autorským. K jejímu užití, s výjimkou bezúplatných zákonných licencí, je nezbytný souhlas autora.

Odkaz na tuto práci

Haifler, František. *Identifikace oblastí společného významu v naskenovaném dokumentu*. Bakalářská práce. Praha: České vysoké učení technické v Praze, Fakulta informačních technologií, 2015.

Abstrakt

Dokument pojednává o problému identifikace oblastí společného významu v naskenovaných dokumentech. První část rozebírá jednotlivé kroky těchto postupů včetně předzpracování obrazu. Další částí je popis zvolené metody pro řešení tohoto problému a její implementace. Závěrem jsou rozebrány výsledky naimplementované metody, případné návrhy úprav pro její vylepšení.

Klíčová slova analýza dokumentu, počítačové vidění, aglomerativní hierarchické shlukování, analýza rozložení, segmentace strany, zpracování obrazu

Abstract

This document discusses the problem of segments identification, which consist of data of similar meaning. First part analyses individual steps of such approaches including image preprocessing. Next part describes chosen method for solving this problem and its implementation. In conclusion are discussed achieved results and eventual adjustments for their improvement.

Keywords document analysis, computer vision, agglomerative hierarchical clustering, layout analysis, page segmentation, image processing

Obsah

Seznam obrázků

Seznam tabulek

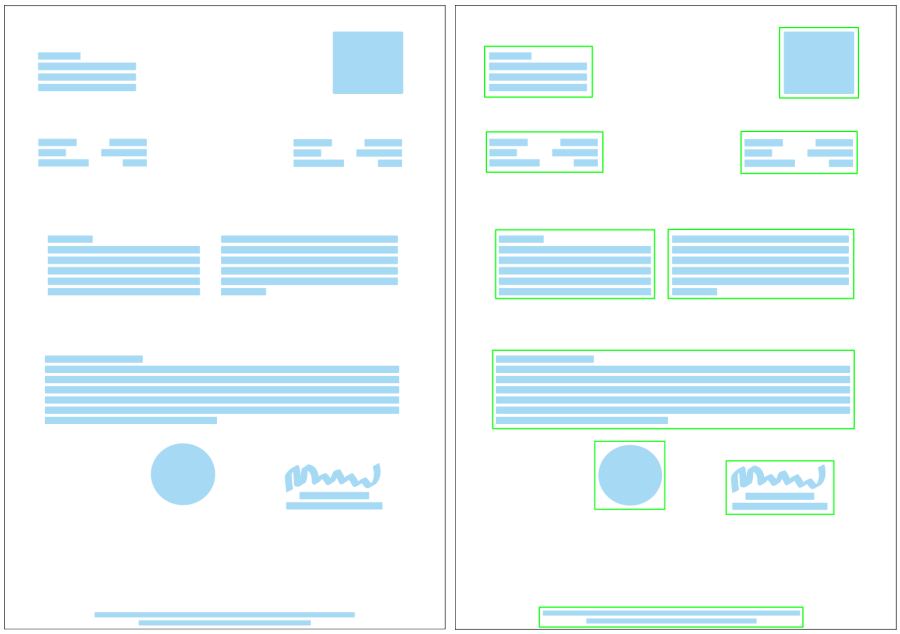
Úvod

Identifikací oblastí v naskenovaném dokumentu se chápe problém, kde na vstupu je dokument v obrazové podobě a na výstupu množina oblastí (v nejlepším případě popsaných obdélníky), které byly v tomto dokumentu nalezeny. Pro tento problém existuje mnoho různých postupů řešení (více v 1.3). Z důvodu vysoké rozdílnosti dokumentů jsou však tyto postupy mnohdy specializované pouze na určitý druh dokumentu a na jiné druhy může takovýto mít postup pouze limitované využití. Proto je výběr správné metody stěžejní problém pro získání korektních a především smysluplných výsledků. Obrázek 0.1 znázorňuje možné rozložení textu vstupního dokumentu a označené identifikované oblasti.

Hlavní složkou této práce je analýza obrazových dat dokumentu, které spadá pod disciplínu zpracování obrazu, které je poddisciplínou počítačového vidění a případně i strojové inteligence. Dále se bude v této práci těchto disciplín a jejich zjištění intenzivně využívat.

Pro potřebu dostatečného porozumění a přesnosti bude také dále popsáno, případně odkazováno, několik metod pro zpracování obrazu zabývajících se především předzpracováním obrazu nutného pro některé z dále probíraných metod pro identifikaci oblastí.

Neposlední částí práce je implementace zvolené metody a diskuze výsledků, kterých se její pomocí dokázalo dosáhnout. Případné návrhy úprav pro její vylepšení a odstranění chyb.



Obrázek 0.1: Ilustrativní příklad dokumentu; vlevo příklad vstupního dokumentu, vpravo příklad dokumentu s označenými oblastmi

Kapitola 1

Analýza problému a existujících řešení

Pro tento obor nejsou k datu ustálené české pojmy — alespoň ne příliš rozšířené, a proto se dále v této práci budou užívat jejich anglické verze, případně jejich český překlad, bude-li dostatečně výmluvný.

1.1 Počítačové vidění a dokumenty

V oboru počítačového vidění je odvětví zabývající se zpracováním a porozuměním lidmi čitelných dokumentů, jak v nativní (digitální) formě v podobě dokumentu ve formátu `.pdf`, tak též v naskenované formě a tedy dokumentu v čistě obrazové formě. Toto odvětví se označuje jako Document Analysis, pro analýzu dokumentu v obrazové formě se mnohdy používá pojem Document Image Analysis.

Tato disciplína se dále skládá z mnoha dalších pod-disciplín, které jsou znázorněny v tabulce 1.1.

Každá tato pod-disciplína řeší jeden či více klíčových problémů nutných pro strojové porozumění dokumentu, ať v nativní či obrazové formě.

Mezi autory neexistuje konsensus jak dělit a označovat jednotlivé kroky, a proto se některé tyto pojmy navzájem překrývají či jsou synonymní. Některé tyto překryvy budou nastíněny později.

Tyto disciplíny dohromady tvoří postup zpracování a porozumění dokumentu, kde každá část se orientuje na porozumění určité části dokumentu. Mnoho těchto disciplín je navzájem závislých a mohou vyžadovat výstupy z předchozího zpracování dokumentu. Například klasifikace bloků v dokumentu (*block classification*) vyžaduje nejdříve identifikaci jednotlivých bloků v dokumentu, které zajišťuje segmentace strany (*page segmentation*). V některých případech jsou tyto dvě disciplíny spojeny realizovány jednou metodou.

Document Classification *(klasifikace dokumentu)*	identifikace druhu dokumentu (článek, kniha, noviny, atp.)
Document Layout Analysis *(analýza rozložení dokumentu)*	rozložení strany dokumentu do jednotlivých oblastí a specifikaci formy a účelu těchto oblastí
Page Segmentation *(segmentace strany)*	rozložení strany dokumentu do jednotlivých oblastí (podproblém analýzy rozložení)
Block Classification *(klasifikace bloků)*	specifikace formy a účelu oblastí v dokumentu
OCR *(Optical Character Recognition)*	získání textové informace z obrazových dat

Tabulka 1.1: Pod-disciplíny spadající pod Document Image Analysis

Tyto dvě disciplíny jsou mnohdy označované souhrnně jako analýza rozložení dokumentu.

1.2 Analýza rozložení dokumentu

Tato analýza se dělí na dvě části — geometrickou (někdy též fyzickou) a logickou.

- **Geometrická** analýza rozložení dokumentu s zabývá možností jeho rozdělení na samostatné bloky, kde blok reprezentuje sadu objektů v dokumentu které spolu pozičně souvisí. Může se tedy jednat například o sloupec textu, obrázek či tabulku. Někteří autoři pro tuto analýzu používají také pojem segmentace strany.

- **Logická** analýza rozložení dokumentu se snaží popsat funkce daného bloku v dokumentu. Tento krok je velmi závislý na typu dokumentu, který se zpracovává. V případě článku mohou funkce bloků být např. nadpis, titulek, obrázek, atp. V případě faktury tyto funkce mohou být např. identifikace odběratele a příjemce, atp. Někteří autoři tuto analýzu označují jako klasifikace bloků případně značení bloků (*block labeling*).

Tato práce se dále bude zabývat pouze geometrickou analýzou a z důvodu jednoduchosti a jednoznačnosti se bude označovat jako segmentace strany.

1.3 Segmentace strany

V této chvíli již lze přistoupit k samotnému problému identifikace oblastí v dokumentech. Tento problém je již velmi dlouho známý a zkoumaný. Největší zájem o tuto problematiku byl v 90. letech, kdy byly publikovány desítky článků o možných přístupech k řešení problému. V přehledech [1, 2] jsou zahrnuty některé významnější metody pro segmentaci stran.

Tyto metody lze rozdělit dle přístupu k řešení problému na:

1. **Shora-dolů** (*top -down*)

   Metody s tímto přístupem obecně začínají s jedním blokem reprezentujícím celou stranu, či nějaké její triviální rozdělení. Dále dochází k rekurzivnímu rozdělování bloků na stále menší podbloky, které postupně reprezentují sloupce, řádky, slova, písmena. Tyto metody jsou obecně postaveny na analýze pozadí a hledání bílých mezer na obrazu dokumentu.

   Tyto metody bývají výpočetně méně náročné, neboť algoritmus pro hledání bílých mezer bývá většinou lineární [3, s. 1.]. Navíc hledáme-li pouze bloky textu, příp. obrázků, lze algoritmus poměrně brzy ukončit (není nutné pokračovat v dělení až k jednotlivým písmenům).

   Některé z těchto metod jsou popsány v [4, 5].

2. **Zdola-nahoru** (*bottom -up*)

   Metody zdola-nahoru začínají s jednotlivými pixely či s jejich množinami v podobě spojitých komponent (viz 1.3.1). Tyto objekty postupně spojujeme do písmen, slov, řádek a sloupců. Oproti metodám shora-dolů (viz 1) tyto metody pracují s obsahem a jsou s ním přímo spjaté. Z tohoto důvodu jsou více robustní v případě zpracovávání dokumentů s komplexnějším formátem.

   Tyto typy metod jsou však výpočetně náročnější z důvodu výpočtu vzdáleností mezi jednotlivými objekty (jejichž počet bývá zpočátku vysoký). Tyto algoritmy mívají zpravidla kvadratickou složitost [3, s. 1. - 2.]. V průběhu zpracovávání se však jejich počet rapidně snižuje díky jejich sjednocování do větších celků.

   Mezi metody zdola-nahoru patří například [3, 6, 7, 8, 9, 10, 11, 12].

3. **Smíšené** (*hybrid*)

   Tyto metody se snaží udržet robustnost přístupu zdola-nahoru s výpočetní efektivitou blížící se možnostem přístupu shora-dolů.

   Některé z těchto algoritmů jsou představeny v [13, 14, 15, 16].

4. **Ostatní**

   Existují ještě další metody, které se nedají jednoznačně zařadit do žádné předchozí skupiny. Jedním příkladem je využití fraktálového podpisu pro analýzu dokumentu [17]. Dalším je použití voronoiových diagramů pro rozdělení strany do vzájemně sousedících oblastí, které mohou obsáhnout i dokumenty ne-manhattanského typu — dokumenty které nelze rozdělit pomocí vzájemně pravoúhlých čar [18].

Toto dělení je však poněkud nucené a některé přístupy nelze přiřadit ani do jedné z těchto tříd, případně se i stává, že jedna metoda je různými autory zařazena do jiné skupiny [1, s. 17.]. Pro většinu metod však toto členění dostačuje. S ohledem na cíl této práce, který se orientuje spíše na seznámení s použitelnými metodami, než na jejich exaktní popis a přesné členění, bude dále tohoto členění využito.

1.3.1 Předzpracování

Převážná většina představených metod nepracuje s čistými obrazovými daty, ale vyžaduje jejich určité úpravy. Tento krok se nazývá předzpracování (*angl. preprocessing*). Některé techniky předzpracování jsou:

- **Binarizace obrazu** (také tresholding)

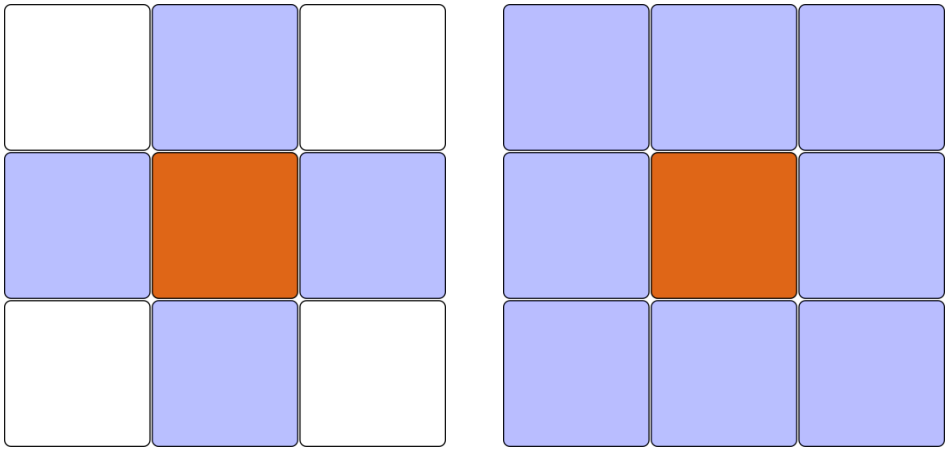
  Binarizace obrazu je proces, při kterém se barevný obraz či obraz v odstínech šedi převede na binární podobu vyjadřující u každého pixelu pouze informaci pozadí/popředí, resp. 0/1. Vzorový obraz se převede do této formy pomocí porovnání intenzity barvy s explicitně zadaným prahem, který jednoznačně rozdělí intenzitu na dva intervaly. Jinak řečeno, je-li intenzita barvy menší než zadaný práh, potom se jedná o pozadí, jinak se jedná o popředí.

  Výběr prahu pro binarizaci je velmi důležitý, neboť pokud je práh příliš nízký, může dojít ke splynutí pozadí s popředím kvůli šumu způsobeným scanováním, případně ztrátovou komprimací. Na druhou stranu je-li práh příliš vysoký, může dojít ke ztrátě relevantních částí dokumentu, které nedosahují dostatečného kontrastu. Více o binarizaci lze najít v [19, s. 9. - 18.].

  Další metody pro binarizaci s více robustním zvolením prahu lze najít v [20, 21]. Velmi známou a používanou metodu volby prahu pro binarizaci pomocí histogramů šedého spektra publikoval Otsu [22].

- **Redukce šumu**

  Po binarizaci se obvykle provede redukce šumu. Díky špatné kvalitě při skenování mohou být ve vzniklém binárním obrazu chybně určené pixely, které můžou při dalším zpracování negativně ovlivnit výsledek. Pro redukci šumu existují různé způsoby, v rámci této práce je tento proces



Obrázek 1.1: Ilustrace 4-okolí a 8-okolí

pouze vylepšení kvality, a proto nebude dále probírán. Větší podrobnosti najít např. v [19, s. 19. - 24.].

- **Nalezení spojitých komponent**

  Spojitá komponenta (angl. connected component) je skupina pixelů popředí, které jsou vzájemně tranzitivně sousedící. V praxi se používají dva druhy okolí — *4-okolí* a *8-okolí*, které jsou ilustrované na obrázku 1.1. V rámci 4-okolí jsou dva pixely sousedící právě tehdy, když oba jsou pixely popředí a leží na stejné vertikální, resp. horizontální poloze a v horizontálním, resp. vertikálním směru mají vzdálenost 1. 8-connectivity je rozšíření o sousednost v diagonálním směru [10, s. 31.].

  Mnohdy se v rámci spojitých komponent využívá jejich ohraničující obdélník, který se získá triviálně pomocí nejlevějšího, nejpravějšího a nejsvrchnějšího, nejspodnějšího pixelu, který obsahují.

1.3.2 Popis metod pro segmentaci stran

1. **Shora-dolů**

   Breuel [5] přistupuje k problému segmentace strany nalezením největších prázdných obdélníků. Obdélník je prázdný tehdy, pokud neprotíná žádný z ohraničujících obdélníků žádné spojité komponenty na straně. Následně se spojí jednotlivé komponenty, pokud je od sebe neodděluje žádný z nalezených obdélníků.

   Další metoda popsaná Nagym et al. [4] využívá formální gramatiky dle Chomského hierarchie. Tato gramatika je specifická pro každý typ dokumentu, neboť se její pomocí segmentuje strana do bloků a zároveň se získané bloky klasifikují. Z tohoto důvodu je nutné pro každý typ dokumentu vytvořit gramatiku pro jeho rozčlenění. Každá z těchto gramatik

se skládá z několika blokových gramatik, kde každá bloková gramatika rozděluje blok horizontálně nebo vertikálně. Toto rozdělení je representováno tzv. X-Y stromem.

2. **Zdola-nahoru**

   Základem mnoha metod pro segmentaci je Run-Length Smearing Algorithm (*RLSA*). Tento algoritmus pracuje s binarizovaným obrazem tímto způsobem. V případě že se nalezne sekvence pixelů pozadí mezi dvěma pixely popředí taková, že počet pixelů pozadí je menší než stanovená hranice, potom se přepíší pixely pozadí na pixely popředí, jinak se pokračuje. Tento postup se provede nejdříve v horizontálním směru s hranicí C_h, to samé se provede ve vertikálním směru s hranicí C_v na vzorovém binarizovaném obrazu. Tyto dva výstupy se spojí binárním operátorem `AND` a posléze se provede finální smearing ve vertikálním směru s hranicí C_f, pro spojení diakritických znamének s písmenem, či pro spojení jednotlivých řádek, dle potřebné detailnosti výsledného rozdělení. Tohoto algoritmu se využívá např. [6, 7, 9, 10].

   Ferilli et al. [6] využívá drobně upravené verze *RLSA*, která namísto operátoru `AND` používá operátor `OR`, bez potřeby finálního vertikálního použití smearingu. V tomto případě lze provést vertikální smearing na výsledku předchozího v horizontálním směru.

   Shih et al. [7] uvádí optimalizace *RLSA*, které se zaměřují na minimalizaci nutných průběhů, které pomocí omezujících předpokladů na vstupní obraz dokáže snížit jejich počet až na 1.

   Další metoda popsaná Simonem et al. [3] využívá teorie grafů pro shlukování spojitých komponent. Každá spojitá komponenta je vrchol grafu, hrany představují vzájemnou polohu jednotlivých komponent, kde váha hrany je maximální vzdálenost mezi komponentami ve vertikálním a horizontálním směru. K tomuto grafu se nalezne minimální kostra pomocí Kruskalova algoritmu, čímž se identifikují přímo sousedící komponenty. Poté se tyto komponenty shlukují do slov, řádek atd.

   Zlatopolsky [8] k segmentaci strany přistupuje zcela jiným způsobem. Tato metoda pracuje nad spojitými komponentami. Pro každou komponentu nalezne jejího pravého nejbližšího souseda, a pokud je jejich vertikální a horizontální vzálenost menší předem daných hranic, potom spojíme tyto komponenty do segmentu řádky. V opačném případě vytvoříme nový segment řádky. Dále se odhadne pootočení obrazu pomocí průměrné orientace jednotlivých segmentů řádek. Poté se přepočítá poloha segmentů, pro získání obrazu dokumentu bez pootočení. Pro každý nejlevější nezpracovaný segment řádky se nalezne jeho nejbližší pravý soused, který není vertikálně vychýlen a nejbližší horní soused, jehož obraz na horizontální osu se překrývá s obrazem zpracovávaného segmentu.

3. **Smíšené**

   Hirayama [13] předkládá metodu založenou na *RLSA*, pro který vypočtou vertikální a horizontální hranice pomocí histogramů vzdáleností mezi spojitými komponentami a jejich výšek. Poté se provede analýza hraničních čar oddělující jednotlivé sloupce textu a dojde ke sjednocení bloků patřících do stejných sloupců.

   Okamoto et al. [14] popisuje obdobnou metodu, která po použití *RLSA* nalezne všechny oddělovací čáry a mezery oddělující sloupce. Toto může vést k vysokému počtu nalezených oddělovačů, a proto se následně provede jejich redukce. Redukce se realizuje pomocí sloučení hraničících oblastí, případně jejich rozlomení pokud se překrývají. Dalším krokem je analýza spojitých komponent získaných z *RLSA* za účelem jejich rozdělení do skupin (šum, část textu, obrázek). Jednotlivé spojité komponenty následně spojují do textových bloků s ohledem na oddělovače získané v jednom z předchozích kroků.

   Esposito et al. [15] představuje metodu založenou opět na *RLSA*, pro který se spočte horizontální hranice pomocí histogramu komplexity dokumentu v horizontálním směru. Poté se proces dělí na globální a lokální analýzu. Jako první dojde ke globální analýze, která zahrnuje detekci sloupců pomocí vertikálních histogramů. Posléze se pomocí horizontálního histogramu detekují jednotlivé sekce a odstavce v každém sloupci. Poté se pomocí poměru pokrytí identifikují grafické bloky (pokrytí obvykle více jak 50 %). Dále následuje lokální analýza, která má za úkol analyzovat identifikované bloky. Z důvodu použití *RLSA* mohou být některé bloky rozděleny do více menších bloků. Dojde k identifikaci těchto bloků pomocí podmínky zarovnání a jejich sjednocení.

   Lin et al. [16] přistupuje k řešení z jiného hlediska. Nejdříve je dokument rozčleněn předem určeného počtu malých homogenních bloků a pro každý blok je spočtena energie, entropie, sdružená entropie, rozdíl entropie a standardní odchylka. Pomocí algoritmu k-středů se vytvoří shluky pro skupiny bloků — text, grafika, bílá oblast. Algoritmus má předem k dispozici prototyp pro každou skupinu. Tyto prototypy jsou použity jako středy a v každé iteraci upravovány podle průměru hodnot v jim přilehlých shlucích. Bloky jsou posléze vzájemně sjednocovány pokud jsou ve vzájemném 8-okolí a náleží do stejného shluku.

Kapitola 2

Zvolené řešení

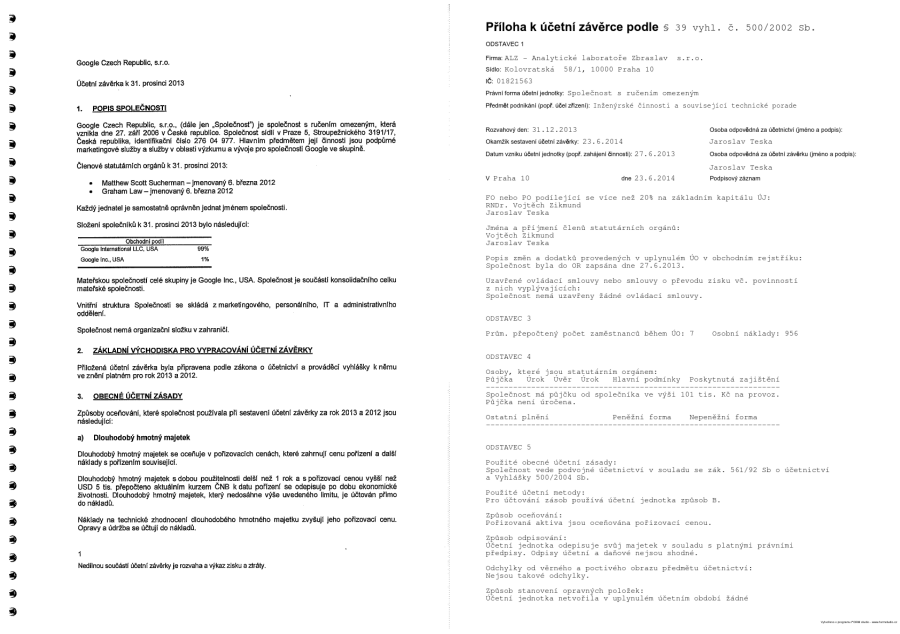
Dokumenty, na které je tato práce cílena jsou velmi různorodé, jak v náplni, tak i v jejím formátování. Mohou se v nich objevovat graficky zpracované a ohraničené tabulky, či jejich čistě textové podoby, mnohdy jsou přítomná loga firem, razítka, či jiné grafické objekty. Text se vyskytuje v jednom či více sloupcích, je zarovnaný do buněk tabulek, případně může obsahovat popisky pootočené o 90 °. Formát textu v rámci jedné strany také nemusí být jednotný a může se velmi lišit. Především v rámci faktur je značná variace, co se týče formátu textu, ať už se jedná o typ písma, velikost, či jeho barvu. Obrázky 2.1, 2.2 a 2.3 ukazují některé z typů formálních dokumentů, pro které by měla zvolená metoda správně identifikovat oblasti.

Z důvodu různorodosti obsahu dokumentu není snadné exaktně definovat všechny možné typy oblastí, které by strana dokumentu mohla obecně obsahovat, nemluvě o strategii jeho identifikace, např. v podobě nastavení hladin a konstant pro popsané metody segmentace strany. Tomuto brání především fakt, že pro určité části dokumentu by byli potřebné určité hodnoty, ale pro jinou část dokumentu by byli tyto hodnoty potřebné zcela jiné.

Z tohoto důvodu je nutné použít postup, který dokáže vyřešit i tyto lokální nesrovnalosti. Po diskuzi s vedoucím práce byl zvolen postup založení na shlukové analýze, který byl inspirován [16, 12]. Tyto postupy nevyužívají shlukovou analýzu a podobných technik přímo pro segmentaci strany, ale těží z generalizačních schopností těchto technik pro určité mezikroky pro samotnou segmentaci. V této práci bude však shluková analýza využita na samotnou segmentaci.

2.1 Shluková analýza

Shluková analýza je proces seskupování množin objektů do tzv. shluků. V každém shluku jsou takové objekty, které jsou si vzájemně předem definovaným způsobem podobné — objekty v rámci jednoho shluku jsou si více podobné než objekty z různých shluků. Tento postup lze použít i pro objekty, jejichž



Google Czech Republic, s.r.o.

Účetní závěrka k 31. prosinci 2013

1. POPIS SPOLEČNOSTI

Google Czech Republic, s.r.o., (dále jen „Společnost") je společnost s ručením omezeným, která vznikla dne 27. září 2006 v České republice. Společnost sídlí v Praze 5, Stroupežnického 3191/17, Česká republika, identifikační číslo 276 04 877. Hlavním předmětem její činnosti jsou podpůrné marketingové služby a služby v oblasti výzkumu a vývoje pro společnosti Google ve skupině.

Členové statutárních orgánů k 31. prosinci 2013:

- Matthew Scott Sucherman – jmenovaný 6. března 2012
- Graham Law – jmenovaný 6. března 2012

Každý jednatel je samostatně oprávněn jednat jménem společnosti.

Složení společníků k 31. prosinci 2013 bylo následující:

Obchodní podíl	
Google International LLC, USA	99%
Google Inc., USA	1%

Mateřskou společností celé skupiny je Google Inc., USA. Společnost je součástí konsolidačního celku mateřské společnosti.

Vnitřní struktura Společnosti se skládá z marketingového, personálního, IT a administrativního oddělení.

Společnost nemá organizační složku v zahraničí.

2. ZÁKLADNÍ VÝCHODISKA PRO VYPRACOVÁNÍ ÚČETNÍ ZÁVĚRKY

Přiložená účetní závěrka byla připravena podle zákona o účetnictví a prováděcí vyhlášky k němu ve znění platném pro rok 2013 a 2012.

3. OBECNÉ ÚČETNÍ ZÁSADY

Způsoby oceňování, které společnost používala při sestavení účetní závěrky za rok 2013 a 2012 jsou následující:

a) Dlouhodobý hmotný majetek

Dlouhodobý hmotný majetek se oceňuje v pořizovacích cenách, které zahrnují cenu pořízení a další náklady s pořízením související.

Dlouhodobý hmotný majetek s dobou použitelnosti delší než 1 rok a s pořizovací cenou vyšší než USD 5 tis. přepočtenou aktuálním kurzem ČNB k datu pořízení se odepisuje po dobu ekonomické životnosti. Dlouhodobý hmotný majetek, který nedosáhne výše uvedeného limitu, je účtován přímo do nákladů.

Náklady na technickou zhodnocení dlouhodobého hmotného majetku zvyšují jeho pořizovací cenu. Opravy a údržba se účtují do nákladů.

1

Nedílnou součástí účetní závěrky je rozvaha a výkaz zisku a ztráty.

Příloha k účetní závěrce podle § 39 vyhl. č. 500/2002 Sb.

ODSTAVEC 1

Firma: ALZ - Analytické laboratoře Zbraslav s.r.o.
Sídlo: Kolovratská 58/1, 10000 Praha 10
IČ: 01821563
Právní forma účetní jednotky: Společnost s ručením omezeným
Předmět podnikání (popř. účel zřízení): Inženýrské činnosti a související technické porade

Rozvahový den: 31.12.2013 — Osoba odpovědná za účetnictví (jméno a podpis): Jaroslav Teska
Okamžik sestavení účetní závěrky: 23.6.2014
Datum vzniku účetní jednotky (popř. zahájení činnosti): 27.6.2013 — Osoba odpovědná za účetní závěrku (jméno a podpis): Jaroslav Teska
V Praha 10 dne 23.6.2014 — Podpisový záznam

FO nebo PO podílející se více než 20% na základním kapitálu ÚJ:
RNDr. Vojtěch Zikmund
Jaroslav Teska

Jména a příjmení členů statutárních orgánů:
Vojtěch Zikmund
Jaroslav Teska

Popis změn a dodatků provedených v uplynulém ÚO v obchodním rejstříku:
Společnost byla do OR zapsána dne 27.6.2013.

Uzavřené ovládací smlouvy nebo smlouvy o převodu zisku vč. povinností z nich vyplývajících:
Společnost nemá uzavřeny žádné ovládací smlouvy.

ODSTAVEC 3

Prům. přepočtený počet zaměstnanců během ÚO: 7 Osobní náklady: 956

ODSTAVEC 4

Osoby, které jsou statutárním orgánem: Půjčka	Úrok	Úvěr	Úrok	Hlavní podmínky	Poskytnutá zajištění
Společnost má půjčku od společníka ve výši 101 tis. Kč na provoz. Půjčka není úročena.					

Ostatní plnění	Peněžní forma	Nepeněžní forma

ODSTAVEC 5

Použité obecné účetní zásady:
Společnost vede podvojné účetnictví v souladu se zák. 561/92 Sb o účetnictví a Vyhlášky 500/2004 Sb.

Použité účetní metody:
Pro účtování zásob používá účetní jednotka způsob B.

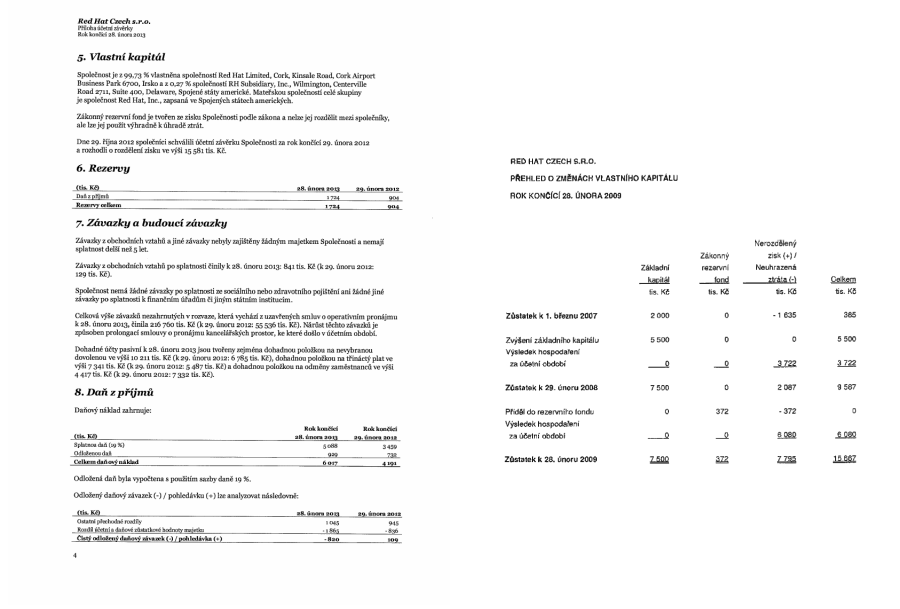
Způsob oceňování:
Pořizovaná aktiva jsou oceňována pořizovací cenou.

Způsob odpisování:
Účetní jednotka odepisuje svůj majetek v souladu s platnými právními předpisy. Odpisy účetní a daňové nejsou shodné.

Odchylky od věrného a poctivého obrazu předmětu účetnictví:
Nejsou takové odchylky.

Způsob stanovení opravných položek:
Účetní jednotka netvořila v uplynulém účetním období žádné

Obrázek 2.1: Příklad účetní závěrky



Red Hat Czech s.r.o.
Příloha účetní závěrky
Rok končící 28. února 2013

5. Vlastní kapitál

Společnost je z 99,73 % vlastněna společností Red Hat Limited, Cork, Kinsale Road, Cork Airport Business Park 6700, Irsko a z 0,27 % společností RH Subsidiary, Inc., Wilmington, Centerville Road 2711, Suite 400, Delaware, Spojené státy americké. Mateřskou společností celé skupiny je společnost Red Hat, Inc., zapsaná ve Spojených státech amerických.

Zákonný rezervní fond je tvořen ze zisku Společnosti podle zákona a nelze jej rozdělit mezi společníky, ale lze jej použít výhradně k úhradě ztrát.

Dne 29. října 2012 společníci schválili účetní závěrku Společnosti za rok končící 29. února 2012 a rozhodli o rozdělení zisku ve výši 15 581 tis. Kč.

6. Rezervy

(tis. Kč)	28. února 2013	29. února 2012
Daň z příjmů	1 724	924
Rezervy celkem	**1 724**	**924**

7. Závazky a budoucí závazky

Závazky z obchodních vztahů a jiné závazky nebyly zajištěny žádným majetkem Společnosti a nemají splatnost delší než 5 let.

Závazky z obchodních vztahů po splatnosti činily k 28. únoru 2013: 841 tis. Kč (k 29. únoru 2012: 129 tis. Kč).

Společnost nemá žádné závazky po splatnosti ze sociálního nebo zdravotního pojištění ani žádné jiné závazky po splatnosti k finančním úřadům či jiným státním institucím.

Celková výše závazků nezahrnutých v rozvaze, která vychází z uzavřených smluv o operativním pronájmu k 28. únoru 2013, činila 216 760 tis. Kč (k 29. únoru 2012: 55 536 tis. Kč). Nárůst těchto závazků je způsoben prolongací smlouvy o pronájmu kancelářských prostor, ke které došlo v účetním období.

Dohadné účty pasivní k 28. únoru 2013 jsou tvořeny zejména dohadnou položkou na nevybranou dovolenou ve výši 10 211 tis. Kč (k 29. únoru 2012: 6 785 tis. Kč), dohadnou položkou na třináctý plat ve výši 7 341 tis. Kč (k 29. únoru 2012: 5 487 tis. Kč) a dohadnou položkou na odměny zaměstnancům ve výši 4 417 tis. Kč (k 29. únoru 2012: 7 332 tis. Kč).

8. Daň z příjmů

Daňový náklad zahrnuje:

(tis. Kč)	Rok končící 28. února 2013	Rok končící 29. února 2012
Splatná daň (19 %)	3 088	2 439
Odložená daň	929	738
Celkem daňový náklad	**4 017**	**4 191**

Odložená daň byla vypočtena s použitím sazby daně 19 %.

Odložený daňový závazek (-) / pohledávku (+) lze analyzovat následovně:

(tis. Kč)	28. února 2013	29. února 2012
Ostatní přechodné rozdíly	1 045	945
Rozdíl zůstatkové a daňové zůstatkové hodnoty majetku	- 1 865	- 836
Čistý odložený daňový závazek (-) / pohledávka (+)	**- 820**	**109**

4

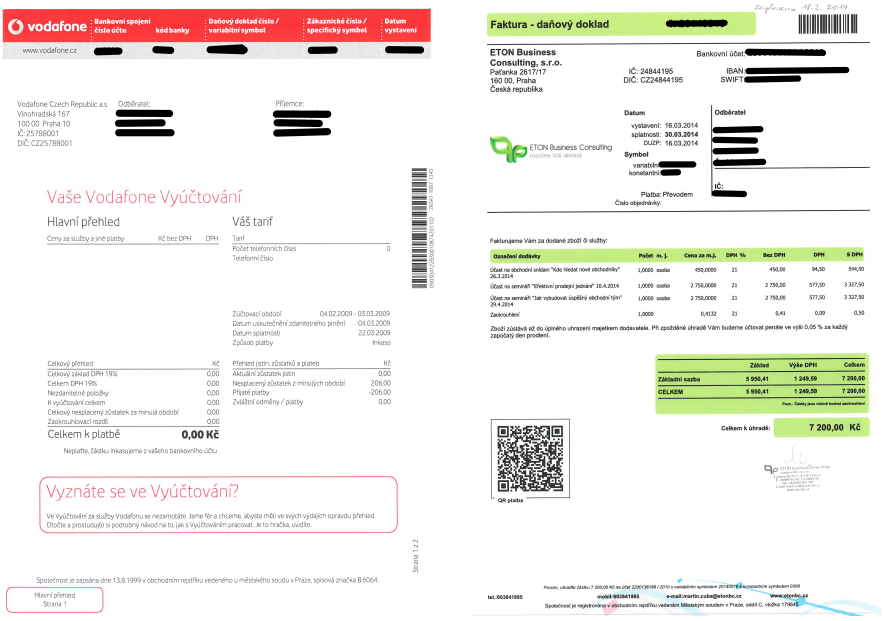
RED HAT CZECH S.R.O.

PŘEHLED O ZMĚNÁCH VLASTNÍHO KAPITÁLU

ROK KONČÍCÍ 28. ÚNORA 2009

	Základní kapitál	Zákonný rezervní fond	Nerozdělený zisk (+) / Neuhrazená ztráta (-)	Celkem
	tis. Kč	tis. Kč	tis. Kč	tis. Kč
Zůstatek k 1. březnu 2007	2 000	0	- 1 635	365
Zvýšení základního kapitálu	5 500	0	0	5 500
Výsledek hospodaření za účetní období	0	0	3 722	3 722
Zůstatek k 29. únoru 2008	7 500	0	2 087	9 587
Příděl do rezervního fondu	0	372	- 372	0
Výsledek hospodaření za účetní období	0	0	6 080	6 080
Zůstatek k 28. únoru 2009	7 500	372	7 795	15 667

Obrázek 2.2: Příklad výroční zprávy



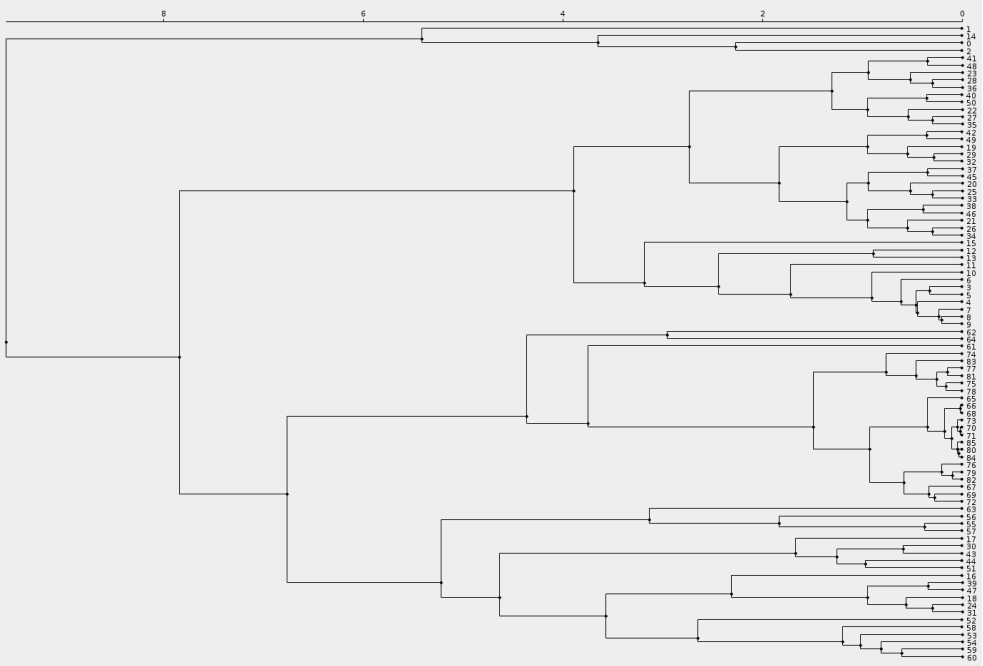
Obrázek 2.3: Příklady faktur

vzájemné vztahy neznáme. V rámci shlukové analýzy existuje mnoho algoritmů pro její řešení, které se vzájemně mohou velmi lišit, jak v datech nad kterými pracují, tak výsledky, které poskytují. Pro účel shlukování objektů do oblastí, byl zvolen přístup hierarchického shlukování. Pro výběr hierarchického shlukování vede fakt, že samotný dokument je ve své podstatě hierarchií, kde se jednotlivé pixely shlukují do písmen, slov, vět, odstavců atp. A díky tomuto faktu je možnost vytvoření modelu této hierarchie pomocí shlukové analýzy.

2.1.1 Hierarchické shlukování

Hierarchické shlukování je přístup ke shlukové analýze, který je založený na principu, kdy objekty, které mají navzájem menší vzdálenost, jsou si navzájem podobnější. Tyto objekty seskupovány do shluků s ohledem na jejich vzájemnou vzdálenost, případně vzdálenost mezi objekty a okolními shluky. Jinými slovy jsou dva objekty seskupeny do shluku v bodě, kde je jejich vzájemná vzdálenost rovna hodnotě tohoto bodu. Tyto objekty mohou být například body v n-rozměrném prostoru a jejich podobnost je definována jako vzdálenost těchto bodů v daném prostoru. Jednotlivé shluky se poté též vzájemně seskupují. Tento proces pokračuje do doby, kdy jsou všechny objekty seskupený do jednoho shluku.

Tímto procesem dojde k vytvoření hierarchie shluků, kterou lze vizualizovat pomocí tzv. dendrogramu. Příklad hierarchie shluků vizualizované dengrogramem lze vidět na obrázku 2.4. Pro získání množiny shluků odpovídající



Obrázek 2.4: Vizualizace hierarchického shlukování pomocí dendrogramu

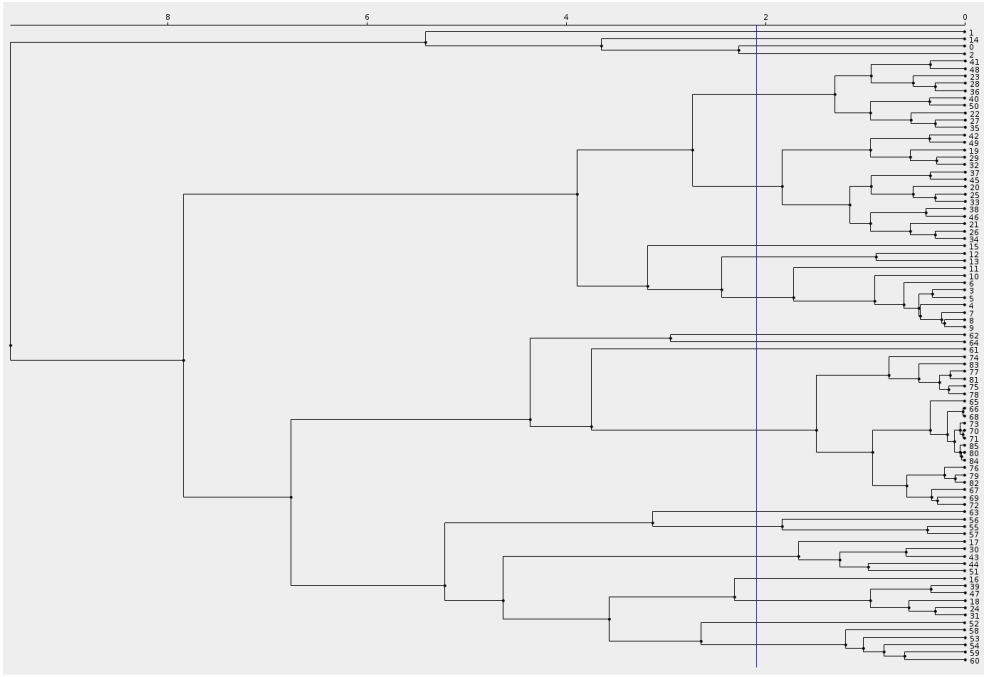
určité hodnotě vzdálenosti (podobnosti) mezi shluky se provede řez takto získané hierarchie v dané vzdálenosti. Takovýto řez je ilustrován na obrázku 2.5.

Pro hierarchické shlukování existují dva přístupy jeho sestavení:

- **Divizní** — začíná s jedním shlukem, který dělí do menších shluků
- **Aglomerativní** — začíná s jednotlivými objekty, které seskupuje do shluků

V rámci této práce bude využito aglomerativního přístupu z důvodu menší výpočetní náročnosti a jednodušší extrakce atributů. Další částí této problematiky je výpočet vzdálenosti mezi jednotlivými shluky. Tento problém je řešen několika metodami:

- **metoda nejbližšího souseda** — vzdálenost shluků je vzdálenost dvou nejbližších objektů z těchto shluků
- **metoda nejvzdálenějšího souseda** — vzdálenost shluků je naopak vzdálenost dvou nejvzdálenějších objektů těchto shluků
- **centroidní metoda** — vzdálenost shluků je vzdáleností mezi těžišti těchto shluků



Obrázek 2.5: Řez hierarchie shluků v daném bodě

- **metoda průměrné vazby** — vzdálenost shluků je rovna průměru vzdáleností všech párů objektů mezi shluky

- a další

2.2 Popis metody

Jak již bylo zmíněno, zvolená metoda pro identifikaci oblastí bude založena na aglomerativním hierarchickém shlukování. Před tím, než lze využít hierarchického shlukování je však nutné převést obrazová data do formy, se kterou dokáže shluková analýza pracovat.

Vstupní obraz naskenovaného dokumentu se v první řadě převede na standardní velikost pro zaručení správné funkčnosti *RLSA*, jelikož tento algoritmus pracuje s fixními hranicemi. Poté se tento obraz překóduje z barevného spektra do stupňů šedi. Obraz ve stupních šedi se poté pomocí Otsuovy metody [22] hledání prahu binarizuje.

Nad takto připravenými daty se použije upravený *RLSA* s operací `OR` [6]. Z důvodu optimálnosti byl zvolena verze tohoto algoritmu, kde se namísto operace `OR` provede vertikální smearing na výsledku předem provedeného horizontálního. Použití tohoto algoritmu bylo zvoleno z důvodu snížení počtu

elementárních objektů pro shlukování. Tímto dojde ke sjednocení triviálně sousedících objektů.

Ve výstupu předchozího kroku se naleznou spojité komponenty odpovídající spojeným blokům textu či grafických objektů. Z těchto komponent se následně vypočítají jejich atributy jako obsah jejich obrysu, souřadnice je ohraničujícího obdélníku, těžiště tohoto obdélníku, jeho obsah a poměr obsahu obrysu a obsahu obdélníku. U těchto komponent může nastat případ, kdy jejich obrysy mohou být vzájemně disjunktní, ale jeden z ohraničujících obdélníků může ležet uvnitř druhého. Dříve než se přistoupí k dalšímu kroku, vyhledají se takovéto případy a dojde k absorpci uvnitř ležící komponenty. Dále se pro vzájemnou stabilitu těchto atributů normalizují jejich hodnoty pomocí min-max metody pro normalizaci. Tato normalizace je lineární zobrazení z dosavadního intervalu hodnot do nově definovaného. Toto zobrazení je dáno předpisem:

$$hodnota_{\text{norm}} = \frac{hodnota - min}{max - min} \cdot (max_{\text{norm}} - min_{\text{norm}}) + min_{\text{norm}} \ [23, s.\ 18]$$

Na normalizované atributy se přinásobí váhy udávající míru důležitosti daného atributu na podobnost dvou komponent. Pomocí normalizovaných, vážených hodnot atributů se vypočte matice vzdáleností jednotlivých komponent. Za předpokladu, že $d(A, B)$ je funkce, vracející vzdálenost atributu A od atributu B, vypadá matice následujícím způsobem:

$$D = \begin{pmatrix} 0 & d(A_1, A_2) & \cdots & d(A_1, A_n) \\ d(A_2, A_1) & 0 & \cdots & d(A_2, A_n) \\ \vdots & \vdots & \ddots & \vdots \\ d(A_n, A_1) & d(A_n, A_2) & \cdots & 0 \end{pmatrix},$$

kde n je počet nalezených komponent.

Tato je matice symetrická, jelikož $d(A, B) = d(B, A)$. Vzdálenost dvou komponent je spočtena jako euklidovská vzdálenost v R^m, kde m je počet extrahovaných atributů, která je definována jako:

$$d(A, B) = \sqrt{\sum_{i=1}^{m} (a_i - b_i)^2}$$

Tato matice již lze využít pro algoritmus hierarchického shlukování, který je dalším krokem. Pomocí shlukování se sestaví hierarchie shluků, na které poté může být proveden řez v požadované vzdálenosti.

Pomocí tohoto řezu se získá množina shluků, které jsou ekvivalentem požadovaných oblastí. Dále následuje pouze zpracování těchto shluků do formy vyjadřující souřadnice oblastí. Pro každý shluk se zjistí množina jej tvořících komponent. Pro tyto komponenty se poté nalezne ohraničující obdélník, obsahující všechny tyto komponenty. Tento obdélník je hledanou oblastí, je však

nutné přepočítat jeho souřadnice, neboť získaný obdélník je v souřadnicovém systému obrazu standardní velikosti.

Kapitola 3

Implementace

Implementace popsané metody je v jazyce Java nad projektem *OBBB* [24], který je postavený na platformě NetBeans [25]. Implementace této metody je v rámci modulu ImageProcessingSuite projektu *OBBB*. V rámci tohoto projektu existuje souprava tzv. bloků, kde každý z těchto bloků zajišťuje nějakou specifickou funkci — například načtení obrazu ze souboru, binarizace obrazu atd. Každý blok přijímá množinu vstupů, se kterými pracuje a dále poskytuje množinu výstupu vhodných pro další zpracování. Jak již bylo zmíněno, projekt je napsaný v programovacím jazyce Java, který je silně staticky typovaný, a proto jsou datové typy pro tyto vstupy a výstupy předem definovány a pro správnou funkci těchto bloků je nutné tyto konvence dodržovat.

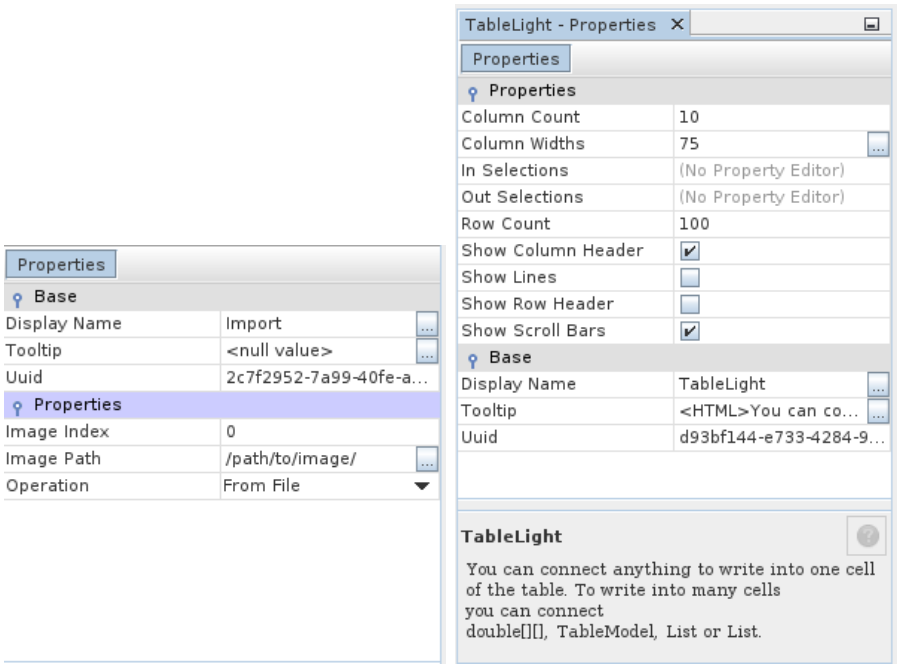
Bloky v sobě zaštiťují množinu parametrů, pomocí kterých se řídí a upravuje průběh zpracování. Bloky tyto parametry poskytují uživateli pro prozkoumání či jejich přenastavení. Rozhraní pro práci s parametry lze vidět na obrázku 3.1.

Další jejich funkčností je schopnost jejich vizualizace — struktura i obsah této vizualizace je zcela v rámci potřeb daného bloku. Příklady takovéto vizualizace bloku lze vidět na obrázku 3.2.

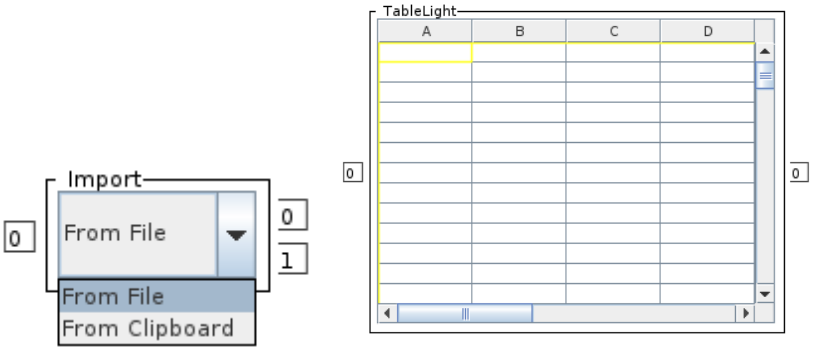
V neposlední řadě je každý blok zodpovědný za provedení jeho účelové funkce — zpracování vstupních dat do výstupní formy, načtení externích zdrojů, vizualizace dat atd. Každý blok má tedy ve své podstatě 3 hlavní úkoly:

1. Provést svou funkci
2. Vizualizovat své rozhraní
3. Reagovat na změny parametrů a nutnost přepočtu

Z tohoto důvodu a pro zajištění modularity je použit návrhový vzor *MVC* pro interní návrh a implementaci jednotlivých bloků, který rozčleňuje implementaci v souladu s dříve zmíněnými úkoly bloků. Tento návrhový vzor rozčleňuje kód do tří komponent:



Obrázek 3.1: Příklad rozhraní pro manipulaci s parametry bloků; vlevo parametry bloku pro načtení obrazu, vpravo pro práci s tabulkami



Obrázek 3.2: Příklad vizualizace bloků; vlevo blok načtení obrazu, vpravo pro práci s tabulkami

1.	`"x"`	`"y"`	`"šířka"`	`"výška"`
2.	x_1	y_1	w_1	h_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$n+1.$	x_n	y_n	w_n	h_n

Tabulka 3.1: Formát tabulky se souřadnicemi oblastí

1. **Model** — stará se o reprezentaci informací a jejich zpracování
2. **View** — má za úkol prezentovat data získaná od modelu
3. **Controller** — reagující na události a zajišťuje změny

Metoda pro segmentaci obrazu dokumentu tedy bude implementována jako jeden z těchto bloků. Vstupem tohoto bloku bude obraz strany dokumentu, nad kterou bude provedena segmentace. Výstupem budou získané segmenty v podobě označených oblastí v obraze dokumentu a v podobě tabulky souřadnic těchto oblastí. Formát této tabulky je popsán v tabulce 3.1.

3.1 Externí knihovny

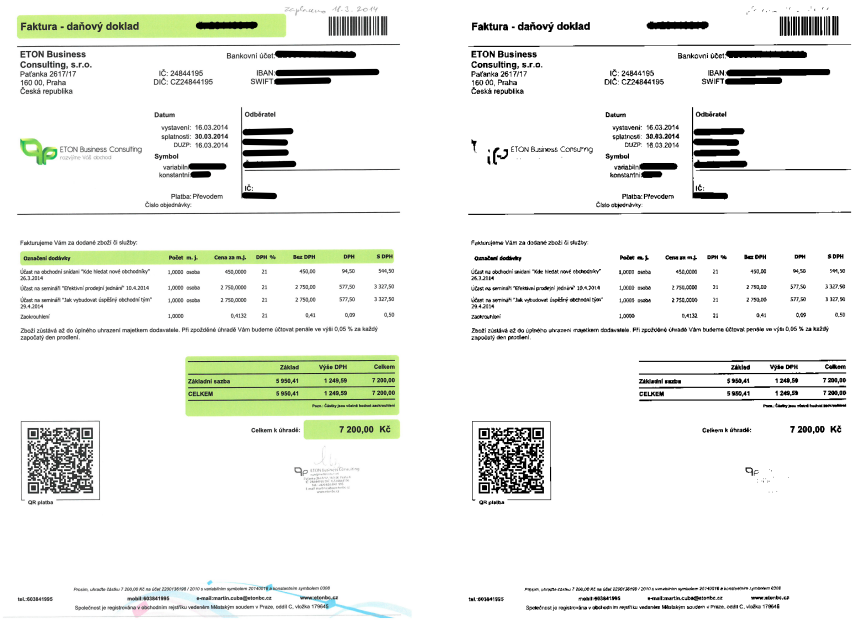
3.1.1 Využití knihovny OpenCV

V rámci projektu *OBBB* se pro práci s obrazovými daty obecně i pro specifické úlohy pro zpracování obrazu využívá svobodně dostupná knihovna pro počítačové vidění *OpenCV* [26], která je napsaná v jazyce C++ a v Javě je dostupná pomocí wrapperu. Tato knihovna obsahuje mnoho užitečných funkcí pro úpravu, zpracování a analýzu obrazu a je v rámci projektu hojně využívána.

V rámci implementace metody byly využity především definované datové typy pro reprezentaci a manipulaci s obrazovými daty v podobě třídy `Mat`, která je základním typem popisujícím obraz ve formě bitové mapy předem definované hloubky a rozměru.

Dále byly využity funkce pro změnu rozměrů takto reprezentovaných obrazových dat poskytující několik používaných metod pro interpolaci hodnot. Pro obraz dokumentu byla zjištěna metoda interpolace pomocí oblasti, která využívá převzorkování pomocí vztahu pixelu a okolní oblasti. Tato metoda poskytuje dobré výsledky při zmenšování obrazu, jelikož nedochází k přílišné ztrátě informace a je odolná vůči moaré efektu [1]. V případě zvětšování obrazu dosahuje obdobných výsledků jako interpolace pomocí nejbližšího souseda a tedy nedochází k přílišnému šíření šumu. Standardní velikost obrazu je v základu nastavena na 800 horizontálních pixelů (dojde k zachování poměru stran), a tedy u převážné většiny moderních scannerů je tato hodnota

[1] více informací lze nalézt na `http://cs.wikipedia.org/wiki/Moaré`



Obrázek 3.3: Výsledek binarize s hladinou zjištěnou pomocí Otsuovy metody

významně menší, než je rozlišení přístroje, a proto by ve většině případů mělo docházet k případu zmenšování obrazu, v čemž zvolená metoda interpolace exceluje.

Další využitou funkcí je transformace obrazu do šedého spektra, implementace Otsuovy [22] metody hledání hladiny a následné binarizace obrazu. Tato metoda dosahuje pro dokumenty dobrých výsledků. Nedochází k přílišné ztrátě informace, pokud je obraz dokumentu dostatečně kontrastní a zároveň tato metoda eliminuje dostatečné množství šumu. Výsledky binarize obrazu dokumentu pomocí této metody lze vidět na obrázku 3.3.

Další využitou funkcí knihovny *OpenCV* je metoda pro vyhledání obrysů ve výsledku *RLSA* na binarizovaném obraze. Tato funkce nalezne všechny obrysy, které obraz obsahuje, a uspořádá je do hierarchického stromu (z důvodu možnosti obrysu uvnitř obrysu). Pro účel dalšího zpracování mají však využití pouze obrysy v nejvyšší vrstvě této hierarchie, a proto dojde k jejich filtraci. Nepotřebné obrysy se zahodí. Pro každý získaný obrys se dále spočte jeho ohraničující obdélník pomocí k tomu určené knihovní funkce.

3.1.2 Aglomerativní hierarchické shlukování

Pro aglomerativní hierarchické shlukování je využita Behnkeho implementace [27], která podporuje 3 metody výpočtu vzdálenosti jednotlivých shluků:

1. metoda nejbližšího sousede

2. metoda nejvzdálenějšího souseda

3. metoda průměrné vazby

Tato implementace bere na vstup dříve popsanou matici vzdáleností a její pomocí vytvoří hierarchii shluků reprezentovanou kořenovým shlukem, která byla sestavena pomocí vybrané metody výpočtu vzdáleností shluků.

3.2 Objektový návrh metody

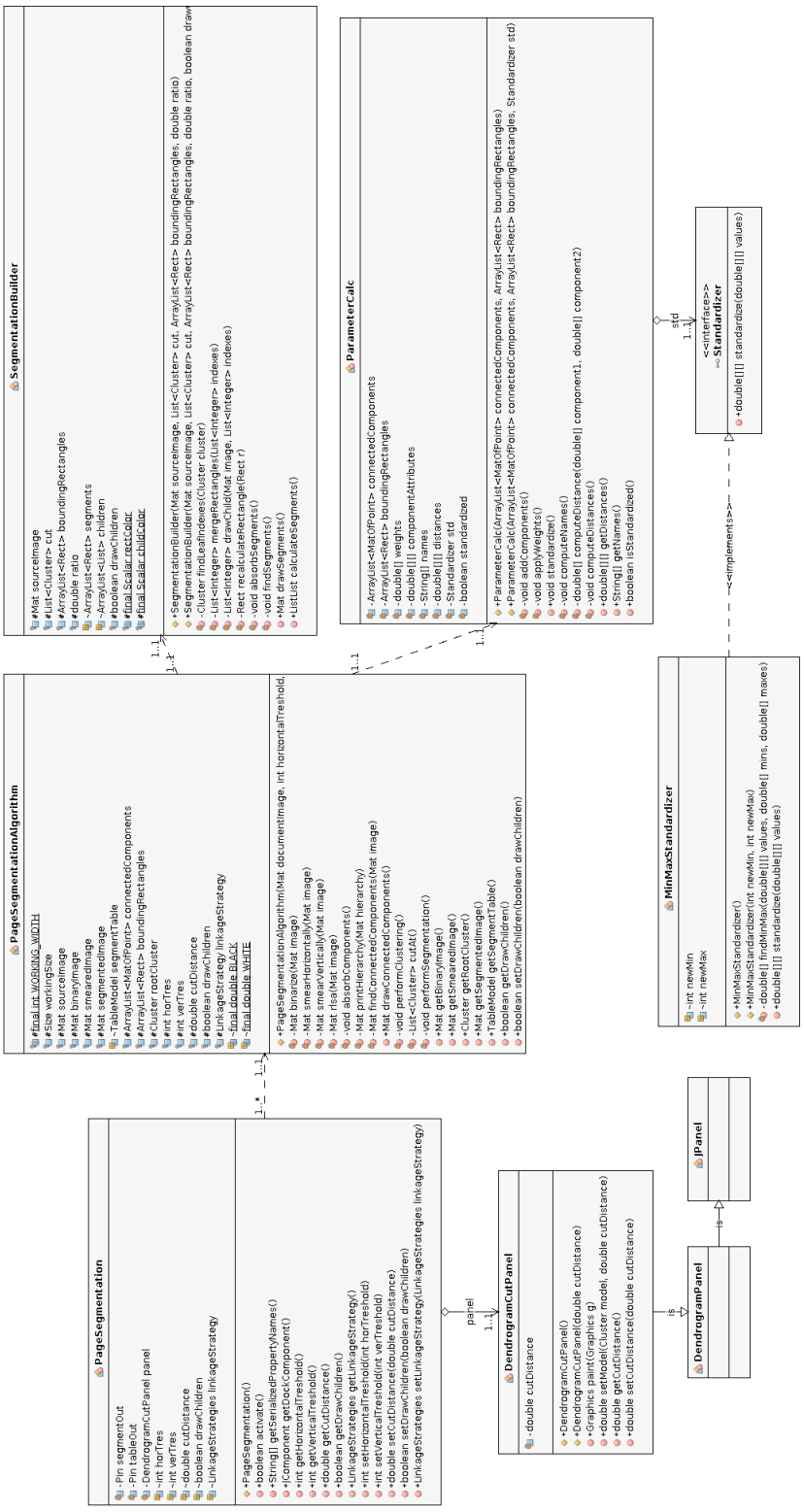
Zde bude probráno rozložení implementace do jednotlivých objektů a jejich příslušnosti ke komponentám *MVC*.

Základní rozdělení objektů bylo navrženo dle specifikací projektu *OBBB*, tedy primární třídou je `PageSegmentation`, která definuje vstupy, výstupy a jejich datové typy. Dále je také zodpovědná za uchování parametrů pro blok a umožnění jejich změny. Poskytuje také informace o tom, jaké parametry schraňuje. V neposlední řadě také obstarává reakci na aktivaci (spuštění) bloku a delegaci výsledků na výstupy. Svou funkčností tedy spadá do komponenty **controller** architektury *MVC*.

Tento objekt vytváří na zažádání instanci třídy `DendrogramCutPanel`, který má za úkol zprostředkování vizualizace bloku. Tento objekt je podtřída `JPanel`, který slouží pro `javax.swing` především jako kreslící plátno či jako agregátor pro další grafické komponenty. Tato podtřída zajišťuje vykreslení dendrogramu pro předanou hierarchii shluků, a také vizualizuje hladinu řezu této hierarchie. Spadá tedy do komponenty **view** architektury *MVC*.

Následující třídy spadají do komponenty **model** architektury *MVC* a zajišťují logiku pro předzpracování obrazu a segmentaci strany. Prvním je třída `PageSegmentationAlgorithm`, která zajišťuje hlavní část výpočtu segmentace strany. Tato třída jako jakýsi agregátor dat zpracovaných různými algoritmy, případně tyto data sama upravuje. V rámci této třídy se provedou veškeré nutné kroky pro segmentaci obrazu. Další třídou je `ParameterCalc`, která se stará o extrakci parametrů z nalezených komponent. Tyto parametry poté modifikuje do požadované formy pro další zpracování. Do této modifikace patří normalizace hodnot parametrů, výpočet vzdáleností a sestavení matice vzdáleností mezi jednotlivými komponentami. Poslední třídou je `SegmentationBuilder`. Tato třída se stará o transformaci předané množiny shluků na samotné oblasti. Tyto oblasti poté poskytuje ve formě tabulky, jejíž formát byl popsán na počátku této kapitoly, nebo ve formě nakreslených obdélníků navrchu zpracovávaného obrazu dokumentu.

Na obrázku 3.4 lze vidět *UML* diagram popsaných tříd a jejich vzájemných vztahů.



Obrázek 3.4: UML diagram vytvořených tříd

Kapitola 4

Diskuse výsledků

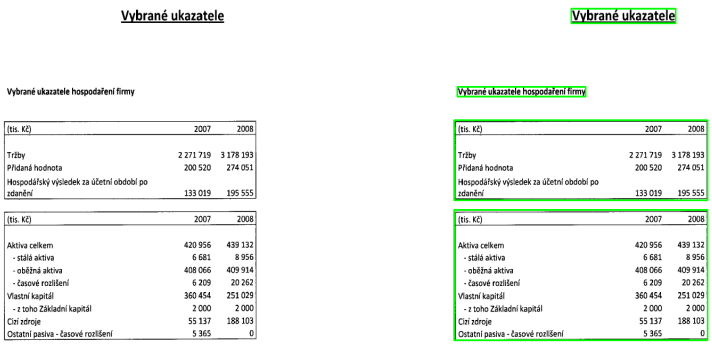
V této kapitole bude probráno jakých výsledků implementovaná metoda dosahuje. V jakých případech identifikuje oblasti správně, v jakých naopak selže. Dále budou navrhnuty úpravy pro odstranění případů selhání, či případných vylepšení.

4.1 Popis výsledků

Dokumenty na obrázcích 4.1, 4.2 a 4.3 jsou příklady dokumentů s jednoduchým rozvržením. Jak lze vidět metoda na těchto typech dokumentů dosahuje dostačujících výsledků. Identifikované oblasti dávají smysl a jejich obsah má opravdu související význam. Dokument na obrázku 4.1 je díky svému formátování (ohraničení v tabulce) velmi jednoduše segmentovatelný, neboť hlavní část segmentace v tomto případě provede *RLSA* a na to navazující hledání obrysů. V případě dokumentu na obrázku 4.2 je díky vodícím čarám způsob segmentace obdobný, jelikož významná část objektů je seskupena před proběhnutím shlukové analýzy. Dokument na obrázku 4.3 již žádné takovéto pomocné prvky nemá a seskupování je tedy prací především shlukové analýzy. Avšak i v tomto případě se metoda dobrala dobrého výsledku. Při menší zvolené hladině řezu je výsledek stále smysluplný, ale určitým způsobem chaotický.

Dokumenty na obrázcích 4.4 a 4.5 jsou komplikovanějšího rozvržení, ale opět díky ohraničujícím čarám metoda správně identifikovala oblasti podobného významu správně. Na obrázku 4.5 je vidět, že šedé pozadí bylo bráno jako textová informace a bylo seskupeno společně s textem.

Na fakturách na obrázcích 4.6 a 4.7 se objevují nesrovnalosti v některých oblastech, kdy komponenta intuitivně náležící do jedné oblasti byla shluknuta do jiné, či dva shluky visuální i pozicí si velmi podobné jsou rozděleny do samostatných oblastí. Na obrázku 4.7 je ilustrována snaha o seskupení visuálně si podobných oblastí pomocí zvýšení hladiny pro řez. Tento krok však vede i k seskupení dalších a nežádoucích oblastí.



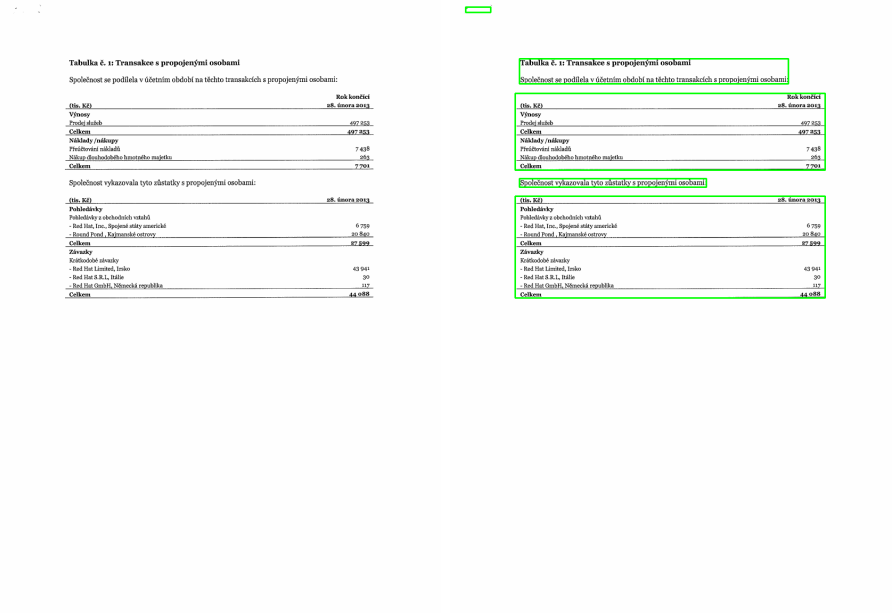
Vybrané ukazatele

Vybrané ukazatele hospodaření firmy

(tis. Kč)	2007	2008
Tržby	2 271 719	3 178 193
Přidaná hodnota	200 520	274 051
Hospodářský výsledek za účetní období po zdanění	133 019	195 555

(tis. Kč)	2007	2008
Aktiva celkem	420 956	439 132
- stálá aktiva	6 681	8 956
- oběžná aktiva	408 066	409 914
- časové rozlišení	6 209	20 262
Vlastní kapitál	360 454	251 029
- z toho Základní kapitál	2 000	2 000
Cizí zdroje	55 137	188 103
Ostatní pasiva - časové rozlišení	5 365	0

Obrázek 4.1: Příklad jednoduchého rozvržení s ohraničením



Tabulka č. 1: Transakce s propojenými osobami

Společnost se podílela v účetním období na těchto transakcích s propojenými osobami:

(tis. Kč)	Rok končící 28. února 2013
Výnosy	
Prodej služeb	497 253
Celkem	**497 253**
Náklady /nákupy	
Přeúčtování nákladů	7 438
Nákup dlouhodobého hmotného majetku	263
Celkem	**7 701**

Společnost vykazovala tyto zůstatky s propojenými osobami:

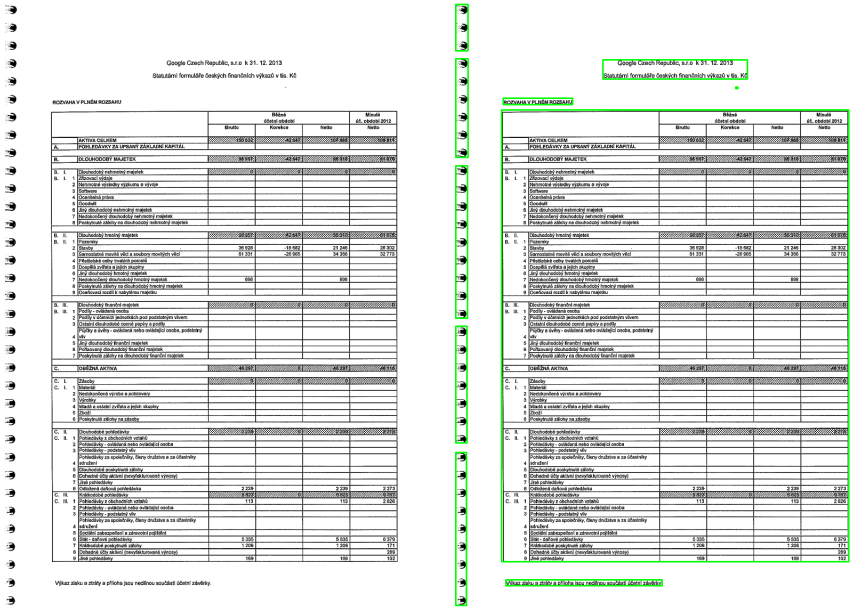
(tis. Kč)	28. února 2013
Pohledávky	
Pohledávky z obchodních vztahů	
- Red Hat, Inc., Spojené státy americké	6 759
- Exozet Prod , Rajtarské ostrovy	20 840
Celkem	**27 599**
Závazky	
Krátkodobé závazky	
- Red Hat Limited, Irsko	43 941
- Red Hat S.R.L., Itálie	30
- Red Hat GmbH, Německá republika	117
Celkem	**44 088**

Obrázek 4.2: Příklad jednoduchého rozvržení s vodícími čarami

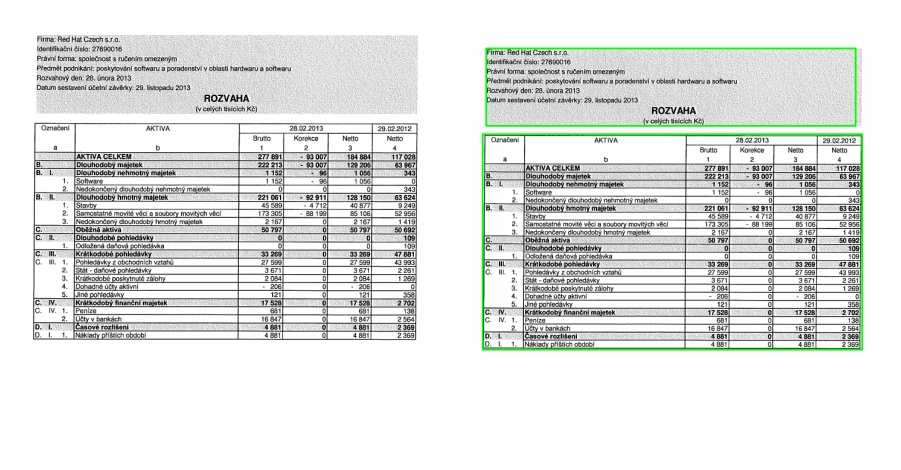
Obrázek 4.3: Příklad jednoduchého rozvržení bez ohraničení; dole proveden řez v menší vzdálenosti

Na obrázcích 4.8 a 4.9 jsou dokumenty jak s plynulým textem, tak s tabulkami. V případě obrázku 4.8 proběhla segmentace strany se vcelku dobrým výsledkem. U obrázku 4.9 byl dokument segmentován na nesmyslné oblasti, které se navíc navzájem protínají. Zde je poznat nevýhoda obecné shlukové analýzy pracující nad extrahovanými parametry. Extrakcí dojde ke ztrátě informace a kontextu jednotlivých komponent.

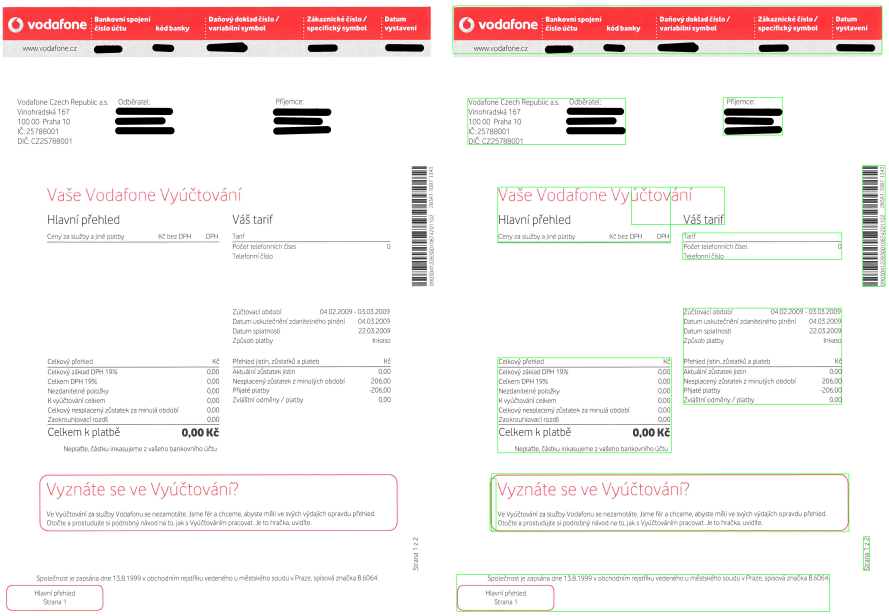
Dokument na obrázku 4.10 poukazuje na problém vznikající v případě, že obraz dokumentu obsahuje nepůvodní grafické objekty, které se poté v rámci shlukové analýzy seskupují společně s původními textovými daty.



Obrázek 4.4: Příklad komplexnějšího rozvržení s ohraničením



Obrázek 4.5: Příklad komplexnějšího rozvržení s šedým pozadím



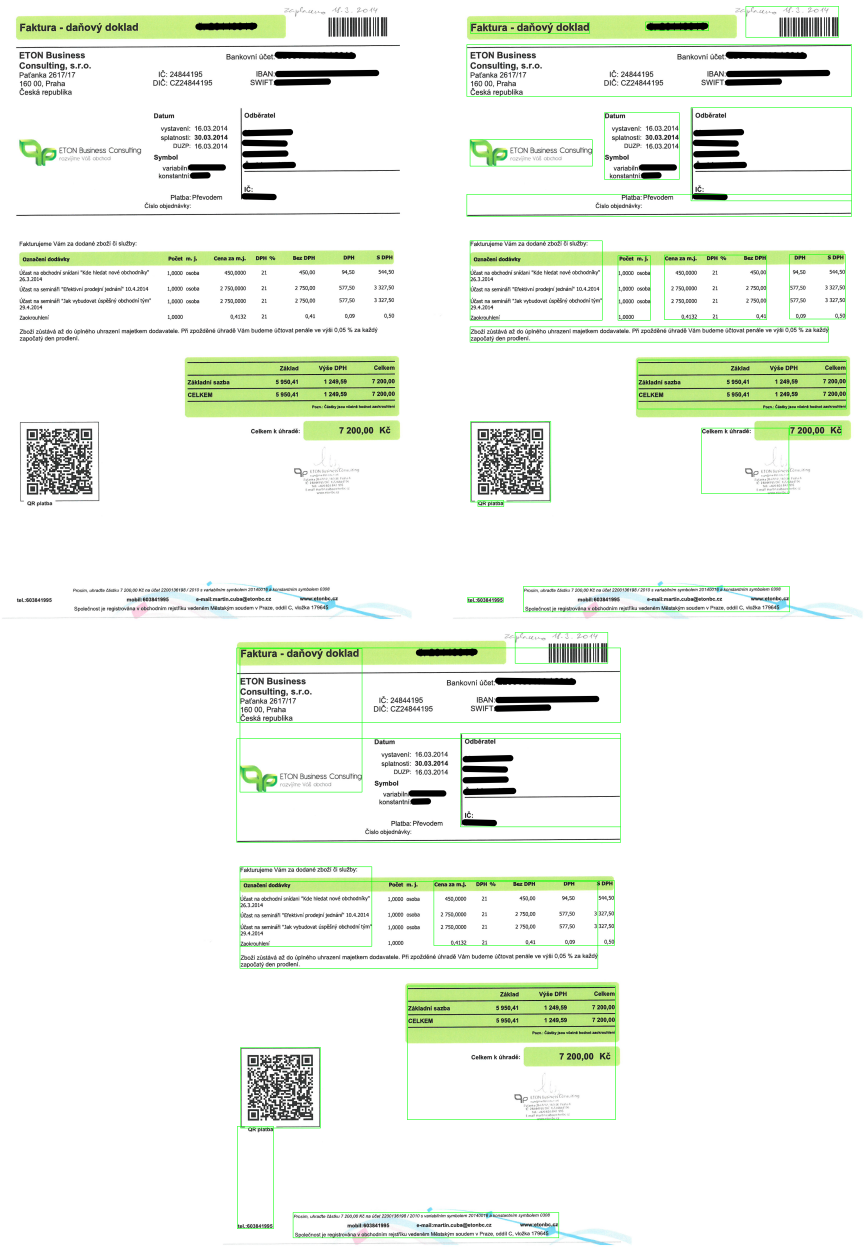
Obrázek 4.6: Příklad faktury s jednoduchým grafickým zpracováním

4.2 Návrh změn a vylepšení

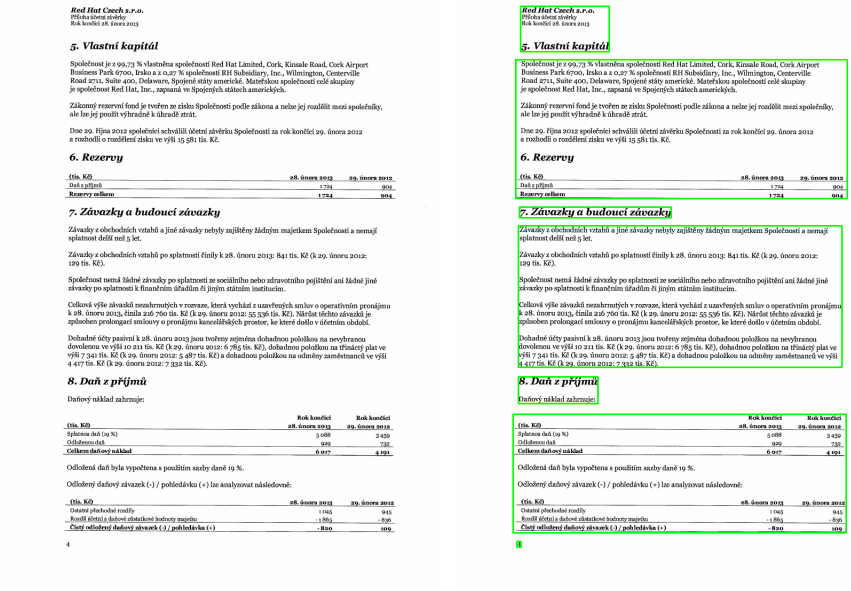
Jedním z možných vylepšení by mohla být klasifikace nalezených komponent do tříd (grafické objekty, text). Poté pomocí těchto tříd odfiltrovat nežádoucí komponenty způsobující chybné seskupené shluky. Případně spustit shlukovou analýzu zvlášť na jednotlivé třídy.

Další změnou by mohla být přeměna obyčejného algoritmu hierarchického shlukování pracujícího čistě nad souřadnicemi z $\mathbb{R}^n$ na jeho jeho informovanější verzi pracující nad samotnými oblastmi, které by se dokázali dynamicky za běhu algoritmu seskupovat a vzájemně mezi sebou dopočítávat jejich vzdálenosti, které by mohli být tím pádem definovány komplexněji — např. vzájemná vzdálenost jejich ohraničujících obdélníků atp.

Další možnost je rozšířit zpracování následující shlukovou analýzu o sjednocování oblastí, jejich vzájemná plocha je větší než definovaná konstanta. Případné „lámání“ oblastí v případě, že se navzájem protínají.



Obrázek 4.7: Příklad faktury se složitějším grafickým zpracováním; dole proveden řez ve větší vzdálenosti



Red Hat Czech s.r.o.
Příloha účetní závěrky
Rok končící 28. února 2013

5. Vlastní kapitál

Společnost je z 99,73 % vlastněna společností Red Hat Limited, Cork, Kinsale Road, Cork Airport Business Park 6700, Irsko a z 0,27 % společností RH Subsidiary, Inc., Wilmington, Centerville Road 2711, Suite 400, Delaware, Spojené státy americké. Mateřskou společností celé skupiny je společnost Red Hat, Inc., zapsaná ve Spojených státech amerických.

Zákonný rezervní fond je tvořen ze zisku Společnosti podle zákona a nelze jej rozdělit mezi společníky, ale lze jej použít výhradně k úhradě ztrát.

Dne 29. října 2012 společníci schválili účetní závěrku Společnosti za rok končící 29. února 2012 a rozhodli o rozdělení zisku ve výši 15 581 tis. Kč.

6. Rezervy

(tis. Kč)	28. února 2013	29. února 2012
Daň z příjmů	1 724	904
Rezervy celkem	**1 724**	**904**

7. Závazky a budoucí závazky

Závazky z obchodních vztahů a jiné závazky nebyly zajištěny žádným majetkem Společnosti a nemají splatnost delší než 5 let.

Závazky z obchodních vztahů po splatnosti činily k 28. únoru 2013: 841 tis. Kč (k 29. únoru 2012: 129 tis. Kč).

Společnost nemá žádné závazky po splatnosti ze sociálního nebo zdravotního pojištění ani žádné jiné závazky po splatnosti k finančním úřadům či jiným státním institucím.

Celková výše závazků nezahrnutých v rozvaze, která vychází z uzavřených smluv o operativním pronájmu k 28. únoru 2013, činila 226 760 tis. Kč (k 29. únoru 2012: 55 536 tis. Kč). Nárůst těchto závazků je způsoben prolongací smlouvy o pronájmu kancelářských prostor, ke které došlo v účetním období.

Dohadné účty pasivní k 28. únoru 2013 jsou tvořeny zejména dohadnou položkou na nevybranou dovolenou ve výši 10 211 tis. Kč (k 29. únoru 2012: 6 765 tis. Kč), dohadnou položkou na přináležitý plat ve výši 7 341 tis. Kč (k 29. únoru 2012: 5 487 tis. Kč) a dohadnou položkou na odměny zaměstnanců ve výši 4 417 tis. Kč (k 29. únoru 2012: 7 332 tis. Kč).

8. Daň z příjmů

Daňový náklad zahrnuje:

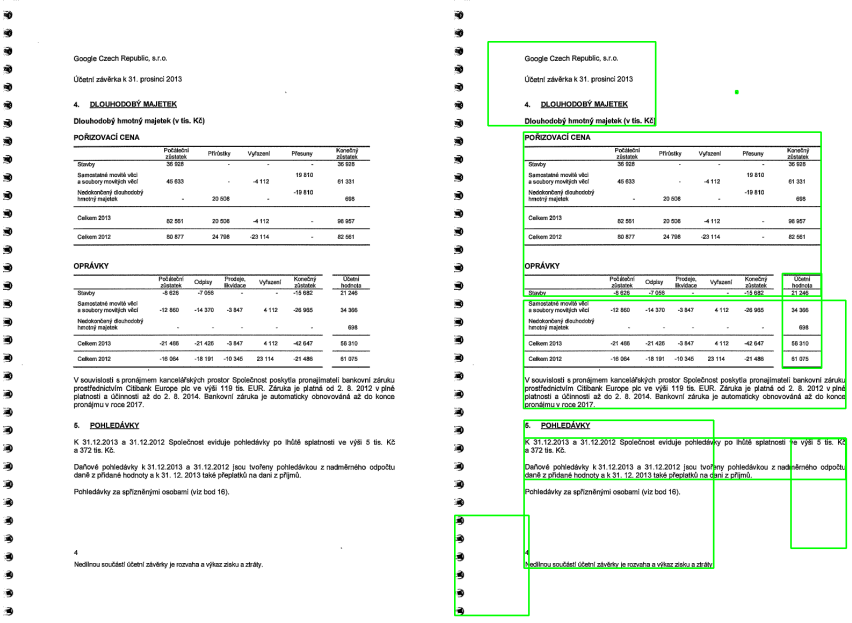
(tis. Kč)	Rok končící 28. února 2013	Rok končící 29. února 2012
Splatná daň (19 %)	5 088	3 459
Odložená daň	939	736
Celkem daňový náklad	**6 027**	**4 195**

Odložená daň byla vypočtena s použitím sazby daně 19 %.

Odložený daňový závazek (-) / pohledávku (+) lze analyzovat následovně:

(tis. Kč)	28. února 2013	29. února 2012
Ostatní přechodné rozdíly	1 043	943
Rozdíl účetní a daňové zůstatkové hodnoty majetku	- 1 863	- 839
Čistý odložený daňový závazek (-) / pohledávka (+)	**- 820**	**104**

4



Google Czech Republic, s.r.o.

Účetní závěrka k 31. prosinci 2013

4. DLOUHODOBÝ MAJETEK

Dlouhodobý hmotný majetek (v tis. Kč)

POŘIZOVACÍ CENA

	Počáteční zůstatek	Přírůstky	Vyřazení	Přesuny	Konečný zůstatek
Stavby	36 928	-	-	-	36 928
Samostatné movité věci a soubory movitých věcí	46 633	-	-4 112	19 810	61 331
Nedokončený dlouhodobý hmotný majetek	-	20 606	-	-19 810	696
Celkem 2013	82 561	20 606	-4 112	-	98 957
Celkem 2012	60 877	24 798	-23 114	-	82 561

OPRÁVKY

	Počáteční zůstatek	Odpisy	Prodeje, likvidace	Vyřazení	Konečný zůstatek	Účetní hodnota
Stavby	-6 626	-7 056	-	-	-13 682	21 246
Samostatné movité věci a soubory movitých věcí	-12 860	-14 370	-3 847	4 112	-26 955	34 366
Nedokončený dlouhodobý hmotný majetek	-	-	-	-	-	696
Celkem 2013	-21 486	-21 426	-3 847	4 112	-42 647	56 310
Celkem 2012	-16 064	-16 191	-10 345	23 114	-21 486	61 075

V souvislosti s pronájmem kancelářských prostor Společnost poskytla pronajímateli bankovní záruku prostřednictvím Citibank Europe plc ve výši 119 tis. EUR. Záruka je platná od 2. 8. 2012 v plné platnosti a účinnosti až do 2. 8. 2014. Bankovní záruka je automaticky obnovována až do konce pronájmu v roce 2017.

5. POHLEDÁVKY

K 31.12.2013 a 31.12.2012 Společnost eviduje pohledávky po lhůtě splatnosti ve výši 5 tis. Kč a 372 tis. Kč.

Daňové pohledávky k 31.12.2013 a 31.12.2012 jsou tvořeny pohledávkou z nadměrného odpočtu daně z přidané hodnoty a k 31. 12. 2013 také přeplatků na dani z příjmů.

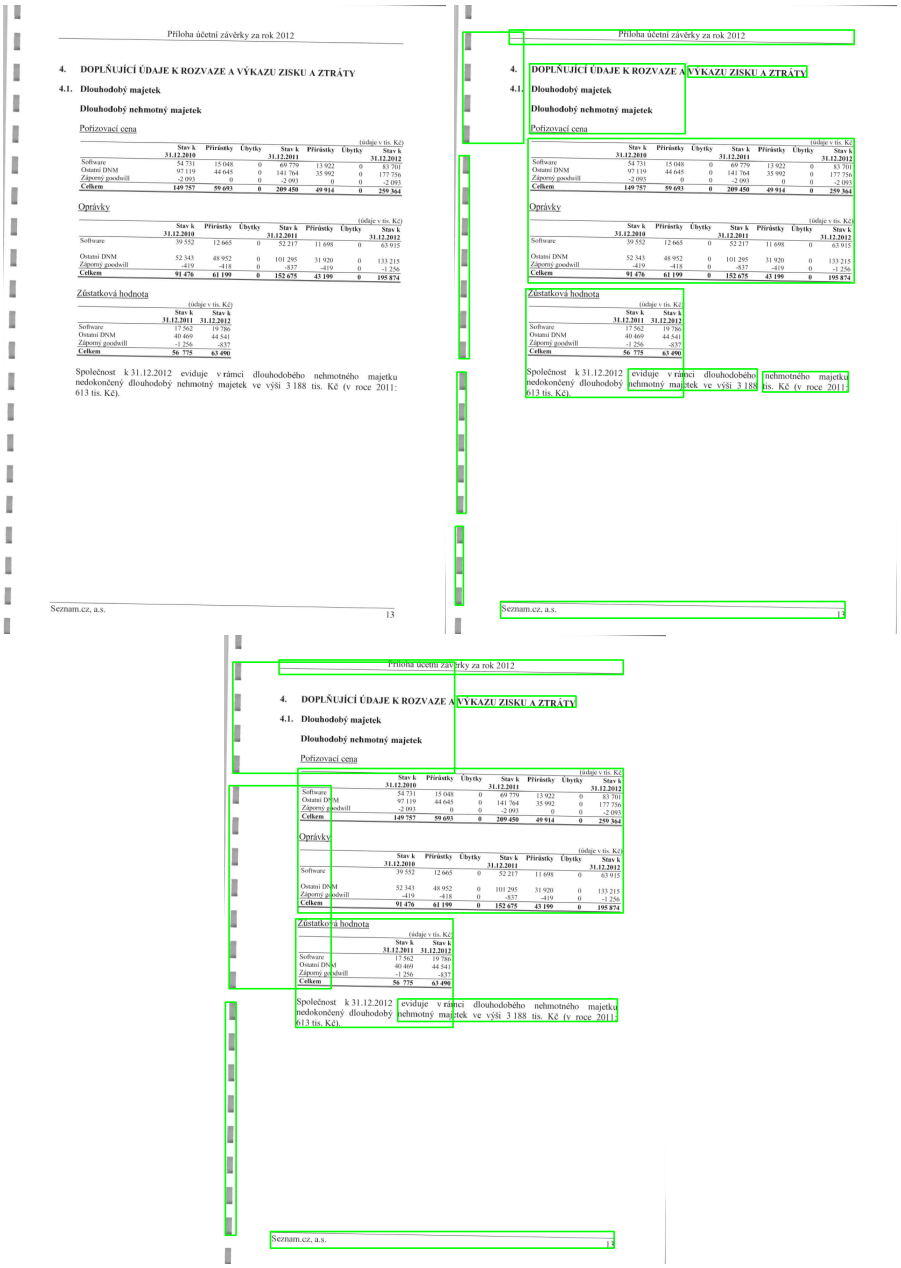
Pohledávky za spřízněnými osobami (viz bod 16).

4

Nedílnou součástí účetní závěrky je rozvaha a výkaz zisku a ztráty.

Obrázek 4.8: Příklad rozvržení s textem a tabulkami

Obrázek 4.9: Příklad rozvržení s textem a tabulkami 2



Obrázek 4.10: Příklad rozvržení obsahující nepůvodní grafické objekty

Závěr

Pro účely této práce byla provedena rešerše problému segmentace strany a z jejích poznatků byla navržena metoda řešící tento problém pro formální dokumenty.

Tato metoda byla úspěšně naimplementována nad projektem *OBBB* v jazyce Java a pro určité typy dokumentů se ukázala jako dostačující. Úspěšně identifikovala oblasti v dokumentu vzájemně souvisejících dat. Její výstupy lze flexibilně upravovat pomocí parametrů, ovlivňující její průběh, a tím dosáhnout požadovaného výsledku, ne však za všech okolností.

U určitých dokumentů se nejeví identifikace oblastí jako příliš spolehlivá a objevili se i typy dokumentů, na kterých selhala, neboť vyprodukovala poněkud chaotické množiny mnohdy navzájem se protínajících oblastí. Tyto problémy nebylo možné vyřešit pouhou změnou parametrů provádění metody a z tohoto důvodu byly navrženy možné úpravy algoritmu pro vyřešení alespoň některých z těchto problémů.

Literatura

[1] Cattoni, R.; Coianiz, T.; Messelodi, S.; aj.: Geometric layout analysis techniques for document image understanding: a review. *ITC-IRST*, January 1998. Dostupné z: `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.105.3741`

[2] Mao, S.; Rosenfeld, A.; Kanungo, T.: Document structure analysis algorithms: a literature survey. In *Document Recognition and Retrieval*, 2003, s. 197–207.

[3] Simon, A.; Pret, J.-C.; Johnson, A.: A fast algorithm for bottom-up document layout analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, ročník 19, č. 3, March 1997: s. 273–277, ISSN 0162-8828, doi:10.1109/34.584106.

[4] Nagy, G.; Seth, S.; Viswanathan, M.: A prototype document image analysis system for technical journals. *Computer*, ročník 25, č. 7, July 1992: s. 10–22, ISSN 0018-9162, doi:10.1109/2.144436.

[5] Breuel, T.: Two Geometric Algorithms for Layout Analysis. In *Document Analysis Systems V*, *Lecture Notes in Computer Science*, ročník 2423, editace D. Lopresti; J. Hu; R. Kashi, Springer Berlin Heidelberg, 2002, ISBN 978-3-540-44068-0, s. 188–199, doi:10.1007/3-540-45869-7_23. Dostupné z: `http://dx.doi.org/10.1007/3-540-45869-7_23`

[6] Ferilli, S.; Biba, M.; Esposito, F.; aj.: A Distance-Based Technique for Non-Manhattan Layout Analysis. In *Document Analysis and Recognition, 2009. ICDAR '09. 10th International Conference on*, July 2009, ISSN 1520-5363, s. 231–235, doi:10.1109/ICDAR.2009.37.

[7] Shih, F.; Chen, S.-S.; Hung, D.; aj.: A document segmentation, classification and recognition system. In *Systems Integration, 1992. ICSI '92., Proceedings of the Second International Conference on*, Jun 1992, s. 258–267, doi:10.1109/ICSI.1992.217295.

[8] Zlatopolsky, A.: Automated document segmentation. *Pattern Recognition Letters*, ročník 15, č. 7, 1994: s. 699 – 704, ISSN 0167-8655, doi: http://dx.doi.org/10.1016/0167-8655(94)90074-4. Dostupné z: `http://www.sciencedirect.com/science/article/pii/0167865594900744`

[9] Wahl, F. M.; Wong, K. Y.; Casey, R. G.: Block segmentation and text extraction in mixed text/image documents. *Computer Graphics and Image Processing*, ročník 20, č. 4, 1982: s. 375 – 390, ISSN 0146-664X, doi: http://dx.doi.org/10.1016/0146-664X(82)90059-4. Dostupné z: `http://www.sciencedirect.com/science/article/pii/0146664X82900594`

[10] Rege, P. P.; Chandrakar, C. A.: Text-Image Separation in Document Images Using Boundary/Perimeter Detection. *ACEEE International Journal on Signal and Image Processing*, ročník 3, č. 1, January 2012: str. 5. Dostupné z: `http://doi.searchdl.org/01.IJSIP.3.1.70`

[11] Sun, H.-M.: Page segmentation for Manhattan and non-Manhattan layout documents via selective CRLA. In *Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on*, August 2005, ISSN 1520-5263, s. 116–120 Vol. 1, doi:10.1109/ICDAR.2005.185.

[12] O'Gorman, L.: The Document Spectrum for Page Layout Analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, ročník 15, č. 11, November 1993: s. 1162–1173, ISSN 0162-8828, doi:10.1109/34.244677. Dostupné z: `http://dx.doi.org/10.1109/34.244677`

[13] Hirayama, Y.: A block segmentation method for document images with complicated column structures. In *Document Analysis and Recognition, 1993., Proceedings of the Second International Conference on*, October 1993, s. 91–94, doi:10.1109/ICDAR.1993.395775.

[14] Okamoto, M.; Takahashi, M.: A hybrid page segmentation method. In *Document Analysis and Recognition, 1993., Proceedings of the Second International Conference on*, October 1993, s. 743–746, doi:10.1109/ICDAR.1993.395630.

[15] Esposito, F.; Malerba, D.; Semeraro, G.: A knowledge-based approach to the layout analysis. In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, ročník 1, August 1995, s. 466–471 vol.1, doi:10.1109/ICDAR.1995.599037.

[16] Lin, M. W.; Tapamo, J. R.; Ndovie, B.: A texture-based method for document segmentation and classification. *South African Computer Journal*, ročník 36, 2006: s. 49–56. Dostupné z: `http://www-direction.inria.fr/international/arima/006/pdf/arima00604.pdf`

[17] Tang, Y.; Ma, H.; Mao, X.; aj.: A new approach to document analysis based on modified fractal signature. In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, ročník 2, August 1995, s. 567–570 vol.2, doi:10.1109/ICDAR.1995.601960.

[18] Kise, K.; Sato, A.; Iwata, M.: Segmentation of Page Images Using the Area Voronoi Diagram. *Computer Vision and Image Understanding*, ročník 70, č. 3, 1998: s. 370 – 382, ISSN 1077-3142, doi:http://dx.doi.org/10.1006/cviu.1998.0684. Dostupné z: `http://www.sciencedirect.com/science/article/pii/S1077314298906841`

[19] O'Gorman, L.; Kasturi, R.: *Document Image Analysis*. IEEE Computer Society Press, 1995.

[20] Don, H.-S.: A noise attribute thresholding method for document image binarization. In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, ročník 1, August 1995, s. 231–234 vol.1, doi:10.1109/ICDAR.1995.598983.

[21] Liu, Y.; Srihari, S.: Document image binarization based on texture features. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, ročník 19, č. 5, May 1997: s. 540–544, ISSN 0162-8828, doi:10.1109/34.589217.

[22] Otsu, N.: A Threshold Selection Method from Gray-Level Histograms. *Systems, Man and Cybernetics, IEEE Transactions on*, ročník 9, č. 1, January 1979: s. 62–66, ISSN 0018-9472, doi:10.1109/TSMC.1979.4310076.

[23] Kordík, P.; Borkovec, J.: BI-VZD - Přednáška č. 3 - Předzpracování dat. Září 2011, [cit. 2015-05-10].

[24] OBBB: Open Black Box Builder [software]. 2011–. Dostupné z: `https://sourceforge.net/projects/obbb/`

[25] Oracle Corporation: NetBeans Platform [software]. [přístup 10. května 2015]. Dostupné z: `https://netbeans.org/features/platform/`

[26] Itseez: OpenCV: Open-source Computer Vision [software]. [přístup 10. května 2015]. Dostupné z: `http://opencv.org`

[27] Behnke, L.: hierarchical-clustering-java [software]. [přístup 10. května 2015]. Dostupné z: `https://github.com/lbehnke/hierarchical-clustering-java`

Příloha **A**

Seznam použitých zkratek

OCR Optical Character Recognition

RLSA Run-Length Smearing Algorithm

OBBB Open Black Box Builder

MVC Model View Controller

OpenCV Open-source Computer Vision

UML Unified Modeling Language

Příloha **B**

Obsah přiloženého CD

- `README.txt` ..... stručný popis obsahu CD
- `src`
  - `impl` ..... zdrojové kódy implementace
  - `thesis` ..... zdrojová forma práce ve formátu LaTeX
- `dist`
  - `impro2`
    - `bin` ..... adresář se spustitelnou formou implementace
- `text` ..... text práce
  - `ZZP.pdf` ..... zadání závěrečné práce
  - `BP_Haifler_Frantisek_2015.pdf` ..... text práce ve formátu PDF