

МЕТОДИ ЗА АНАЛИЗ НА ПОДОБИЕ И ОТКРИВАНЕ НА СХОДНИ ИЗОБРАЖЕНИЯ

О. Наков, М. Лазарова, И. Станков, Д. Гаджев,

i.stankov@abv.bg

Абстракт: Методите за анализ на подобие и откриване на сходни изображения са част от методите за извличане на мултимедийни данни или „Multimedia Data Mining”. Тези методи се използват при анализ на големи обеми от информация и често тяхното бързодействие има важна роля за работния процес. Методите, които са разгледани в този доклад, използват няколко подхода за анализ на изображения, отличават се един от друг по своето бързодействие и често водят до различни резултати. Разгледаните методи са имплементирани и са извършени експериментални изследвания за оценка, на параметри на производителност и качество на разпознаване за всеки от тях.

Ключови думи: Бързодействие; Графични форми; Извличане на мултимедийни данни; Топология на изображение; Цветова хистограма

METHODS OF ANALYSIS OF SIMILARITY AND DETECTION OF SIMILAR IMAGES

O. Nakov, M. Lazarova, I. Stankov, D. Gadzhev

i.stankov@abv.bg

Abstract: Methods of analysis of similarity and detection of similar images are part of the methods to retrieve multimedia data or "Multimedia Data Mining". These methods are used in the analysis of large volumes of information and often their performance has an important role in workflow. The methods that are discussed in this report, used several approaches to image analysis. They are distinguished from one another by their performance and often lead to different results. The considered methods are implemented and carried out experimental research evaluation of parameters of performance and quality of recognition for each of them.

Keywords: Color histogram; Graphic forms; Image topology; Multimedia Data Mining; Performance

1. СЪЩЕСТВУВАЩИ МЕТОДИ ЗА АНАЛИЗ НА ПОДОБИЕ И ОТКРИВАНЕ НА СХОДНИ ИЗОБРАЖЕНИЯ.

Съществуват различни методи за анализ на изображения, които се използват при анализа на подобие и откриване на сходство между изображения. Тези методи се използват в извличането и обработката на информация от мултимедийни данни в Multimedia Data Mining [1] областта. Методите за определяне на подобие, които намират приложение за анализ на подобие между изображения, могат да бъдат разделени на две групи според характера на използваните данни за оценка и определяне на сходство.

Първата група методи използват допълнителна информация, като тя най-често включва ключови думи от името или описанието на файловете, размери на изображенията, час и дата на тяхното създаване. Методите от тази група не са насочени пряко към анализа на изображения и могат да бъдат използвани при анализ на други видове файлове, поради което те не са основна цел на този доклад.

Втората група от методи е тази, към която е насочен текущият доклад. Методите от тази група използват данни, които са извлечени от анализирани изображения. Най-често данните се представят, като цветови хистограми, наличие на текстури, топология на изображението и характеристики на изобразяваните обекти.

1.1 Етапи на анализ и откриване на подобия и сходства в изображения.

Независимо от това, към коя група са насочени методите за анализ и какъв вид информация използват, за да открият подобия или сходства между изображения, съществува едно основно правило за тяхната работа. Според правилото тези методи се прилагат на два основни етапа.

В първия етап методите анализират подадените изображения или файлове, за

да получат необходимата за работата им информация, като ключови думи, размери, час и дата на създаване, цветови хистограми, наличие на текстури, топология на изображението, характеристики на изобразяваните обекти или други видове информация, в съответствие със спецификата и изискванията на всеки от методите. В резултат от анализа, извършен в първия етап и извлечената информация, за всяко изображение се съставя множество от характеристики, което го представя и което съхранява по-малък размер данни в сравнение с оригиналното изображение.

Във втория етап на работа се извършва сравнение на характеристиките, получени за всяко от изображенията от първия етап. По този начин се откриват подобия или сходства между изображения, като те се дължат на подобия в информацията получена от анализа и съхранена в множествата от данни за всяко изображение.

Благодарение на използването на множества от характеристики, които представят анализирани изображения и които имат по-малки размери, се ускорява изпълнението на втория етап от работата на методите. Изпълнението на първия етап е тази част от работата, която се извършва по-бавно. Тъй като първият и вторият етап могат да бъдат извършвани самостоятелно, често анализа на изображения се прави предварително и получените множества от данни представящи изображенията се съхраняват. По този начин съхранените характеристики могат да бъдат използвани в произволен момент, за да бъде потърсено подобие или сходство между предварително анализирани и ново постъпили изображения.

1.2 Основни проблеми на разглежданите методи.

Методите, които са разгледани в този доклад, използват няколко подхода за анализ на изображения, отличават се един от друг по своето бързодействие и често получените от тяхната работа резултати се различават. Резултатите се

различават не само при използването на различни методи, но и при разлика в настройването на метода, чрез избор на конкретни стойности на използваните параметри. От това следва, че едни от основните проблеми на тези методи е това, как те са настроени за работа, от което зависи и качеството на получените резултати. Съответно един от основните аспекти, които ще бъдат разгледани в този доклад, са различните характеристики на методите, това как те могат да бъдат настройвани и как промяната на работните им параметри се отразява на работата на методите.

Тъй като методите, които се разглеждат често анализират голям брой изображения, които обикновено са съхранявани във файлове с не малки размери, то използваните методи анализират голямо количество информация. От това следва още един проблем на разглежданите методи, а именно тяхното бързодействие. Бързодействието при анализа на изображения зависи от бързодействието на използваните методи и от големината на анализираните изображения. Бързодействието на методите от своя страна, зависи от броя операции, които трябва да бъдат направени с едно изображение, за да бъде то обработено. Повишаването или намаляването на броя операции използвани от един метод, могат да повлияят на резултатите получени от анализа. Големината на анализираните изображения също може да бъде променяна, с което се цели ускоряване на анализа, но това може да повлияе на получените крайни резултати.

Както отбелязахме до момента, основните проблеми на методите за анализ на изображения са тяхното бързодействие, качеството на получените от работата им резултати и начините, по които те са настроени да работят. Анализа на тези три основни проблема на методите е целта на този доклад. За да улесним представянето на направения анализ, ще разделим разглежданите методи на две групи. Първата група методи ще наречем прости методи за

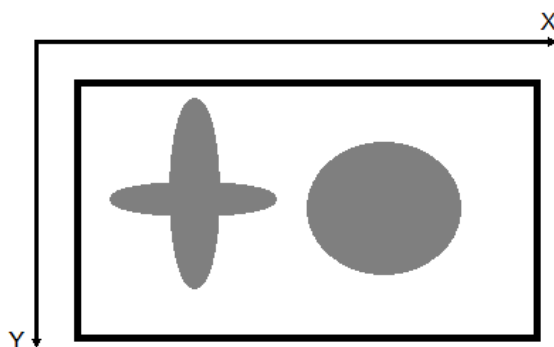
анализ, а втората ще наречем сложни методи за анализ. В първата група ще попаднат тези методи, които имат по-високо бързодействие заради своята същност на изпълнение, а във втората група ще попаднат съответно тези методи, които имат по-ниско бързодействие.

2. ПРОСТИ МЕТОДИ ЗА АНАЛИЗ НА ПОДОБИЕ И ОТКРИВАНЕ НА СХОДНИ ИЗОБРАЖЕНИЯ.

В групата на простите методи за анализа на изображения попадат методи, които изграждат цветни хистограми или цветни топологии на подадените им изображения. Те принадлежат към тази група, заради високото бързодействие, което имат. Високото бързодействие в работата на методите се поражда от факта, че те обработват данните от подаденото им изображение само няколко пъти. За разлика от методите с ниско бързодействие, които се нуждаят от много операции, за да извършат своя анализ, методите с високо бързодействие се нуждаят само от единици до десетки операции за своя анализ. По-долу са представени двата метода с високо бързодействие, които бяха изследвани.

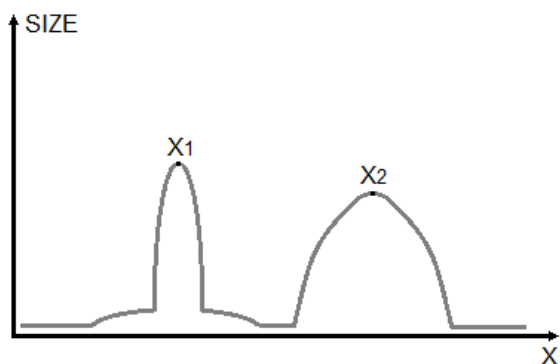
2.1 Метод за съставяне на цветни хистограми.

Методът за съставяне на цветни хистограми анализира цветовете и тяхното разположение в подаденото му изображение. В резултат на анализа, за всеки от търсените цветове се правят хистограми. Хистограмите на цветовете се правят и за двете измерения на подадените изображения, както това е показано на фигура 1.

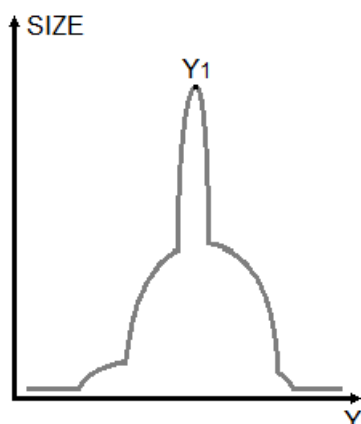


Фиг. 1. Примерно изображение подлежащо на анализ.

Примерни хистограми от анализа на изображението, показано на фигура 1, са представени на фигура 2 и фигура 3. В някои случаи е възможно анализа да се усложни, като освен хистограми за двете измерения на подаваните изображения, се правят и хистограми за разпределението на цветовете в отделни под области на изображенията.



Фиг. 2. Хистограма на цветовете по оста X за примерното изображение.



Фиг. 3. Хистограма на цветовете по оста Y за примерното изображение.

Цветовите хистограми могат да бъдат направени за различни цветове. Най-често тези хистограми се правят за трите основни цвята, с които се представят цифрови изображения, а именно: червен, зелен и син. Възможно е използването на други цветове в случаите, когато те имат важно значение за анализа.

Важен момент в изграждането на цветовите хистограми е избирането на начин или метод, по който да се отчита всеки от цветовете. Когато има ясно изразени червени, зелени и сини пиксели в едно изображение, те лесно могат да бъдат отчетени. Но когато има пиксели с цветове, получени от смесване на основните, е необходимо да бъде

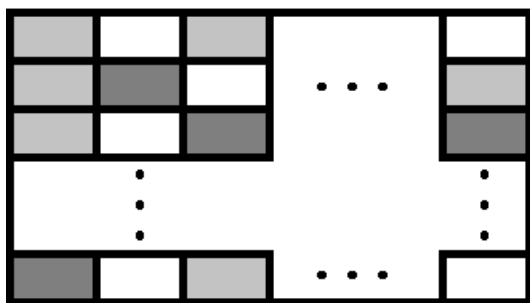
направена оценка, от какви цветове са изградени те и как техните стойности да повлияят на данните в хистограмите на цветовете. Пример за това може да бъде жълт пиксел, цвета на който е получен от смесването на основните цветове червено и зелено. Различни подходи при изграждането на цветови хистограми биха довели до различни варианти за отчитане на пиксела. Пикселът може да не бъде взет под внимание, може да бъде взет под внимание, като се отчете само основният цвят с най-висока стойност или може да се вземе под внимание за всички хистограми, като се отчете и стойността на всеки от основните цветове в пиксела. При отчитането на цветовете за всеки пиксел е възможно и поставянето на прагови стойности. Праговите стойности определят дали цвета на един пиксел отговаря на определени условия, за да се приеме, че този пиксел ще бъде отразен в хистограмите на цветовете.

След като бъдат установени цветовете, които биха били подходящи за анализ, начина и праговите стойности, за отчитане цветовете на пикселите, се съставят цветовите хистограми. Хистограмите се съставят както за всяка от осите на изображението, така и за всеки от основните цветове. След като хистограмите са изготвени, те се анализират допълнително и получените данни се явяват крайни данни за анализа на изображението според метода на цветовете хистограми. Допълнителният анализ, който се прави, отчита максимумите в графиките за всеки цвят. За всеки максимум се взима под внимание неговото разположение в изображението и стойността на максимума. От всички максимуми се избират тези, които имат най-големи стойности, и се установява къде в изображението се намират те и следователно къде се натрупва определен цвят. На фигура 2 е показана примерна хистограма с наличие на повече от един максимум. Важно е да са отбележи, че могат да бъдат използвани толкова максимуми, колкото са необходими, за да се подобри работата на анализа.

Един от особените проблеми при установяването на максимуми е да се прецени, кои от тях са важни и кои по-скоро ще попречат на анализа и не трябва да бъдат взети под внимание. Също така важни проблеми са и неправилното отчитане на максимуми, установяването на платовидни елементи в хистограмите и определянето на броя максимуми, които да се вземат под внимание.

2.2 Метод за съставяне на цвeтова топология.

Методът за съставяне на цвeтови топологии, подобно на метода за създаване на цвeтови хистограми, анализира цвeтовете и тяхното разположение в подаденото му изображение. За разлика от метода за създаване на цвeтови хистограми, в резултат от работата на метода за създаване на цвeтови топологии не се изготвят хистограми, а се получава информация за това, какви цвeтове преобладават в различните части на подаденото за анализ изображение. За да се анализира изображението, то се разделя на под области, които могат да обхващат различни по размер и разположение части на изображението. След като се установи начина, по който ще се формират под областите, за всяка от тях се прави анализ на цвeтовете разположени в нея. Пример за това, как изображение може да бъде разделено на под области, е показан на фигура 4.



Фиг. 4. Разделение на изображение на под области.

Анализът на цвeтовете в под областите, също както анализа на цвeтовете при изграждането на цвeтови хистограми, може да става по различни начини. Два примерни варианта са използването на преобладаващи цвeтове за всяка под

област и отчитането на стойности за всички цвeтове, за всяка под област.

Отчитането на всички цвeтове в една под област дава по-ясна информация за наличието на цвeтове в областите. По този начин се позволява използване на допълнителни методи за анализ на получената информация, за да се установи сходство между изображения. Пример за подобен допълнителен анализ е възможността за използване на клъстериране на получените данни от множество изображения. Проблем при отчитането на стойностите, за всички цвeтове в под областите, се получава, когато не се използват допълнителни методи за анализ, а се прилага директно сравнение между стойностите на данните за цвeтовете в под областите. Проблемът се явява в това, че стойностите за един цвят в изображения, които се сравняват, могат да бъдат много ниски или нулеви, при което сравнимите изображения ще имат подобие. Но откритото подобие, може да е погрешно, защото то е направено на базата на липсата на цвeтове, а не на базата на съществуващите в изображението цвeтове. Пример за това могат да бъдат зелено и червено изображение между които е открито подобие, заради липсата на син цвят.

Приемането на преобладаващи цвeтове, за всяка от формираните под области, е лесен метод. Данните за преобладаващите цвeтове, получени от прилагането на метода, могат директно да бъдат използвани за сравнение с данните, получени за други изображения. Предимството на този метод е възможността, лесно да се избегне подобие между изображения, което е породено от липсата на цвeтове. Това става, като се отчита наличие на подобие само между цвeтовете, които са приети за преобладаващи за конкретната под област. Отчитането само на преобладаващи цвeтове, може да се яви и като проблем, защото се губи информацията за наличието на другите цвeтове. В резултат на това събраната информация след анализа на множество

изображения не винаги е подходяща за клъстериране.

Двете разгледани по-горе разновидности на метода за съставяне на цетова топология, могат да бъдат комбинирани, както е възможно и тяхното изменение в зависимост от това, за какво ще се използват получените от тях резултати. Възможно е поставяне на прагови коефициенти за отчитане на определени цетове и разпознаването на различни цетове, а не само основните. С всички тези промени, провежданият анализ може да бъде конфигуриран по такъв начин, че да подобри качеството си.

3. СЛОЖНИ МЕТОДИ ЗА АНАЛИЗ НА ПОДОБИЕ И ОТКРИВАНЕ НА СХОДНИ ИЗОБРАЖЕНИЯ.

В групата на сложните методи за анализа на изображения попадат методи, които откриват наличие на текстури или разпознават обекти от подадените им изображения. Тези методи принадлежат към тази група, заради ниското си бързодействие. Ниското бързодействие в работата на методите се поражда от факта, че те обработват данните от подаденото им изображение многократно, за да изпълнят сложните си анализи. Методите с ниско бързодействие често се нуждаят от стотици итерации до стотици хиляди итерации. Това забавя многократно тяхното изпълнение в сравнение с простите методи, които разгледахме.

Методът с ниско бързодействие, който е изследван, е предназначен за намиране на обекти или форми в изображения. Изпълнението на този метод изисква множество обработки над изображенията, които са подадени за анализ. Освен анализа на цетове, който споменахме и при простите методи за анализ на подобие, за откриването на обекти е необходимо използване на методи за откриване на контури, откриване на форми и анализ на откритите форми. В направеното изследване са използвани различни методи, които са представени по-долу.

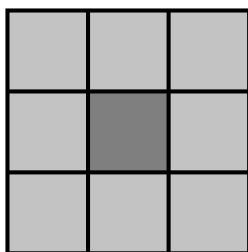
3.1 Методи за откриване на контури.

Методите за откриване на контури се използват, за да разпознаят области в изображения, които са гранични между различни заснемани обекти. Съществуват различни методи за откриване на контури [2]. Някои от методите позволяват намирането на скрити контури, възстановяването на липсващи контури или определянето на дълбочина в изображението и установяване на това, как са подредени контурите в реалното триизмерно пространство. Тъй като тези методи са сложни и не са тясно свързани с темата на този доклад, те няма да бъдат представени по-подробно. За целите на текущият анализ са използвани методи за откриване на контури, които се ограничават до това, да намерят гранични области между обекти в анализираните изображения.

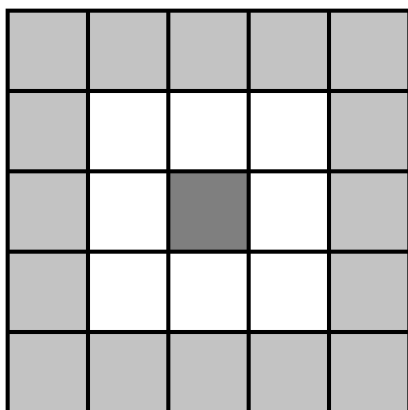
Един от най-простите начини за намиране на гранични области между обекти в изображение, е анализа на цетове на съседни пиксели. Когато разликата в цетата на съседни пиксели е голяма, може да бъде предположено, че тези пиксели принадлежат на различни обекти или форми. В някои случаи е необходимо, да бъдат проверени цетове на пиксели, които не са преки съседи, а са отдалечени на малко разстояние един от друг. По този начин могат да бъдат установени гранични области между форми, дори когато граничните пиксели имат цетове, преливащи се от околните форми. Подобни методи са подходящи за използване при анализ на изображения с влошено качество. Пример за подобни изображения са тези, съхранени във формат JPEG, които се създават с компресия, водеща до загуба на качество.

Начинът, по който ще се сравняват цетове между пиксели, зависи от спецификите на анализа и може да бъде променян за да се достигне до по-качествени резултати. Пикселите участващи в сравнението могат да са преки съседи, както това е показано на фигура 5, да бъдат непреки съседи, както това е показано на фигура 6, или да бъдат

разположени по други начини, които са подходящи за провежданите анализи.



Фиг. 5. Анализ на пиксели, разположени в пряко съседство.



Фиг. 6. Анализ на пиксели, разположени в непряко съседство.

Методите за откриване на контури, които разгледахме до момента се използват за да идентифицират пиксели, които са част от контур. Друг вид методи за определяне на контури са методите, за верифициране на контури. Те се прилагат над вече намерените пиксели, които са отчетени като части от контури. Методите за верифициране целят да премахнат пикселите, които вероятно не са контури, и да добавят пиксели, които трябва да принадлежат на контурите, но не са разпознати като такива до момента.

В резултат от направените изследвания, с различни методи за откриване на пиксели принадлежащи на контури и тяхното верифициране, може да се каже, че изследваните методи могат да бъдат оптимизирани и настроени. С помощта на оптимизация на методите, може да се повиши тяхното бързодействие, което, както отбелязахме до момента, е един от основните проблеми на тези методи. Възможностите за настройване на работата на методите за откриване на контури, могат да подобрят качеството на анализите, което е другият основен

проблем на методите, които разглеждаме. Примери за възможни оптимизации и настройки на методите за откриване на контури са разгледани по-долу.

Оптимизирайки методите за откриване на контури, може да се намали броя итерации, които се извършват за да се анализират пикселите в едно изображение. Пример за подобна оптимизация е възможността, пикселите да не се сравняват с всичките си съседи. Ако един пиксел се сравни с левият и десният си съсед, когато е подаден на методите за анализ, то в процеса на анализ, той ще бъде сравнен още веднъж с тях. Това се дължи на факта, че той ще бъде използван за сравнение, когато неговите съседи бъдат подадени на методите за анализ. Възможна оптимизация в този случай е възможността да не се прави сравнение спрямо дясно разположените пиксели, а на подобен принцип може да се пропусне сравнение с половината съседни пиксели. Тази оптимизация би довела до премахване на половината от изпълняваните сравнения в работата на един метод.

Настройване на методите за откриване на контури, може да се приложи при анализа на цветове на съседни пиксели. Подобни настройки могат да подобрят работата на методите и да доведат до по-точно разпознаване на това, дали съседни пиксела принадлежат на една или на различни форми от анализираното изображение.

3.2 Методи за откриване на форми.

Методите за откриване на форми са една от най-големите причини за ниското бързодействие на сложните методи. Тези методи могат да бъдат изпълнявани по два основни начина, а именно чрез рекурсивни или чрез итеративни алгоритми. При изпълнението си с рекурсивни алгоритми, методите работят по-бързо в сравнение с изпълнението си с итеративни алгоритми.

Изпълнението с рекурсивни алгоритми има проблеми при работа с големи

изображения. Това се дължи на факта, че формите, разположени в анализираните изображения, често са формирани от няколко милиона пиксела, което води до няколко милиона рекурсивни извиквания на използваните методи и до възможността за претоварване на използваните системи.

Итеративното изпълнение се използва за да се избегнат проблемите, породени при рекурсивно изпълнение. В резултат на итеративното изпълнение на методите за откриване на форми, често се достига до необходимостта за многократно обработване на едно изображение. Резултати от направените изследвания показват, че за изображения съставени от няколко милиона пиксела, често е необходимо да се правят от стотици обработки до стотици хиляди обработки.

Необходимостта от много обработки при итеративното изпълнение на методите за откриване на изображения ги забавя. За да се подобри производителността на тези методи се препоръчва те да бъдат комбинирани с частично рекурсивно изпълнение и да бъдат оптимизирани по различни начини. Важно е да се отбележи, че възможностите за комбинирано итеративно и рекурсивно изпълнение и оптимизация на използваните методи трябва да бъдат използвани по начини, подходящи за конкретният анализ, ако това е възможно. По този начин производителността на методите за откриване на форми може да се повиши.

3.3 Методи за анализ на форми.

Методите за анализ на форми се прилагат, след методите за откриване на форми и техните контури. Анализът на формите подготвя данни за размера, формата, ориентацията, цвета или цветовете на формата, нейното разположение в изображението и други нейни характеристики. Данните, които са събрани от този анализ за една форма, служат за сравнение с данните събрани за други форми. Благодарение на подобие в събраните данни за формите в

анализираните изображения може да се открие подобие между изображения.

4. КОМБИНИРАНЕ НА МЕТОДИ ЗА АНАЛИЗ НА ПОДОБИЕ И ПРЕДСТАВЯНЕ НА ПОЛУЧЕНИТЕ ОТ ТЯХ РЕЗУЛТАТИ.

След приключване на анализа на изображения, получените за тях данни се сравняват. Често данните получени от различните методи, които са представени до момента, намират подобие между различни изображения. За да се комбинират данните от различните методи за анализ, може да бъде направено сечение между тях или да се процедира по други начини. Забелязва се, че, при неправилно използване на данни от повече от един метод, възниква възможността за влошаване на крайните резултати от анализа. Пример за това са случаите, когато изображения с малки нива на подобие, получени от различните методи, имат сумарно по-високи нива на подобие от изображения с високи нива на подобие, получени за един от методите.

5. ИЗВОДИ

Често различните методи, които представихме до момента, намират подобие между различни изображения. Това се получава заради различните подходи, които те използват. В различни случаи, някои методи се справят по-добре или по-лошо от други. Използването на подходящите методи в подходяща ситуация, тяхната правилна конфигурация и комбинирането на резултати от работата на различните методи, може да доведе до по-качествен анализ. Важно при използването на тези и подобни методи е да се познават техните специфики и да се предвиди, кои от тях биха били по-подходящи за конкретните поставени цели.

ЛИТЕРАТУРА

- [1] Jiawei Han and Micheline Kamber, *Data Mining Concepts and Tehnique*, Morgan Kaufmann and Elsevier, San Francisco, ISBN 13: 978-1-55860-901-3
- [2] Szeliski R., *Computer Vision: Algorithms and Applications*, Springer, 2008