СЪВРЕМЕННИ МЕТОДИ ЗА РАЗПОЗНАВАНЕ НА ЧОВЕШКА ПОЗА ОТ ИЗОБРАЖЕНИЯ С ДЪЛБОЧИННА ИНФОРМАЦИЯ

Александър Маринов

STATE OF THE ART IN HUMAN POSE RECOGNITION FROM DEPTH IMAGES

Alexander Marinov

***Abstract**: The recent immersion of depth/range cameras triggered new wave of research in the area of human pose recognition. Depth images provide grounds for a new feature space that can be used to build strong and robust classifiers. The paper presents a short overview on state of the art approaches to human pose estimation from depth images. The main focus is over Jamie Shotton's approach based on per pixel classification into a set of body parts.*

***Key words**: depth images, Image Recognition, body tracking, segmentation*

1. Въведение

Проследяването на състоянието на човешкото тяло, неговата поза и движение в пространството е приложимо в сферите на роботиката, човеко-машинните интерфейси, компютърни игри, системи за сигурност, филмовата индустрия и др. Едни от основните цели при изграждането на система за определяне на човешката поза са работа в реално време и висока стабилност по отношение на вариациите в облеклото и формата на тялото, многообразието от възможни движения, и характеристиките на средата. Задачата е опростена значително с появата на камери с дълбочинна информация за сцената, но комплексността на движенията и необходимите изчисления за работа в реално време остават основно предизвикателство.

Изображенията с дълбочинна информация добавят ново измерение както към RGB пространството така и към подходите в компютърното зрение. Докладът прави обзор на няколко метода за идентификация и проследяване на човешки части във времето с най-голям акцент върху работата на Jamie Shotton, която е съществен компонент в Kinect гейм платформата.

2. Съвременни методи

Появата на камери с дълбочинна информация за сцената води до изследването на нови методи за проследяване и оценка на човешката поза, които неминуемо заимстват техники от визуално базираните подходи.

2.1 Използване на виртуален 3D модел

Grest et al. (4) използват Iterative Closest Point (ICP) алгоритъма за проследяване на скелет с известни размери и начална позиция. Авторите създават плътни мрежови модели за всички сегменти на тялото и се опират на MPEG4 стандарта, който дефинира до 180 степени на свобода на движение (DOF). Така създадения MPEG4 модел на тялото е комбинация от кинематични вериги: движение на точка от тялото може да се представи като последователност от ротации около съответни оси. С други думи, позицията на точка от даден сегмент на тялото може да се изчисли за дадени ставни ъгли и начална позиция на тялото. При известни съответствия на начални точки X^0 от мрежовия модел и наблюдаваните точки Y от дълбочинното представяне на тялото, задачата за оценяване човешката поза се състои в намирането на тези параметри на трансформацията (ставните ъгли), които проектират X върху Y. Авторите предлагат нелинейни оптимизационни техники, които използват аналитично изведен Якобиан (Jacobian matrix). За изчисляване на съответствието между 3D множествата от точки от наблюдаваното дълбочинно изображение и MPEG4 модела, се използва оптимизиран алгоритъм за итеративно търсене на най-близките съседи чрез ICP.

За изследване точността и ефективността на своя подход, авторите правят следните 2 теста с видео клипове върху 3Ghz Pentium 4:

- движения в горната част на тялото с 14 DOF. Най-бързи и точни резултати са постигнати при 800 точки за фрейм – средно 200ms за кадър.
- синтетични движения включващи цялото тяло (ръце, крака, глава) с 26 DOF. При 1000 точки за фрейм, времето за обработка на кадър се увеличава до 250ms.

Предложеният подход работи в близо до реално време – 4 fps (кадри в секунда). Едни от недостатъците му са преинициализиращите стъпки: офлайн напасване на мрежовия модел върху дълбочинното изображение на човека със скалиране на всеки съответен компонент от тялото; задаване на начална /референтна/ поза. Точното напасване на модела дава отражение върху точността на разпознаване на позата. Фиг.1 дава кратка представа за работата на подхода.

Фиг. 1. Дълбочинно изображение (в ляво), оригиналното RGB изображение с наложен изчисления модел на позата (в средата), 3D модела погледнат от страни.

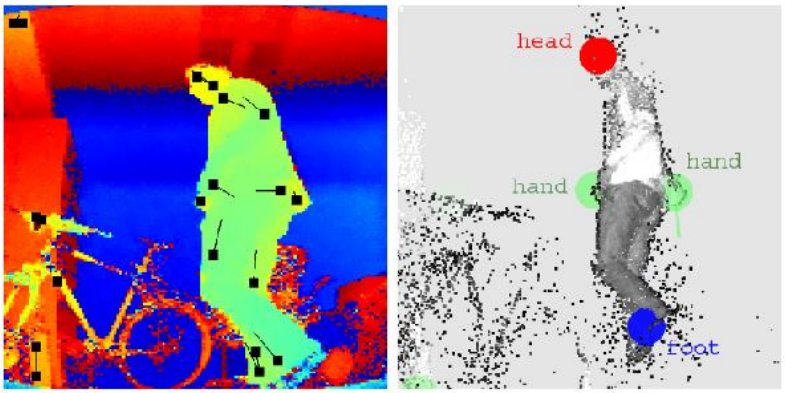
Горният подход се използва в (5), където авторите добавят и цветова информация – напасват силуета на тялото от цветното изображение към 2D силуета на мрежовия модел. Недостатъците в избора на този подход са сходни с тези от (4). Постигнатата ефективност при 10 DOF е 5 fps за мрежови модел от 10 000 точки. По тежкия и плътен модел от 90 000 точки се нуждае от 1.5s за обработката на 1 кадър.

2.2 Идентификация чрез специални геодезични точки (interest points)

Plagemann et al. (6) работят по идентификация и локализация на части от човешко тяло (глава, длани и стъпала) в дълбочинни изображения чрез анализ и класификация на части от 3D мрежи. Първоначалната покриваща дълбочинното изображение мрежа бива разкъсвана по места, където разстоянието между точките превишава праг $\delta connected$. В така образуваните под-мрежи се търсят специални геодезични точки (interest points), такива които съответстват на едни и същи точки от тялото дори след различни промени в стойката или жестикулации. Идеята на алгоритъма за откриване на множество от такива точки се състои в последователно прибавяне на геодезично най-отдалечената от текущо натрупаните точки като се започва от точката, намираща се в геодезичния център. За оценка за ориентация на точката се приема вектора, започващ на разстояние $\delta orient$, тръгвайки по обратния най-къс път в посока на първоначалната. Специалните точки се приемат за хипотези на принадлежност към някоя от търсените части. Самото класифициране се извършва с предварително обучен Boost класификатор. В класификатора като признаци за класифициране се разглеждат прозорци от дълбочинното изображение, центрирани в специалните точки и ориентирани спрямо намерената оценка за ориентация на точката.

Сериозен недостатък на този подход е че специалните точки лимитират набора от части на човешкото тяло, които могат да бъдат локализирани до тези,

които се намират на екстремални геодезични позиции– глава, краища на ръце и крака. Зависи от настройка на параметрите $\delta connected$ и $\delta orient$. Не се различава ляв от десен крайник. Предимство е че алгоритъма дава оценка за ориентацията на намерените части, работи в реално време и представя по-добри резултати от алтернативни подходи с плъзгащи прозорци поради по целесъобразното им позициониране в изследваните специални геодезични точки. Фиг. 2 демонстрира пример от резултата на работата на алгоритъма, където е изобразен облак от намерените специални точки и идентифицираните части от тялото.



Фиг. 2. Цветно кодирано дълбочинно изображение с откритите специални точки (в ляво). 3D облак от точки на същото изображение с откритите части на тялото.

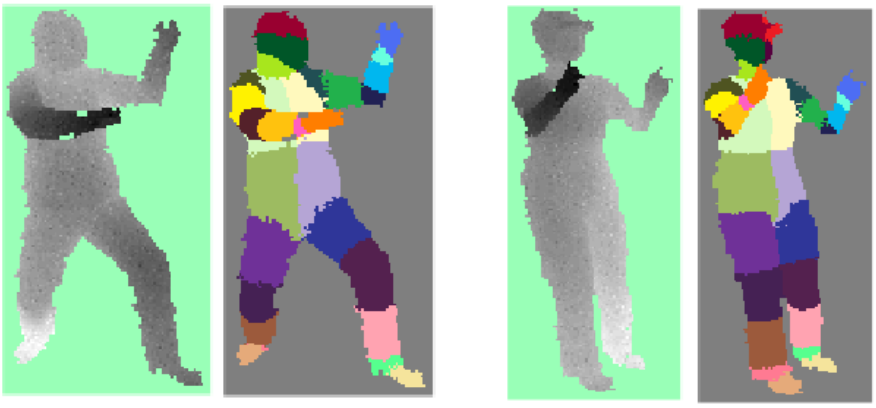
3. Разпознаване на човешки части с по-пикселово класифициране

За разпознаване на човешката поза, Shotton et al. (1) предлагат метод за бързо и точно намиране на вероятните 3D позиции на скелетни стави от уникално дълбочинно изображение. Използвайки подход за разпознаване на обекти, те представят тялото като комбинация от части, всяка от които е асоциирана с уникален етикет и локирана пространствено около конкретна става (Фиг. 3, 4). Задачата се свежда до класифициране на отделен пиксел. За класификатор се използва рандомизирана гора от класификационни дървета, която е построена над огромна база от многообразни синтезирани изображения с модели на човешкото тяло като по този начин се избягва пренагаждане към данните. Чрез пре-проектиране на класификационните резултати с mean shift (2) алгоритъма се генерират 3D предложения за няколко скелетни стави. Този подход, фокусиран над уникален фрейм за инициализация и възстановка, може да се използва в допълнение към съществуващи алгоритми за проследяване на движението, където да се включи времева и кинематична последователност. Освен това, той е достатъчно обобщаващ и сполучливо класифицира нови (нетренирани) пози. Следващите секции разглеждат в детайли предложения подход.

3.1 Тренировъчни данни

Тъй като човешкото тяло е в състояние да заеме голям брой от пози, чието разнообразие трудно би могло да се симулира, екипа на Shotton заснема няколко стотин видео клипа с дълбочинна информация на човек в действие: при шофиране, танцуване, тичане, управление на меню, и др. Тъй като предложения от тях подход се интересува от статични изображения, те отсяват тези пози от последователните фрейми, които са максимално различни върху дадено множество от стави – използва се клъстеризиране по най-далечния съсед (3). Чрез методи от компютърната графика създават 3D модели на човешкото тяло, които максимално добре да имитира човешките пози от реалния свят. С тези модели се синтезират дълбочинни и по части етикирани изображения при различни параметри: форма и размер на тялото, облекло и прическа, позиция на камерата и шум от камерата. Този подход позволява научаването на инвариантност по отношение на горните параметри.

Един от приносите на предложения подход за разпознаване на човешка поза е избора на телесни части, които изцяло да покрият тялото. Някои части са подбрани така, че директно да локализират важна скелетна става, докато други или се използват в комбинация за намиране на стави или просто запълват непокритите пространства. Всяка телесна част е асоциирана с уникален етикет, което позволява главната задача за разпознаване на позата да се сведе до ефикасен класифициращ алгоритъм. Така дефинирани частите се използват като текстури, които покриват съответните зони от 3D модела на тялото. Фиг. 3 представя няколко синтезирани дълбочинни изображения и съответното им представяне по части.



Фиг. 3. Синтезирани дълбочинни изображения и тяхните разделения по части.

3.2 Признаци върху дълбочинна информация

За постигане на 3D транслационна инвариантност, авторите предлагат използването на прости дискриминативни признаци чрез сравняване на дълбочинната характеристика на набор от пиксели. За всеки пиксел от тялото се изчисляват признаци от типа

(1) $$f_\theta(I,x) = d_I\left(\mathrm{x} + \frac{\mathrm{u}}{d_I(\mathrm{x})}\right) - d_I\left(\mathrm{x} + \frac{\mathrm{v}}{d_I(\mathrm{x})}\right),$$

където $d_I(x)$ е дълбочината на пиксел x от изображение I, а параметъра $\theta = (u, v)$ описва двойка офсети в сценичното пространство, а нормализацията на офсетите осигурява дълбочинна инвариантност. Сам за себе си всеки от признаците е слабо информативен относно принадлежността на даден пиксел към част от тялото. Но комбинирането им в гора от класификационни дървета е достатъчно за разграничаването на отделните части.

Голямо предимство в избора на тези признаци е тяхната изчислителна ефективност, а едновременното им калкулиране може да бъде възложено на многоядрен графичен процесор.

3.3 Рандомизирани класификационни гори

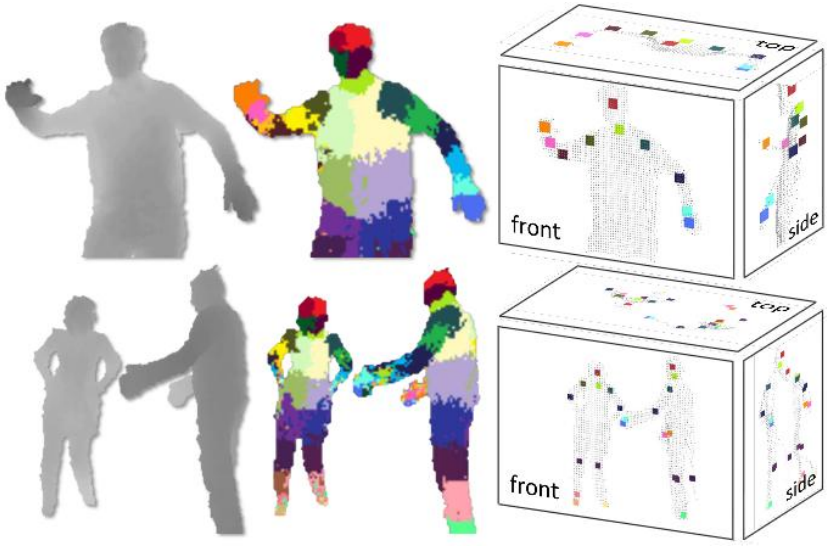
Рандомизираните класификационни гори са мощен мулти-клас класификатор. Гората се състои от класификационни дървета, чиито възли държат по един f_θ признак и разделящ праг τ. Листата съдържат съответното научено вероятностно разпределение $P_t(c|I,x)$ над етикетите на частите на тялото c. Получените разпределения за даден пиксел се усредняват върху резултатите от всички дървета, т.е. върху цялата гора, за да се получи крайното класифициране

(2) $$P(c|I,x) = \frac{1}{T}\sum_{t=1}^{T} P_t(c|I,x).$$

Рандомизиращия елемент се състои в генерирането на класификационните дървета. Всяко дърво се тренира върху произволно избрано множество от синтезирани изображения, всяко от които е представено от 2000 произволно избрани пиксела $Q = \{(I,x)\}$. За всяко дърво се генерира множество кандидат признаци представени от параметрите θ и τ. За всеки такъв признак множеството пиксели Q се разделя на ляво и дясно под-множества в зависимост от стойността на уравнение (1) в отношение с прага τ. За възел на дървото се взема този признак, при който информационната печалба от разделянето на пикселите е максимална. Генерирането на дървото продължава рекурсивно до стигането на някаква заложена максимална дълбочина, или до достигането на недостатъчно добра информационна печалба. Самата информационна печалба се изчислява на база на ентропията на Shannon над нормализираната хистограма върху етикетите на частите на тялото.

3.4 Скелетни стави

След класифицирането на всички пиксели от входното изображение се пристъпва към последния етап от предложения подход за разпознаване на поза на тялото – намиране 3D позициите на скелетните стави. Shotton et al. предлагат използването на mean-shift алгоритъма със специална функция за оценка на плътността, базирана на Гаусово ядро с тегла – вероятностната оценка от класификатора (1). Теглата в Гаусовото ядро са дефинирани така, че да осигуряват инвариантност по дълбочина. Mean-shift се използва за ефективно намиране на няколко локални притегателни центрове над плътността (атрактори), а сумата от теглата на притеглените точки към даден атрактор дава ниво на доверие да е център на става. Фиг.4 демонстрира резултати от по-пикселовото класифициране на входни дълбочинни изображения и предполагаемата 3D позиция на ставите.



Фиг. 4. Уникално дълбочинно изображение и резултат от по-пикселово класифициране. Цветовете отговарят на най-вероятния етикет на частта от тялото. Предложенията за 3D позициите на ставите са представени в последната колона.

3.5 Резултати

За оценка на ефективността на предложения подход се правят тестове както със синтезирани изображения (5000), така и с реални изображения (8808 от 15 различни телосложения), които са етикирани от човек. Разглежда се максимална ротация на тялото от ± 120°. Тестовата извадка съдържа плътните райони на частите на тялото и 7 скелетни стави за горната част на тялото. Авторите измерват средната точност на класифициране за всеки клас (част на тялото), както и средната точност в предполагаемите позиции на ставите. Резултатите показват

високо сходство в поведението на класификатора при синтезирани и реални изображения.

За анализ на качеството на предложения класификатор се изследва и влиянието на няколко трениращи параметъра – брой дървета, дълбочина на дърветата, брой тренировъчни изображения, максимален офсет θ за дълбочинните признаци от уравнение (1). Оптимални резултати, които намират компромисно решение по отношение на качество на класифициране, изчислително време и необходима работна памет са: 3 дървета с дълбочина 20, обучени над 300 000 изображения при 2000 кандидат признаци с 50 прага за всеки.

Алгоритъмът работи с приблизителна скорост от 50 fps върху модерен 8-ядрен процесор, като постига усреднена точност около 0.731 mAP (mean average precision) върху всички стави. Оптимизирана версия с изчисления върху GPU работи с ~200 fps върху конзолата Xbox 360, за която е измерен 0.677 mAP. Авторите използват тестовата извадка от (6) и установяват, че точността при намирането на предполагаемите позиции на ставите е значително по-добра от тази с подхода на Plagemann (6), както и че е поне 10 пъти по-бърз.

Подходът на Shotton се справя добре и генерира съответни ставни позиции при кадри с няколко човека. Това е доказателство, че по пикселовият класификатор обобщава добре различни сценарии, без да е изрично трениран за тях. Резултатите са значително добри и при пози, които не са участвали в обучението. Експеримент с 360° ротация показва засилена несигурност за ляво и дясно. Този проблем обаче може да се реши като класификационните резултати се предадат на проследяващ алгоритъм с множество хипотези.

4. Заключение

В статията разгледахме няколко подхода за разпознаване и проследяване на части от човешкото тяло без да се претендира за изчерпателност по темата. Представени бяха 3 нови и в своята същност съвсем различни подхода. Grest et al. (4) апроксимира човешката фигура в 3D с виртуален костюм, а Plagemann et al. (6) с теория на графите изследва специални точки с вероятност да представляват човешки крайници. С особено добри резултати се отличава работата Shotton et al. (1), която се приема за State of Art в областта. Тя демонстрира силата на рандомизираните класификационни гори (Random Forest), които с голяма точност класифицират изключително слабия признак – разликата в дълбочината на 2 точки.

Благодарности

Тази разработка е подкрепена финансово от Проект No: BG 051 PO 001-3.3.04/13, "Подкрепа на творческото развитие на докторанти, пост-докторанти и млади учени в областта на компютърните науки", финансиран от ЕВРОПЕЙСКИ СОЦИАЛЕН ФОНД, Оперативна програма "Развитие на човешките ресурси" 2007-2013.

Литература

1. Shotton, J, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. Real-Time Human Pose Recognition in Parts from Single Depth Images. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2011.
2. Comaniciu, D. and P. Meer. Mean shift: A robust approach toward feature space analysis. IEEE Trans. Pattern Anal. Machine Intell., 24(5), 2002.
3. Gonzalez, T. Clustering to minimize the maximum intercluster distance. Theor. Comp. Sci., 38, 1985.
4. Grest, D, J. Woetzel, and R. Koch. Nonlinear body poses estimation from depth images. In Proc. DAGM, 2005.
5. Grest, D, V. Krüger, and R. Koch. Single view motion tracking by depth and silhouette information. In Scandinavian Conference on Image Analysis, 2007.
6. Plagemann, C., V. Ganapathi, D. Koller, and S. Thrun. Real-time identification and localization of body parts from depth images. In Proc. ICRA, 2010.