Doc. RNDr. Andrej FERKO, PhD., Mlynská dolina, FMFI UK, SK-842 48 Bratislava, Slovakia

doctoral thesis review

Ing. Jiří SKÁLA

Algorithms for manipulating large geometric data

Pilsen: University of West Bohemia 2012, 124 pages

In computer graphics and visualization, there are still many problems open. Among them, there is a general problem – how to process very complex models in real time. The author – under the supervision of Prof. Ivana Kolingerová – focused on a particular subproblem: geometry processing of very large datasets, which cannot be stored in central memory. Such data volumes are produced by laser scanning, medical diagnostics, scientific simulations and become commonplace. The goal of the dissertation is to propose novel data-stream based manipulation using hierarchical clustering and dynamic Delaunay triangulation in 2D and 3D. **The subject of the thesis is relevant to current needs of the scientific community, which is indicated by multiple internationally recognized (and cited) publications and, moreover, the doctoral thesis research is highly actual for practical use, as well.**

The English text of the submitted thesis consists of 9 introductory pages, 9 numbered chapters and two appendices (A More examples..., B Activities), 124 pages in total. The chapters are: 1 Introduction, 2 Large Geometric Data, 3 Data Streams, 4 Clustering, 5 Delaunay triangulation, 6 Summary of the proposed solution, 7 Contributions to clustering, 8 Dynamic hierarchical triangulation, and 9 Conclusion. Bibliography (not given in Contents, pp. 113-124) refers to 146 items. Chapters 2-5 are hierarchically clustered to Part I Background and state of the art. Chapters 6-9 were clustered to Part II Contributions.

The methodology of this research project combines selected advantages of well-studied and understood methods, which are fully adequate and fit perfectly to the main stream reserach in the field (the methodology of mathematic modeling, computational geometry, computer graphics and scientific visualization, applied and focused to special properties of the problem(s) to solve, e.g. the clustering and hierarchy nature, coherency, hidden parallelism, quality metrics...). **The methods used in the thesis have been appropriate and led to a successful solution.**

The main contributions are summarized at pages 57-59. In particular, there are six new methods for clustering (7.1-7.6), three novel ideas in dynamic hierarchical triangulation (8.1-8.3). Some of them indicate a promising future work (e.g. elliptical metrics, p. 80, better cluster expansion, p.102). The goal of the dissertation is to propose novel data-stream based manipulation using hierarchical clustering and dynamic Delaunay triangulation in 2D and 3D. To what extent the main objectives of the work have been fulfilled? **The dissertation fulfills the given goal(s) in full extent.**

The work is done and written in a professional quality. The thesis satisfies conditions of a creative scientific work and there is observable both deep erudition and rich experimental experience of the author. „A new approach to handling huge geometric dat is presented" (p. 12), using facility location, a novel format, storing the complete hierarchy on disc, optimalized streaming, multiresolution DT in planar and spatial case, employing ellliptical metrics to macth

clusters better, extraction of nonconvex shapes in 3D, nearly everything carefully tested on prominent datasets. For hierarchical triangulation the author states, that „there is no directly competitive algorithm" 98/2b. **The original contribution is very innovative, competitive and important/promising.**

However, there are some questionable formulations (or maybe even mistakes). The following notions seem to be undefined: infinite amount of data 11/2 (page 11, line 2), topology of the model 15/2b, compact representation and geometric fidelity 16/6, popping effect 17/1b, error of the projected image 19/10, BRIO 54/8, really gigantic data 58/7, artefacts 90/4b, valid triangle 93/7b. Chapters 2-4 would be desirable to finalize by partial conclusions or discussing the pros and cons. Fig. 8.8 would better display the propety shown having one od Ds outside of the hemisphere. In ref. 81 there is missing the name of the second author Mr. Munkres, p. 119. For multiple, among others refs. 22, 135, 137, there is no date of validity given (IS 690). Another minor typos, missing references or brackelets, and „data is" are marked in the book. These remarks do not decrease the valuable contribution of the work.

The author co-authored 3 referred journal papers (one acceptation pending) and he published 4 papers at international conferences, one technical report and one software library, all of them related with the PhD topic. Publications of the author (p. 110, 9 items, one submitted) appeared in the years 2007-2012 and his research was partly supported by 5 national or international projects. There are already four citations collected. The publications and other academic activities are excellent and clearly over average among PhD projects known to me at multiple universities in Central Europe.

In the discussion, it would be desirable to discuss the selection of the following questions: 1. Why m/k can be considered constant 66/9? 2. Why the Gran Canaria in Fig. 7.11 is assigned to Brasil (Africa is much closer)? 3. What is the user model and GUI, mentioned at 91/4? 4. What is the meaning of dashed edges in fig. 8.7? 5. Which another possibilities (e.g. angular) for face indentification do You consider? There are two only in page 96n. 6. Under what measure of quality is Gabriel properiy identification 97/4 better? 7. „... real data usually is naturally sorted" 99 /21. Hm... What means sorting in 2D and 3D and how would the clustering method work without respect to distances? (There is a mergehull algorithm, where unsorted subsets lead to an efficient solution...)

Conclusion

The author of the doctoral thesis, Ing. Jiří SKÁLA, proved to have an ability to perform research and to achieve scientific results. I do recommend the thesis for presentation with the aim of receiving the Degree of Ph.D.

Bratislava, October 17, 2012

Posudek oponenta

Název disertační práce: Algorithms for manipulating large geometric data

Autor: Ing. Jiří Skála

Vedoucí práce: prof. Dr. Ing. Ivana Kolingerová

Disertační práce Ing. Jiřího Skály je tematicky zaměřena do oblasti 2D/3D aplikované výpočetní geometrie. Autor se zabývá se problematikou manipulace s datovými sadami představovanými rozsáhlou bodovou množinou ($n < 1e9$) s variabilní prostorovou hustotou.

Zvolené téma práce považuji za velmi zajímavé a aktuální. Vzhledem k rychle rostoucím datovým objemům je na trhu výrazná absence produktů zaměřených na zpracování rozsáhlých datových množin. Autorem vytvořené řešení začíná tam, kde končí možnosti dnešních špičkových komerčních softwarů, je zaměřeno na zpracování množin o 1-2 řády větších. S podobným problémem se potýká řada geoinformaticky zaměřených pracovišť při zpracování lidarových dat. Doposud byl realizován převodem vstupních dat do prostorové databáze s následným rozdělením na dlaždice a jejich triangulací. Navrhovaný postup je výrazně efektivnější s poskytuje přidanou hodnotu ve formě multi-scale reprezentace.

Navržené řešení vychází z předpokladu, že data takového rozsahu je nutné zpracovat sekvenčním přístupem, k jednotlivým prvkům nelze přistupovat přímo. Řešení je využitelné pro 2D/3D entity, je jednoprůchodové, bez nutnosti předzpracování vstupních dat, podporuje hierarchický přístup s využitím víceúrovňové clusterizace dat. Autor využívá multi-resolution reprezentaci dat, která umožňuje znázornit scénu s variabilní mírou podrobnosti (LOD) v závislosti na požadavku uživatele či externích kritériích. Praktická realizace je založena na rozdělení vstupního streamu na menší bloky, které lze zpracovat přímo. Na každý z bloků je opakovaně aplikován modifikovaný clusterizační algoritmus s variabilním počtem shluků. Dekompozicí vstupní množiny technikou zdola-nahoru vzniká hierarchická struktura odvozených podmnožin se zmenšující se prostorovou hustotou prvků. Tyto podmnožiny jsou následně triangulovány po clusterech s využitím techniky inkrementálního vkládání (přidání clusteru) a inkrementálního výběru (odebrání clusteru), jsou podporovány nekonvexní množiny. Významná část implementovaných algoritmů je paralelizována a umožňuje zefektivnit vlastní řešení problému.

Disertační práce je strukturována do dvou částí zhruba stejného rozsahu. První je věnována současnému stavu poznání v jednotlivých tematických celcích, druhá část vlastnímu autorskému přínosu k tématu. Tento přístup považuji za velmi dobrý a přehledný, příspěvek autora je snadno zjistitelný a zdokumentovatelný. Rešerše literatury je prováděna zvlášť pro jednotlivá témata, vzhledem k šířce záběru práce to považuji za rozumné.

Kapitola 2 je věnována rozsáhlým vstupním množinám a postupům pro jejich datovou redukci technikami generalizace a LOD, autor se také věnuje problematice hierarchických triangulací, kapitola 3 problematice datových proudů. K úvodním kapitolám mám drobnou připomínku: autor v rešerši uvádí přístupy prostým odkazem na literaturu bez jejich explicitního vyjmenování či širší diskuze o výhodách/nevýhodách: "*The sketching techniques [46, 11, 19, 20, 21, 65] compute summary information on the stream.*" (str. 25). "*A comprehensive overview of clustering can be found, e.g., in [33, 55, 31, 72, 7].*" (str. 28). Bylo by vhodné tyto techniky alespoň vyjmenovat a usnadnit čtenáři orientaci.

Kapitola 4 se zaměřuje na problematiku clusterizace, považuji ji za velmi podrobně a kvalitně zpracovanou. V práci je používán modifikovaný adaptabilní algoritmus založený na k-means, u kterého jako vstupní parametr nemusí být zadáván počet centroidů. Vlastní vytváření shluků je realizováno s využitím kombinace metod teorie grafů a lineárního programování. Iterativní přístup je velmi elegantní, byť výpočetně náročný. Autor se zabývá úvahami zaměřenými na vztah maximální velikosti clusteru k jejich celkovému počtu a vliv tohoto parametru na algoritmus. V dalších částech práce provádí praktické experimenty s těmito parametry... Mezi velikostí clusteru a jejich počtem zřejmě existuje kvadratická závislost. Drobnou připomínku mám k uvedeným vzorcům 4-12 až 4-21, které nepůsobí příliš čitelně. Např. u vzorce 4.12 není jasné, která část výrazu se minimalizuje a není uvedena levá strana. Pokud je součástí vzorce nějaká podmínka, nečísluje se zpravidla samostatně (4-14 až 4-21).

Autor v této kapitole také prezentuje eliptickou metriku umožňující modifikovat standardní metriku v Delaunay triangulaci. Aplikaci této nové metody považuji za velmi zajímavý a novátorský přístup. Uvítal bych, kdyby autor uvedl, jak parametry eliptické normy určit v konkrétním bodě. V textu se vyskytl termín "poloměr elipsy", který je v užitých formulacích poněkud zavádějící "*...any radius of the defining ellipse has a unit length.*" (str. 30).

Za dobře zpracovanou považuji i následující kapitolu o Delaunay triangulaci a jejím streamování. Ačkoliv je první část disertační práce seznámením se stávajícím stavem, přináší teoretické i praktické poznatky na velmi vysoké úrovni reflektující aktuální stav poznání. Hloubkou rešerše autor prokázal, že se v problematice dostatečně orientuje a je schopen zachytit poslední trendy v popisovaných oblastech. Vhledem k velmi širokému záběru práce a množství literatury, kterou autor musel prostudovat, toto považuji významné plus disertační práce.

Další kapitoly práce jsou věnovány teoretickému přínosu disertanta do jednotlivých oblastí, tato část je tematicky rozdělena do tří kapitol. V části věnované clusterizaci se autor zabývá víceúrovňovou hierarchickou clusterizací dat. Autor zde představuje nový přístup volby vah, jak takové převážení spojené s normalizací vypadá, přechází poněkud stručnou formulací "*The goal is to keep the weights around one at higher levels too...*". V takových případech je nutné uvést vztah, pomocí kterého budou váhy určovány. Zvláště to platí, jedná -li se o kapitolu v části "Contribution", kde má být uveden příspěvek autora k tématu. Není mi jasný obr. 7.2, str. 61. Autor v textu uvádí, že na obrázcích a-c) jsou znázorněny 3 bloky (nalezl jsem 4). Na obr. d) mají clustery z předcházející úrovně jiné rozmístění a jejich počet je jiný.

V této části nalezneme spoustu zajímavých a přínosných poznatků, např. odvození vztahu pro velikost bloku v závislosti na velikosti vstupních dat a předpokládaném počtu bodů v clusterech. Proč ve vztahu 7.5 není zkrácena proměnná m nacházející se na obou stranách v první mocnině? Pro otestování nezávislosti clusterizačního algoritmu na řazení vstupních datech použil autor dva postupy: čtení streamu o náhodné délce 10-10 000 prvků a 9900-10 000 s malým procentem outliers, které porovnal se sekvenčním přístupem zpracování. Test považuji za zajímavý a přínosný, výsledky podporují předpoklady autora o invarianci vůči vstupním datům. Zabývá se také odhadem paměťové složitosti algoritmu, v této souvislosti prosím o bližší komentář ke skokovému poklesu funkce pro zadanou velikost bloku v grafu 7.6.

Hierarchie clusterů lze uchovat externě a to v binárním souboru, požadované bloky a clustery mohou být podle potřeby načítány z paměťového média a následně triangulovány. Navrhované řešení se tímto krokem stává obecně použitelným pro téměř libovolná data a umožňuje pracovat s různou mírou detailu scény. Ve vyšší úrovni je skupina bodů jednoho clusteru nahrazena centroidem tohoto clusteru, na nižší úrovni každý z bodů clusteru mohl představovat centroid nějakého clusteru.

Autor analyzuje efektivitu clusterizačního algoritmu a shledává současnou verzi jako úzké hrdlo, které významně ovlivňuje výkon algoritmu jako celku. Navrhuje jeho činnost urychlit několika postupy. První přístup spočívá redukci počtu iterací při adaptabilní clusterizaci, druhý přístup je založen na prostorových indexačních strukturách. Pro clusterizaci byl navržen nový randomizovaný postup, který využívá Hungarian algoritmu (varianta Kuhn-Munkres) umožňující snížit počet iterací z hodnoty $N \log N$ na hodnotu $0.1N$ (v textu uvedeno $O(0.1N)$, což je mírně zavádějící). Následně autor porovnává dosažený výsledek s výsledkem vzniklým plným počtem iterací a dochází k závěru, že se průměrný přírůstek váhové funkce činí cca 1.55% a jeví se jako nezávislý na velikosti vstupní množiny. Tento postup ovlivňuje zejména menší bloky, pro které zmenšení množství iterací může vyústit ke skokovému nárůstu váhové funkce o 15-30%. (viz tab. 7.5). Druhá metoda využívá stromové struktury KD-tree, Quad-tree a Oct-tree. Výsledky KD-tree jsou srovnatelné s předcházející metodou, v některých datových sadách jsou výpočetní časy mírně lepší, jindy mírně horší. Tuto část práce považuji za výborně zpracovanou.

Algoritmus clusterizace je plně paralelizován, testování probíhalo vzhledem k porovnání odhadů časové a paměťové složitosti se skutečnými výsledky na různých typech dat. Není mi jasné, jak byla testována varianta bez I/O operací. Časy pro I/O nebyly do výsledku započteny vůbec nebo I/O operace nebyly vůbec provedeny? Autor také implementoval clustrovací algoritmy pro GPU s využitím CUDA frameworku, tento přístup mu přinesl zvýšení výkonu o cca 20%.

Při zpracování clusterizace autor navrhuje speciální efektivnější postup uchování distanční matice (symetrická) s uchováním části prvků nad a pod hlavní diagonálou. Není mi jasné, jaké má toto řešení výhodu oproti uchování horní či dolní trojúhelníkové matice, které lze pro případné další operace považovat za vhodnější. Prosím o vysvětlení u obhajoby.

Poslední kapitola je věnována problematice 2D/3D Delauany triangulace hierarchických množin. Navržené řešení umožňuje provádět vzhledem k triangulaci dvojí operaci s clustrem, a to: přidání či odebrání clusteru. Pro tyto účely autor používá algoritmus inkrementálního vkládání a inkrementálního výběru. Algoritmus inkrementálního výběru se jeví jako výpočetně náročnější (cca. 10x), přesto mohou být tyto operace prováděny takřka v reálném čase. Práce s clustery je automatizována, clustery mimo výřez definovaný souřadnicemi mohou být "vyjmuty" z triangulace. Vyjmutí je doprovázeno vznikem díry, která je následně retriangulována. Posledním krok algoritmu představuje obnovení nekonvexity zpracovávané oblasti pro tento účel autor vyvinul vlastní algoritmus provádějící

postupné "ořezávání" konvexní obálky. Autor uvádí, že je tento postup použitelný pro homogenní množiny. Bylo by zajímavé otestovat, jak funguje pro množiny z různou prostorovou hustotou bodů, u kterých nemusí být dodrženy předpokládané tvary trojúhelníků (větší výskyt tvarově nevhodných trojúhelníků). Takové množiny však nejsou produktem laser scanningu, vznikají spíše terestrickým měřením

K autorovi mám ještě následující dva dotazy, o jejichž zodpovězení prosím při obhajobě:

- Jak budou clusterizační algoritmy fungovat pro nehomogenní množiny, tj. pro množiny s výrazně proměnlivou prostorovou hustotou prvků? Nebude při přechodu mezi jednotlivými úrovněmi docházet k ještě výraznějším disproporcím v plošné hustotě bodů?
- Použité parametry eliptické metrika byly konstantní pro každý objekt nebo jsou tyto parametry lokálně proměnlivé?

Disertační práce působí jako celek velmi dobře, zpracovává velmi podrobně značně rozsáhlou problematiku. Z literatury je patrné, že autor výsledky publikoval v pěti vědeckých článcích, což svědčí o kvalitě navrženého řešení a jeho akceptaci vědeckou komunitou.

Nejvýznamnější připomínka ze strany recenzenta se týká strukturace textu. Text disertační práce je oborově textem technickým, jeho formátování a styl však více připomíná popis. Kombinace dlouhých odstavců bez mezi titulků, popis funkčních vztahů namísto uvedení vzorce a popis algoritmů s absencí pseudokódu v nějakém formalizovaném jazyce působí nepřehledně a stěžují orientaci v textu. Některé algoritmické pasáže se stávají obtížněji pochopitelné, např. modifikace algoritmu k-means, aplikace eliptických metrik. Obecně bych doporučil používat více vzorců, obrázků a formalizovaných popisů pro zlepšení celkové čitelnosti.

Autor uvádí, že podobný systém doposud není k dispozici. Existuje však systém Grifinor vyvíjený v Dánsku, který pro hierarchické triangulace používá 2D teselace (Voronoi diagramy). K prvkům uvnitř buňky přistupuje podobně jako k prvkům sdruženým v clusteru...

Za velmi významný fakt dále považuji, že se autorovi podařilo dovést navržené algoritmy do implementačního stádia a vytvořit plně funkční aplikaci. U takto složitého softwaru, který včetně návrhu, vývoje, implementace, testování a optimalizace představuje práci pro skupinu programátorů, je realizace všech kroků jedním autorem obdivuhodná. Předpokládám, že výsledné dílo nalezne uplatnění na softwarovém trhu, autorovi doporučuji kontaktovat např. resort ČÚZK, kde o problematiku triangulace velkých množin projevují dlouhodobý zájem.

Předložená disertační práce obsahuje originální autorské myšlenky vedoucí k řešení zadaného problému, které se opírají o teoretické zdůvodnění. Práce ing. Jiřího Skály splňuje podmínky kladené na doktorskou disertační práci, představuje původní vědecké dílo s mezinárodním přínosem v oblasti teoretické i implementační. Autor prokázal schopnost tvořivého myšlení, systematického a logického přístupu, schopnost samostatného řešení vědeckého problému a výrazně posunul stav řešené problematiky. Připomínky obsahového i formálního charakteru nesnižují úroveň této práce. Disertační práci považuji celkově za velmi zdařilou a kvalitně zpracovanou.

Doporučuji, aby Ing. Jiřímu Skálovi **bylo umožněno** obhájit jeho doktorskou disertační práci.

V Praze dne 15. listopadu 2012

Ing. Tomáš Bayer, Ph.D.

Katedra aplikované geoinformatiky a kartografie

Přírodovědecká fakulta Univerzity Karlovy