

I. IDENTIFIKAČNÍ ÚDAJE

Název práce:	Analýza obrazu gelové elektroforézy
Jméno autora:	Josef Grus
Typ práce:	bakalářská
Fakulta/ústav:	Fakulta elektrotechnická (FEL)
Katedra/ústav:	Katedra kybernetiky
Oponent práce:	Radim Šára
Pracoviště oponenta práce:	Katedra kybernetiky

II. HODNOCENÍ JEDNOTLIVÝCH KRITÉRIÍ

Zadání	náročnější
<i>Hodnocení náročnosti zadání závěrečné práce.</i>	
Zadání patří k náročnějším, protože bylo nutno správně vyřešit řadu technických detailů a ověřit, že zvolená řešení uspokojivě fungují. Bylo nutno se vyrovnat s poměrně špatnou kvalitou vstupních dat.	

Splnění zadání	splněno s menšími výhradami
<i>Posuďte, zda předložená závěrečná práce splňuje zadání. V komentáři případně uveďte body zadání, které nebyly zcela splněny, nebo zda je práce oproti zadání rozšířena. Nebylo-li zadání zcela splněno, pokuste se posoudit závažnost, dopady a případně i příčiny jednotlivých nedostatků.</i>	
Zadání bylo formálně splněno. V práci ale postrádám alespoň pokus o kritický přehled publikovaných řešení pro automatické zpracování elektroforetických obrázků a alespoň kvalitativní srovnání výsledků navrženého řešení s výsledky jiných řešení. Dále postrádám kvalitní dokumentaci použitých dat.	

Zvolený postup řešení	částečně vhodný
<i>Posuďte, zda student zvolil správný postup nebo metody řešení.</i>	
Student se snažil nalézt vhodná řešení pro jednotlivé podproblémy a věnoval se systematicky ověření funkčnosti modelů a vyhodnocení kvality dílčích výsledků. Podle mého názoru ale měl věnovat více času porozumění podstatě vzniku elektroforetických dat, vlastnostem výsledných obrazů a příčinám pozorovaných distorzí a tyto znalosti využít v konstrukci modelu (viz ještě detailní poznámky dále).	
Hlavní výhrady k metodologii:	
<ul style="list-style-type: none"> Není popsáno, jakým způsobem vznikla data. Kdo je vytvořil, za jakých podmínek, kolik bylo obrázků, kolik v od každého typu podle počtu referenčních žebříčků apod. Jak byla data rozdělena do trénovacích, validačních a testovacích podmnožin? Značky v referenčních žebříčcích tvoří uspořádanou posloupnost, stejně tak detekce. Je tedy na místě použít tuto vlastnost v modelu detekce. Vede to na úlohu dynamického programování. Troufám si říci, že řešení pak bude nejen splňovat přirozené podmínky, ale algoritmus bude i řádově rychlejší než relaxovaná úloha ILP, či maďarský algoritmus (který, mimochodem, nepatří k nejrychlejším algoritmům pro bipartitní párování). Závěr, že metoda lineární interpolace je vhodnější než hyperbolický model v problému odhadu počtu bází (kap. 5.5) je ukvapený. Hyperbolická funkce je parametrický model s malým počtem stupňů volnosti, který v sobě obsahuje nezávislou informaci odvozenou z fyzikální podstaty elektroforézy. Tato informace (omezená flexibilita modelu) slouží jako filtr chyb detekce. Oproti tomu popsaná lineární interpolace je neparametrická metoda s velkým počtem stupňů volnosti, která neobsahuje žádnou další informaci a bude trpět přefitováním. Má, pochopitelně, lepší chybu na trénovacích datech. To ale není žádná přednost. Bylo by třeba provést validaci na nezávislém datasetu. Měl bych menší výhrady k použité metodě, kdy podúloha, která vyžaduje více předpokladů (např. detekce značek) se řeší dříve než podúloha, která vyžaduje méně předpokladů (např. korekce rotace obrazu). 	

Další nejasnosti:

- Není jasné, jak byly získány referenční (*ground-truth*) polohy reálných značek pro účely konstrukce korelačního filtru v sekci 4.1.2.
- Není zřejmé, proč jako základní detekční filtr nevyhovuje analytický filtr, například konvoluce gausiánu s intervalovou funkcí, případně náhrada vhodnou mocninou gausiánu.
- Tvzení „[filtr] je schopen detekovat i zkreslené a zašuměné značky“ na str. 16 je v daném kontextu příliš silné (zevšeobecnující) tvrzení, nepodložené dostatečně experimentálními výsledky.
- Není zřejmé, proč variabilita značek vyžaduje použití tak komplikovaného mechanismu detekce jako jsou detektory popsané v sekci 4.2. Není základní problém v tom, že lineární metody jsou z nějakého důvodu nedostatečné a je nutno použít nelineárních metod (to by zdůvodňovalo použití LBP nebo *Haar-like features*).
- Není jasné, co má čtenář vidět v obr. 5.26. Jak se dospělo k pozorování, že značek s počtem bází kolem 1100 leží na úrovni referenční značky s počtem bází 1400? Vzhledem ke zkreslení, vzájemné blízkosti referenčních značek a k šířce značek není pro takový závěr dostatečná evidence.

Není jasná část terminologie:

1. Co jsou to báze? Kde se jejich počty získají? Toto je podstatné pro porozumění významu práce.
2. Co přesně je referenční žebříček a jaké má zaručené vlastnosti?
3. Co je to „bod“ v (3.1)-(3.3)? Je to detekovaná poloha značky?
4. Není jasné, jestli se pod termínem „korelace“ míní normalizovaná korelace, či nikoliv. Pokud ano, proč je nutné normalizovat rozsah hodnot obrazu?
5. Co je to „heatmap“? Korelační mapa?
6. Co se míní výrazem „k vyřešení [problému] byla použita morfologie“. Morfologie čeho? Neměl autor na mysli to, že byly použity metody matematické morfologie?
7. Nerozumím větě „[...] pro běžné provedení rotace není třeba získat úhel pro konstrukci rotační matice“ (str 35). Jak provedeme rotaci bez znalosti úhlu?
8. Co je to „skupina značek“ (sekce 5.3)?
9. Standardní český termín pro *matching* (ve smyslu teorie grafů) je párování, nikoliv „přiřazení“.
10. Co je to *blob* (str. 70)?

Odborná úroveň

B - velmi dobře

Posuďte úroveň odbornosti závěrečné práce, využití znalostí získaných studiem a z odborné literatury, využití podkladů a dat získaných z praxe.

V práci chybí přehled známých řešení.

Formální a jazyková úroveň, rozsah práce

D - uspokojivě

Posuďte správnost používání formálních zápisů obsažených v práci. Posuďte typografickou a jazykovou stránku.

Práce je nepřiměřeně rozsáhlá. Nebylo třeba popisovat metody, které autor zkoušel než našel tu, kterou nakonec použil. Na druhou stranu cením rozsah experimentální části a mezivýsledků, kde se autor snažil zdokumentovat všechny podstatné vlastnosti použitých řešení.

Práce je psána česky, ale s řadou anglicismů, i v případech, kdy to není nutné (např. termín *lane* je bez problému nahraditelný slovem dráha, *threshold* slovem práh, *scale* slovem měřítko, *bias* slovem posun, apod). V práci je nemálo překlepů. Sazba porušuje pravidla technického stylu. Například:

- Reference na číslované vzorce se standardně uvádějí v závorce, např. (2.3) znamená automaticky odkaz na vzorec. Viz např. na str. 18.
- Na konci řádku nemůže být jednopísmenná předložka (např. vícekrát na str. 11).
- Věta nemůže začít číslicí, matematickým symbolem apod (např. v prvním odstavci na str. 45).
- Pokud čísla ve větě jsou 0-9, potom se všechna píší slovy, jinak se všechna píší číslicemi (viz např. první věta na str. 11, ale i jinde).
- Obrázky mají být přednostně nahoře na stránce, pod hlavičkou, případně dole na stránce před zápatím, případně na samostatné stránce, ale téměř nikdy uvnitř stránky mezi odstavci, kde rozbíjejí text, jako např. na str. 15. Každý

obrázek by měl být odkázán referencí v textu, protože není, na rozdíl od číslovaného vzorce, součástí věty.

Další nedostatky:

- Obr. 2.2 by si zasloužil lepší anotaci, kde je v něm například referenční žebříček?
- Ve vzorci (4.6) chybí druhé mocniny u x a y .
- Není jasné, proč musí existovat afinní rektifikační funkce v případě, že výsledkem detekce je přesně detekovaný žebříček (str. 36). Není to tak, že předpokládáme zkreslení, které je v prvním přiblížení afinní?
- Na str. 37 se píše o kriteriální funkci, která je ale definována až na str. 41. Vzniká tak mylný dojem, že funkce (5.11) na str. 37 je onou kriteriální funkcí, o které se na této straně píše.
- Rozbor, definice úlohy a algoritmus by měly být popsány v posloupnosti (1) analýza problému a algoritmizovatelná informace, (2) definice úlohy a kriteriální funkce, (3) algoritmus. Např. kap. 5 tomuto postupu příliš nevyhovuje.
- V Alg. 1 není jasné, proč je počet iterací roven n (počtu značek v žebříčku). Dále není jasné, co jsou to *random data*, ani proč je náhle použit anglický termín.
- Ve vzorci (5.18) není zřejmé, přes které parametry se minimalizuje. Také se zdá, že α, β mají být α_p, β_p . Stejná výtka platí pro (5.19) a (5.22) a i dále.
- Párovací algoritmy, popisované v sekcích (5.3.2)-(5.3.5) nezaručují, navzdory zdánlivě pečlivé formulaci, to, že párování bude monotonní, tj. že přiřazené značky budou následovat ve stejném pořadí jako je jejich vertikální poloha.
- Není jasné, proč se na straně 43 zavádí vyvážený bipartitní graf, když se v (5.21)-(5.25) explicitně pracuje s obecným úplným bipartitním grafem (a popsané algoritmy vyváženost nevyžadují). Co je v daném kontextu m a n ? Nepředpokládá se $n < m$?
- Proč se na str. 45 mluví o reálné ose, když má pouze jít o nezáporná reálná čísla? Nestačí deklarovat $\beta > 0$?
- Ve vzorci (5.44) chybí parametry a, b, c funkce h , zejména když jsou to parametry, přes které se optimalizuje.

Výběr zdrojů, korektnost citací

B - velmi dobře

Vyjádřete se k aktivitě studenta při získávání a využívání studijních materiálů k řešení závěrečné práce. Charakterizujte výběr pramenů. Posuďte, zda student využil všechny relevantní zdroje. Ověřte, zda jsou všechny převzaté prvky řádně odlišeny od vlastních výsledků a úvah, zda nedošlo k porušení citační etiky a zda jsou bibliografické citace úplné a v souladu s citačními zvyklostmi a normami.

Některé zdroje nejsou citovány podle normy, chybí v nich povinné údaje, jako například afiliace autora v případě technické zprávy [24].

Každé tvrzení, které není podepřeno důkazem nebo experimentem musí být doloženo citací. Např. na str. 11 „Častá metoda pro detekci je [...]“. Množství citací je pak mírou „častosti“.

Další komentáře a hodnocení

Vyjádřete se k úrovni dosažených hlavních výsledků závěrečné práce, např. k úrovni teoretických výsledků, nebo k úrovni a funkčnosti technického nebo programového vytvořeného řešení, publikačním výstupům, experimentální zručnosti apod.

Otázky do diskuse:

1. Je možné, že řada technických problémů by zmizela, pokud by data byla pečlivěji digitalizována?
2. Jak dalece vadí, že normalizace na maximum a minimum v (4.1) není robustní, narozdíl, například, od 99% percentilu?
3. Jakou informaci nese amplituda značky?
4. Existuje nějaká souvislost mezi polohou značky a její šířkou?
5. Není jasné, proč hodnota prahu θ v sekci 4.1.3 není určena jako vyšší percentil, místo konvexní (?) kombinace 85% a 100% percentilu.
6. Není jasné, proč se nepoužilo standardní potlačení nemaxim (*non-maximum suppression*) v detekci poloh značek z korelační mapy.
7. Jaký speciální význam má bod na ROC křivce, který je euklidovskými nejbližší hornímu pravému rohu (1,1)?
8. Je-li znám počet sloupců v elektroforetickém obrazu, mělo by být snadné rozlišit správnou šířku filtru v sekci 4.1.5. případně ji s vyhovující přesností určit.
9. Proč se v úloze segmentace na dráhy (*lanes*) nevyužívá znalosti o tom, že jsou všechny stejně široké (a dokonce je jejich šířka apriori poměrně přesně známa)?

10. Maďarský algoritmus a relaxační algoritmus pro maximální bipartitní párování jsou ekvivalentní (poskytují stejné řešení). Proč bylo nutné zkoušet oba? Pokud šlo o rychlost, proč nebyl použit nějaký rychlejší ekvivalentní algoritmus?
11. V závěru se píše, že „cíle byly z většiny splněny“. Které cíle splněny nebyly? Není autor nepřiměřeně sebekritický? Úkol byl nepochybně rozsáhlý, s množstvím technických problémů, způsobených nízkou kvalitou vstupních dat a nebylo možno očekávat, že v krátké době, za kterou je bakalářský projekt nutno zpracovat, bude možno všechny problémy stoprocentně vyřešit.

III. CELKOVÉ HODNOCENÍ, OTÁZKY K OBHAJOBĚ, NÁVRH KLASIFIKACE

Shrňte aspekty závěrečné práce, které nejvíce ovlivnily Vaše celkové hodnocení. Uveďte případné otázky, které by měl student zodpovědět při obhajobě závěrečné práce před komisí.

Student udělal bezesporu velký kus práce. Oceňuji, že víceméně systematicky validoval zvolené modely a algoritmy. Na druhou stranu, v práci nejsou dobře zdokumentována použitá data. Autor se málo věnoval porozumění podstatě vzniku elektroforetických obrázků. Práce neobsahuje srovnání s jinými metodami zpracování těchto obrázků. Autor dále podcenil ale náročnost přípravy textu práce, zejména přesnost popisu metod.

Předloženou závěrečnou práci předběžně hodnotím klasifikačním stupněm **C - dobře**.

Datum: 2.6.2020

Podpis: