Available online at www.sciencedirect.com

ScienceDirect

Computer Vision and Image Understanding 104 (2006) 90–126

Computer Vision and Image Understanding

www.elsevier.com/locate/cviu

Review article

A survey of advances in vision-based human motion capture and analysis

Thomas B. Moeslund [a,*], Adrian Hilton [b], Volker Krüger [c]

[a] Laboratory of Computer Vision and Media Technology, Aalborg University, 9220 Aalborg, Denmark
[b] Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, GU2 7XH, UK
[c] Aalborg Media Lab, Aalborg University Copenhagen, 2750 Ballerup, Denmark

Received 22 February 2006; accepted 2 August 2006
Available online 2 October 2006

Abstract

This survey reviews advances in human motion capture and analysis from 2000 to 2006, following a previous survey of papers up to 2000 [T.B. Moeslund, E. Granum, A survey of computer vision-based human motion capture, Computer Vision and Image Understanding, 81(3) (2001) 231–268.]. Human motion capture continues to be an increasingly active research area in computer vision with over 350 publications over this period. A number of significant research advances are identified together with novel methodologies for automatic initialization, tracking, pose estimation, and movement recognition. Recent research has addressed reliable tracking and pose estimation in natural scenes. Progress has also been made towards automatic understanding of human actions and behavior. This survey reviews recent trends in video-based human capture and analysis, as well as discussing open problems for future research to achieve automatic visual analysis of human movement.

Keywords: Review; Human motion; Initialization; Tracking; Pose estimation; Recognition

1. Introduction

Automatic capture and analysis of human motion is a highly active research area due both to the number of potential applications and its inherent complexity. The research area contains a number of hard and often ill-posed problems such as inferring the pose and motion of a highly articulated and self-occluding non-rigid 3D object from images. This complexity makes the research area challenging from a purely academic point of view. From an application perspective computer vision-based methods often provide the only non-invasive solution making it very attractive.

Applications can roughly be grouped under three titles: surveillance, control, and analysis. *Surveillance applications* cover some of the more classical types of problems related to automatically monitoring and understanding locations where a large number of people pass through such as airports and subways. Applications could for example be: people counting or crowd flux, flow, and congestion analysis. Newer types of surveillance applications—perhaps inspired by the increased awareness of security issues—are analysis of actions, activities, and behaviors both for crowds and individuals. For example for queue and shopping behavior analysis, detection of abnormal activities, and person identification.

Control applications where the estimated motion or pose parameters are used to control something. This could be interfaces to games, e.g., as seen in EyeToy [3], Virtual Reality or more generally: Human–Computer Interfaces. However, it could also be for the entertainment industry where the generation and control of personalized computer graphic models based on the captured appearance, shape, and motion are making the productions/products more believable.

Analysis applications such as automatic diagnostics of orthopedic patients or analysis and optimization of an athletes' performances. Newer applications are annotation of

* Corresponding author.
E-mail address: tbm@cvmt.aau.dk (T.B. Moeslund).

1077-3142/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2006.08.002

Table 1
Previous surveys

Year	Author	#Papers	Focus
1994	Aggarwal et al. [10]	69/0	Articulated and elastic nonrigid motion
1994	Cedras and Shah [54]	76/0	Motion extraction
1995	Aggarwal et al. [11]	104/0	Articulated and elastic nonrigid motion
1995	Ju [181]	91/0	Motion estimation and recognition
1997	Aggarwal and Cai [9]	51/0	Motion extraction
1997	Gavrila [114]	87/0	Motion estimation and recognition
2000	Moeslund and Granum [247]	155/0	Initialization, tracking, pose estimation, and recognition
2001	Buxton [48]	88/6	Recognition
2001	Wang et al. [388]	164/14	Detection, tracking, and recognition
2003	Hu et al. [157]	185/54	Surveillance
2004	Aggarwal and Park [12]	58/10	Recognition
2006	This survey	424/331	Initialization, tracking, pose estimation, and recognition

Note that the Year is not necessary the publication year but rather the year of the most recent paper in a survey. The two numbers in the #Papers column state the total number of publications and the publications after 2000.

video as well as content-based retrieval and compression of video for compact data storage or efficient data transmission, e.g., for video conferences and indexing. Another branch of applications is within the car industry where much vision research is currently going on in applications such as automatic control of airbags, sleeping detection, pedestrian detection, lane following, etc.

The number of potential applications, the scientific complexity, the speed and price of current hardware, and the focus on security issues have intensified the effort within the computer vision community towards automatic capture and analysis of human motion. This is evident by looking at the number of publications, special sessions/issues at the major conference/journals as well as the number of workshops directly devoted to such topics. Furthermore, the major funding agencies have also focused on these research fields—especially the surveillance area.

The interest in this area has led to a large body of research which has been digested in a number of surveys, see Table 1.

Even though some of these surveys are recent, it should be noted that the number of papers reviewed after 2000 is limited as seen in the table. In the relatively short period since 2000 a massive number of papers have been published advancing state of the art. This indicates increased activity in this research area compared to the number of papers identified in previous surveys.

Recent contributions have among other things addressed the limiting assumptions identified in previous approaches [247]. For example, many systems now address natural outdoor scenes and operate on long sequences of video containing multiple (occluded) people. This is possible, especially, due to more advanced segmentation algorithms. Other examples are model-based pose estimation where the introduction of learnt motion models and stochastic sampling methods have helped to achieved much faster and more precise results. Also within the recognition area there have been significant advances in both the representation and interpretation of actions and behavior.

Due to the significance of recent advances within this field we present the current survey. The survey is based on 352[1] recent papers (2000–2006) and structured using the functional taxonomy presented in the 2001 survey by Moeslund and Granum [247]:

Initialization. Ensuring that a system commences its operation with a correct interpretation of the current scene.
Tracking. Segmenting and tracking humans in one or more frames.
Pose estimation. Estimating the pose of a human in one or more frames.
Recognition. Recognizing the identity of individuals as well as the actions, activities and behaviors performed by one or more humans in one or more frames.

The different papers are further divided into sub-taxonomies. Inspired by [247] we also provide a visual overview of all the recent referenced papers, see Table 2. For readers new to this field it is recommended to read [247] before preceding with the survey at hand. In fact this survey can be seen as a sequel to [247].

2. Model initialization

Initialization of vision-based human motion capture and analysis often requires the definition of a humanoid model approximating the shape, appearance, kinematic structure, and initial pose of the subject to be tracked. The majority of algorithms for 3D pose estimation continue to use a manually initialized generic model with limb lengths and shape which approximate the individual. To automate the initialization and improve the quality of tracking a limited number of authors have investigated the recovery of

[1] Note that this number is different from the one listed in Table 1 (331). The reason being that we also include papers from the last half of 2000 since this is where the previous survey [247] ends.

Table 2
Publications on human motion capture and analysis from 2000–2006 (inclusive)

Year	First author	Initialisation	Tracking	Pose estimation	Recognition
Publications 2000–2006 (inclusive)					
2000	Barron			[26]	
2000	Buades			[45]	
2000	Chang		*	[56]	*
2000	Davis		[84]		
2000	Deutscher		*	[90]	
2000	Elgammal		[96]		
2000	Felzenszwalb			[104]	
2000	Haritaoglu	*	[138]	*	*
2000	Howe		*	[154]	
2000	Ivanov		*	[171]	
2000	Karaulova	*	*	[189]	
2000	Khan	*	[194]		
2000	Oliver		[270]		
2000	Ormoneit	[274]	*	*	
2000	Ricquebourg		[304]		*
2000	Stauffer		*		[355]
2000	Takahashi		[359]	*	
2000	Taylor	[361]		*	
2000	Trivedi		[367]		
2000	Trivedi		*		[368]
2000	Zhao		[417]		
∑	Total = 21	2	9	8	2
2001	Ambrosio			[16]	
2001	Ambrosio			[17]	
2001	Barron			[27]	
2001	Bobick				[38]
2001	Choo			[62]	
2001	Davison			[85]	
2001	Delamarre		*	[86]	
2001	Deutscher			[91]	
2001	Elgammal	*	[100]		
2001	Grammalidis	*		[124]	
2001	Gutchess		[130]		
2001	Haritaoglu		*		[134]
2001	Herda	*	*	[143]	
2001	Hoshino		*	[150]	
2001	Huang		*	[159]	
2001	Intille				[166]
2001	Ioffe		*	[168]	
2001	Khan		[193]		
2001	Li			[218]	
2001	Mikić	*	*	[239]	
2001	Moeslund	*	*	[247]	*
2001	Moeslund	*	*	[248]	
2001	Mohan		[254]		
2001	Moon		*	[258]	
2001	Ogaki		*	[268]	
2001	Pece	*	[284]		
2001	Plänkers		*	[287]	
2001	Prati		[292]		
2001	Rosales		*	[294]	
2001	Sangi		[320]		
2001	Sato		[321]		*
2001	Sidenbladh	*	*	[332]	
2001	Sminchisescu		*	[342]	
2001	Song			[348]	
2001	Song	[349]			
2001	Zhao		[422]		*
∑	Total = 36	1	9	23	3
2002	Allen	[14]			
2002	Atsushi		[21]		

Table 2 (*continued*)

Year	First author	Initialisation	Tracking	Pose estimation	Recognition
Publications 2000–2006 (inclusive)					
2002	Ben-Arie		*	*	[31]
2002	BenAbdelkader		*		[32]
2002	Bradski		*		[40]
2002	Cheng		*		[58]
2002	Davis		*	*	[83]
2002	Fua		*	[112]	
2002	Gleicher			[118]	
2002	Gonzàlez				[121]
2002	Halvorsen		*	[131]	
2002	Hariadi		[133]		
2002	Haritaoglu		[135]		*
2002	Herda	*		[146]	
2002	Huang		*	[163]	
2002	Ijspeert				[165]
2002	Jang		[173]	*	
2002	Jenkins				[175]
2002	Jenkins				[176]
2002	Jenkins				[177]
2002	Lee	*	*	[212]	
2002	Li		*	[219]	
2002	Metaxas	[234]			
2002	Mikić	*	*	[237]	
2002	Mittal		[243]		
2002	Moeslund	[253]	*	*	
2002	Montemerlo		[257]		
2002	Ozer		[275]	*	*
2002	Park		[280]	*	
2002	Pece		[285]		*
2002	Pers		[286]		
2002	Plänkers		*	[288]	
2002	Rao	*	*		[298]
2002	Ren		*	*	[300]
2002	Rittscher	*	*	*	[305]
2002	Roberts		*	[309]	
2002	Ronfard			[314]	
2002	Sidenbladh	*		[335]	
2002	Sminchisescu		*	[339]	
2002	Theobalt	*	*	[364]	
2002	Utsumi	*	[374]		
2002	Yam		*		[402]
2002	Zhao		[418]		
$\sum$	Total = 43	3	12	14	14
2003	Allen	[15]			
2003	Azoz		*	[22]	
2003	Babu				[23]
2003	Barron	[28]		*	
2003	Buxton				[48]
2003	Capellades		[52]		*
2003	Carranza	*	*	[53]	
2003	Cheung	*	*	[59]	
2003	Chowdhury				[63]
2003	Chu	*		[65]	
2003	Comaniciu		[67]		
2003	Cucchiara		[69]		
2003	Davis				[79]
2003	Demirdjian	*		[87]	
2003	Demirdjian	*		[89]	
2003	Efros				[94]
2003	Elgammal		[95]		
2003	Elgammal				[99]
2003	Eng		[101]		*
2003	Foster		*		[111]
2003	Gerard	*		[115]	

(*continued on next page*)

Table 2 (*continued*)

Year	First author	Initialisation	Tracking	Pose estimation	Recognition
Publications 2000–2006 (inclusive)					
2003	Gonzalez		[122]	*	
2003	Grauman			[125]	
2003	Herda		*	[142]	
2003	Jepson		[178]		
2003	Koschan		[198]		
2003	Krahnstoever	[201]	*	*	
2003	Liebowitz	*		[220]	
2003	Masoud				[231]
2003	Mikić	*	*	[238]	
2003	Mitchelson			[241]	
2003	Mitchelson		*	[242]	
2003	Mittal		[244]		
2003	Moeslund	*	*	[245]	
2003	Moeslund		*	[249]	
2003	Moeslund		*	[250]	
2003	Monnet		[256]		
2003	Parameswaran				[277]
2003	Plänkers		*	[289]	
2003	Polat		[290]		
2003	Prati		[293]		
2003	Shah		[325]	*	*
2003	Shakhnarovich			[326]	
2003	Sidenbladh	*	[333]	*	
2003	Sminchisescu		*	[343]	
2003	Sminchisescu		*	[344]	
2003	Song	[350]	*		*
2003	Starck	[351]			
2003	Starck	[352]		*	
2003	Störring		[357]		
2003	Vasvani				[375]
2003	Vecchio				[376]
2003	Viola		[381]		
2003	Wang				[387]
2003	Wang		[388]	*	*
2003	Wang		*	*	[389]
2003	Wang		*		[390]
2003	Wang		[391]		
2003	Wu			[398]	
2003	Yang		[405]		
2003	Zhao		[419]		
2003	Zhong		[423]		
$\sum$	Total = 62	6	21	21	14
2004	Agarwal			[6]	
2004	Agarwal		*	[7]	
2004	Agarwal				[12]
2004	Billard				[34]
2004	Bregler			[43]	
2004	Brostow	[44]			
2004	Chowdhury				[64]
2004	Cucchiara		[70]		
2004	Date			[78]	
2004	Davis				[81]
2004	Davis		[82]		
2004	Demirdjian			[88]	
2004	Elgammal				[97]
2004	Elgammal			[98]	
2004	Figueroa		[105]		
2004	Gao		[113]		*
2004	Giebel			[116]	
2004	Gonzàlez				[120]
2004	Gritai				[127]
2004	Hayashi		[139]		
2004	Heikkila		[140]		

Table 2 (*continued*)

Year	First author	Initialisation	Tracking	Pose estimation	Recognition
Publications 2000–2006 (inclusive)					
2004	Herda			[144]	
2004	Howe	*		[152]	
2004	Hu		[155]		
2004	Hu		[157]	*	*
2004	Huang	*	*	[160]	*
2004	Iwase		[172]		
2004	Junejo	*			[182]
2004	Kang		[186]	*	
2004	Krahnstoever	[200]		*	
2004	Lee	*	*	[210]	
2004	Lee	*	*	[211]	
2004	Leo				[217]
2004	Loy			[225]	
2004	Lu				[226]
2004	Lv		*		[228]
2004	Mikolajczyk		*	[240]	
2004	Moeslund		*	[251]	
2004	Mori			[261]	
2004	Murakita		[263]		
2004	Okuma		[269]		
2004	Pan		[276]		
2004	Parameswaran	[278]		*	
2004	Park				[281]
2004	Porikli				[291]
2004	Remondino	[299]			
2004	Ren				[301]
2004	Roberts			[310]	
2004	Sidenbladh	*	[331]		
2004	Sigal			[336]	
2004	Thalmann	[363]			
2004	Urtasun		*	[373]	
2004	Wu		[397]		
2004	Yang		[406]		
2004	Yang		[407]		
2004	Yi				[409]
2004	Zhao		[420]		
2004	Zhao	*	[421]		
$\sum$	Total = 58	5	19	19	15
2005	Andersen		[18]		
2005	Balan			[25]	
2005	Beleznai		[29]		
2005	Blank				[36]
2005	Boiman				[39]
2005	Bullock	*	*	[47]	*
2005	Calinon				[50]
2005	Calinon				[51]
2005	Chalidabhongse		[55]		
2005	Chen	*	[57]		
2005	Cheung	*		[60]	
2005	Cucchiara		*		[71]
2005	Curio			[73]	
2005	Dahmane		[74]		*
2005	Dalal		[75]		
2005	Deutscher			[92]	
2005	Dimitrijevic	[93]		*	
2005	Fanti		*		[103]
2005	Guha		[129]		
2005	Herda	[145]		*	
2005	Howe			[151]	
2005	Kang		[187]		
2005	Kang		[188]		

(*continued on next page*)

Table 2 (*continued*)

Year	First author	Initialisation	Tracking	Pose estimation	Recognition
Publications 2000–2006 (inclusive)					
2005	Ke				[190]
2005	Kehl			[191]	
2005	Kim		[196]		
2005	Krosshaug			[203]	
2005	Krüger	*	[204]		
2005	Kumar	[205]	*	*	
2005	Lee		[208]		
2005	Lee	*	*	[214]	
2005	Leibe		[215]		
2005	Lim		[222]		
2005	Micilotta			[235]	
2005	Moeslund	*	*	[246]	
2005	Moeslund	[252]	*	*	
2005	Mulligan		*	[262]	
2005	Navaratnam			[265]	
2005	Ong			[272]	
2005	Ormoneit			[273]	
2005	Ramanan	*		[296]	
2005	Ren			[302]	
2005	Robertson				[311]
2005	Roth		[316]		
2005	Sanfeliu		[319]		
2005	Sheikh		[327]		
2005	Sheikh				[328]
2005	Sminchisescu			[340]	
2005	Smith		[345]		
2005	Smith				[346]
2005	Starck	*	*	[353]	
2005	Toyosawa		[366]		*
2005	Ukita		[369]		
2005	Urtasun		*	[370]	
2005	Urtasun		*	[372]	
2005	Veeraraghavan		*		[378]
2005	Viola		[382]		
2005	Wang	*	[385]		
2005	Weinberg				[393]
2005	Wu	*	[396]		
2005	Xu		[401]		
2005	Yang		[404]		
2005	Yang		[408]		
2005	Yilmaz				[410]
2005	Yilmaz				[411]
2005	Yu		*		[412]
2005	Yu				[413]
2005	Zhang			[414]	
2005	Zhao			[415]	
2005	Zhao		[416]		
$\sum$	Total = 70	4	27	23	16
2006	Agarwal		*	[8]	
2006	Ahmad				[13]
2006	Antonini	*	[19]		
2006	Balan			[24]	
2006	Berclaz	[33]			
2006	Bissacco				[35]
2006	Bray			[42]	
2006	Buades	[46]	*	*	
2006	Cuntoor				[72]
2006	Dalal		[76]		
2006	Eng		[102]		
2006	Figueroa		[106]		
2006	Figueroa		[107]		
2006	Fihl		[108]		

Table 2 (*continued*)

Year	First author	Initialisation	Tracking	Pose estimation	Recognition
Publications 2000–2006 (inclusive)					
2006	Fihl		*		[109]
2006	Han			[132]	
2006	Heikkila		[141]		
2006	Howe		*	[153]	
2006	Hu		[156]		
2006	Huang				[161]
2006	Huerta		[164]		
2006	Jaeggli			[174]	
2006	Jiang				[180]
2006	Khan		[195]		
2006	Kristensen		[202]		
2006	Lee		[206]	*	
2006	Lee	[207]			
2006	Lee			[213]	
2006	Leichter		[216]		
2006	Lim	[221]	*		
2006	Liu				[223]
2006	Lv				[229]
2006	Menier	*		[233]	
2006	Micilotta			[236]	
2006	Moon			[259]	
2006	Mori	[260]		*	
2006	Nillius	[266]			
2006	Parameswaran				[279]
2006	Park		[282]	*	
2006	Park		[283]		
2006	Rahman				[295]
2006	Ramanan			[297]	
2006	Reng				[303]
2006	Rius			[306]	
2006	Roh				[313]
2006	Ryoo				[318]
2006	Schindler		[323]		
2006	Shi				[330]
2006	Sigal			[337]	
2006	Sigal			[338]	
2006	Sminchisescu			[341]	
2006	Smith				[347]
2006	Sundaresan	[358]			
2006	Taycher			[360]	
2006	Urtasun			[371]	
2006	Veeraraghavan				[377]
2006	Wang		[386]		
2006	Wang				[392]
2006	Wu	*	[395]		
2006	Wu	*	[399]		
2006	Xiang				[400]
2006	Yamamoto				[403]
$\sum$	Total = 62	7	19	17	19
00–06	Total = 352	28	116	125	83

Papers are ordered first by the year of publication and second by the surname of the first author. Four columns allow the clarification of the contributions of the papers within the four processes. The location of the reference number (in brackets) indicates the main topic of the work and an asterisk (*) indicates that the paper also describes work at an interesting level regarding this process.

more accurate reconstructions of the subject from single or multiple view images.

Initialization captures prior knowledge of a specific person which can be used to constrain tracking and pose estimation. A priori knowledge used in human motion capture can be broken into a number of sources: kinematic structure; 3D shape; color appearance; pose; motion type. In this section, we review recent research which advances estimation of kinematic structure, 3D shape, and appearance. Initialization of appearance is commonly an integral part of the tracking and pose estimation and is therefore also considered in conjunction with specific approaches in Sections 3 and 4.

2.1. Kinematic structure initialization

The majority of vision-based tracking systems assume a priori a humanoid kinematic structure comprising a fixed number of joints with specified degrees-of-freedom. The kinematic initialization is then limited to estimation of limb lengths. Commercial marker-based motion capture systems typically require a fixed sequence of movements which isolate individual degrees-of-freedom. The known correspondence between markers and limbs together with reconstructed 3D marker trajectories during movement are then used to accurately estimate limb lengths. Hard constraints on left-right skeletal symmetry are commonly imposed during estimation. A number of approaches [26,28,278,361] have addressed initialization of body pose and limb lengths from manually identified joint locations in monocular images. Anthropometric constraints between ratios of limb lengths are imposed to allow estimation of the kinematic structure up to an unknown scale factor.

Direct estimation of the kinematic structure from sequences of a moving person has also been investigated. Krahnstover et al. [200,201] present a method for automatically initializing the upper-body kinematic structure based on motion segmentation of a sequence of monocular video images. Song et al. [350] introduce an unsupervised learning algorithm which uses point feature tracks from cluttered monocular video sequences to automatically construct triangulated models of whole-body kinematics. Learnt models are then used for tracking of walking motions from lateral views. These approaches provide more general solutions to the problem of initializing a kinematic model by deriving the structure directly from the scene.

Methods that derive the kinematic structure from 3D shape sequences reconstructed from multiple views have also been proposed. Cheung et al. [59] initialize the kinematic structure from the visual-hull of a person moving each joint independently. A full-skeleton together with the shape of each body part is obtained by alignment of the segmented moving body parts with the visual-hull model in a fixed pose. Menier et al. [233] present an automated approach to 3D human pose estimation from the medial axis of the visual-hull. The kinematic structure is initialized independently at each frame enabling robust tracking. More general frameworks are presented in [44,65] to estimate the underlying skeletal spine structure from a temporal sequence of the 3D shape. The spine is estimated from the shape at each frame and common temporal structures identified to estimate the underlying structure. This work demonstrates reconstruction of approximate kinematic structures for babies, adults, and animals.

Initializing the joint angle limits on the human kinematic structure is an important problem to constrain motion estimation to valid postures. Manual specification of joint angle limits has been common in many motion estimation algorithms using anthropometric data. This does not take into account the complex nature of joint limits and coupling between limits for different degrees-of-freedom. To overcome these limitations recent research has investigated learning models of joint limits and their correlations. Anthropometric models for the relationship between arm joint angles (shoulder, elbow, and wrist) have been used to provide constraints in visual tracking and 3D upper-body pose estimation [248,253,262]. Recent research has investigated the modeling of joint limits from measurements of human motion captured using marker-based systems [144,145] and from clinical data [252]. This is demonstrated to improve the performance of human pose estimation for complex upper-body movement.

Increasingly, human motion capture sequences from commercial marker-based systems have been used to learn prior models of human kinematics and specific motions to provide constraints for subsequent tracking. Similarly motion capture data-bases [1,2,4] have recently been used to syntheses image sequences with know 3D pose correspondence to learn a priori the mapping from image to pose space for reconstruction.

2.2. Shape initialization

A generic humanoid model is used in many video-based human motion estimation techniques to approximate a subject's shape. Representations have used either simple shape primitives (cylinders, cones, ellipsoids, and superquadrics) or a surface (polygonal mesh, sub-division surface) articulated using the kinematic skeleton [247]. A number of approaches have been proposed to refine the generic model shape to approximate a specific person.

In previous research [147] a generic mesh model was refined based on front and side view silhouettes taken with a single camera. Texture mapping was then applied to approximate detailed surface appearance. Recently simultaneous capture from multiple calibrated views has been used [53,289,352] to achieve more accurate shape and appearance. Plänkers and Fua [289] initialize upper-body shape by fitting an implicit ellipsoidal metaball representation to stereo point clouds prior to tracking. Carranza et al. [53] fit a generic mesh model to multiple view silhouette images of a person in a fixed pose prior to tracking whole-body motion. Starck and Hilton [352] reconstruct whole-body shape and appearance for a person in an arbitrary pose by optimizing a generic mesh model with respect to both silhouette, stereo, and feature correspondence constraints in multiple views. These model fitting approaches provide an accurate parameterized approximation of a person provided the assumed shape of the generic model is a reasonable initial approximation. Model fitting methods commonly assume short hair and close fitting clothing which limits their generality.

The availability of sensors for whole-body 3D scans provides accurate measurement of surface shape. Techniques to fit generic humanoid models to the whole-body scans in a specific pose enable a highly detailed representation of a person's shape to be parameterized for animation and tracking

[14,351]. Allen et al. [14] fit a sub-division surface to multiple scans of a person in different poses to parameterize the change in body surface shape with pose. Databases of 3D scans have also been used to learn statistical models of the inter-person variation in whole-body shape [15,363]. Reconstruction of shape from images can then be constrained by the learnt model to improve performance.

2.3. Appearance initialization

Due to the large intra and inter person variability in appearance with different clothing, initialization of appearance has commonly been based on the observed image set. Statistical models of color are commonly used for tracking, see Section 3.3. Initialization of the detailed surface appearance for model-based pose estimation has also used texture maps derived from multiple view images [53,352]. A cost function evaluating the difference in appearance between the projected model and observed images is then used in pose estimation.

Sidenbladh and Black [332,333] address modeling the likelihood of image observations for different body parts. They learn the statistics of appearance and motion based on filter responses for a set of training examples. In a related approach, Roberts et al. [309] learn the likelihood of body part color appearance using multimodal histograms on a 3D surface model. Results are presented for 2D tracking of upper-body and walking motions in cluttered scenes.

A recent trend has been towards the learning of body part detectors to identify possible locations for body parts which are then combined probabilistically to locate people [235,296,310,314], see Section 4.1.1. Initialization of such models requires a large training corpus of both positive and negative training examples for different body parts. Approaches such as AdaBoost have been successfully used to learn body part detectors such as the face [380], hands, arms, legs, and torso [235,310]. Alternatively, Ramanan et al. [296] detect key-frame poses in walking sequences and initialize a local appearance model to detect body parts at intermediate frames.

Lim et al. [221] address the problem of changing appearance due to motion by modeling the dynamics of the appearance for walking humans. This is done by mapping the pixels inside a bounding box to a low dimensionality space (only 3D) using a nonlinear Local Linear Embedding algorithm. In this space the temporal continuity of the appearance is preserved, which allows for learning a dynamic model of the appearance for walking humans. This model can then be used to predict not only the position and 2D shape of a walking human, but also the appearance.

The initialization of models which accurately represent the change in appearance over time due to creases in clothing, hair, and change in body shape with movement remains an open problem. Recent introduction of robust local body part detectors provides a potential solution for tracking and pose estimation.

2.4. Discussion of advances in model initialization

Initialization of shape, appearance, and pose remains an import step to automate the process of human motion capture and analysis. As illustrated in this review significant advances have been made towards automatic solutions. The problem of initializing the kinematic structure and pose from feature tracks for monocular sequences has been addressed [350]. A number of researchers have presented methods for initializing the kinematic structure from multiple view image sequences using an intermediate volumetric reconstruction [59,233]. These approaches provide a solution to the problem of automatic kinematic model initialization for human pose estimation. Learning approaches [145] and anthropometric models [252] have been presented to initialize the joint angle limits on the kinematic structure to constrain tracking and pose estimation.

Over the past 5 years there has been substantial research in the automatic initialization of model shape from multiple view images [53,59,289,352]. These approaches reconstruct an articulated model which approximates the shape of a specific person providing the basis for improved accuracy in tracking. Recent research has also started to address the modeling of changes in human body shape during movement [14]. Similarly multiple view reconstruction techniques have allowed the automatic initialization of model appearance to that of a specific individual.

Initialization of appearance models for monocular tracking and pose estimation remains an open problem. A number of approaches have been proposed for initialization of appearance based on image patch exemplars or color mixture models. Recent work on body part detectors has exploited supervised learning approaches to discriminate individual body part appearance from background [296,310,314]. Only limited research has addressed the problem of modeling changes in a person's appearance during movement. The problem of fully automatic initialization of model kinematics, shape, and appearance for human pose estimation from monocular image sequences remains open for future research.

3. Tracking

Since 2000 tracking algorithms have focused primarily on surveillance applications leading to advances in areas such as outdoor tracking, tracking through occlusion, and detection of humans in still images. In this section we review recent advances in these areas as well as more general tracking problems.

The notion of *tracking* in visual analysis of human motion is used differently throughout the literature. Here we define it as consisting of two processes: (1) *figure-ground segmentation* and (2) *temporal correspondences*. The latter, temporal correspondences, is the process of associating the detected humans in the current frame with those in the previous frames, providing temporal trajectories through the state space. Recent advances are mainly due to processing more natural scenes where multiple people and occlusions are present.

Figure-ground segmentation is the process of separating the objects of interest (humans) from the rest of the image (the background). Methods for figure-ground segmentation are often applied as the first step in many systems and therefore a crucial process. Recent advances are mostly a result of expanding existing methods. We categorize these methods in accordance with the type of image measurements the segmentation is based on: motion, appearance, shape, or depth data. Before describing these we first review recent advances in background subtraction as this has become the initial step in many tracking algorithms.

3.1. Background subtraction

Up until the late 90s background subtraction was known as a powerful preprocessing step but only in controlled indoor environments. In 1998, Stauffer and Grimson [354] presented the idea of representing each pixel by a mixture of Gaussians (MoG) and updating each pixel with new Gaussians during run-time. This allows background subtraction to be used in outdoor environments. Normally the updating was done recursively, which can model slow changes in a scene, but not rapid changes like clouds. The method by Stauffer and Grimson has today become the standard of background subtraction. However, since 1998 a number of advances have been seen which can be divided into *background representation*, *classification*, *background updating*, and *background initialization*.

3.1.1. Background representation

The MoG representation can be in RGB space, but also other color spaces can be applied, see [202] for an overview. Often a representation where the color and intensities are separated is applied, e.g., YUV [394], HSV [69], and normalized RGB [232], since this allows for detecting shadow-pixels wrongly classified as object-pixels [293]. Using a MoG in a 3D color space corresponds to ellipsoids or spheres (depending on the assumptions on the covariance matrix) of the Gaussian representations [232,354,421]. Other geometric representations are truncated cylinders [196] and truncated cones [108].

Conceptually different representations have also been developed. Elgammal et al. [96] use a kernel-based approach where they represent a background pixel by the individual pixels of the last N frames. Haritaoglu et al. [138] represent the minimum and maximum value together with the maximum allowed change of the value in two consecutive frames. Heikkila and Pietikainen [141] represent each background pixel by a bit sequence, where each bit reflects whether the value of a neighboring pixel is above or below the pixel of interest, i.e., a texture operator. This makes the background model invariant to monotonic illumination changes. Oliver et al. [270] also use a pixel's neighbors to represent it. They apply an eigenspace representation of the background and detect new objects by comparing the input image with an image reconstructed via the eigenspace.

Eng et al. [101,102] divide a background model learnt over time into a number of non-overlapping blocks. The pixels within each block are grouped into at most three classes according to homogeneity. The means of these classes are then the representation of the background for this block, i.e., a spatio-temporal representation. Heikkila and Pietikainen have also applied their texture operator for a spatio-temporal block-based (overlapping blocks) background segmentation [140]. Other spatio-temporal approaches are [256] and [423] where the background is represented by a predicted region found by an autoregressive process.

The choice of representation is not only dependant on the accuracy but also on the speed of the implementation and the application. This makes sense since the overall accuracy of background subtraction is a combination of representation, classification, updating, and initialization. For example, Cucchiara et al. [69] use only one value to represent each background pixel, but still good results (and speed) can be obtained due to advanced classification and updating. It should however be noted that the MoG representation is by far the most widely used method.[2] For scenes with dynamic background the MoG representation does not suffice and methods directly aimed at modeling dynamic background should be applied, see e.g., [256,327, and 423].

3.1.2. Classification

A number of false positives and negatives will often be present after a background subtraction, for example due to shadows [293]. Using standard filtering techniques based on connected component analysis, size, median filter, morphology, and proximity can improve the result [69,96,129,232,408,420]. Alternatively, the fact that neighboring pixels are likely to be both foreground *or* background can be used in classification. Markov Random fields have been applied to implement this idea [323,327].

Recent methods have tried to directly identify the incorrect pixels and use classifiers to separate the pixels into a number of sub-classes: unchanged background, changes due to auto iris, shadows, highlights, moving object, cast shadow from moving object, ghost object (false positive), ghost shadow, etc. [57,69,149]. Classifiers have been based on color, gradients [232], flow information [69], and hysteresis thresholding [101].

3.1.3. Background updating

In outdoor scenes, in particular, the value of a background pixel will change over time and an update mechanism is therefore required. The slow changes in the scene can be updated recursively by including the current

[2] See [208,424] for optimizations of the MoG representation.

pixel value into the model as a weighted combination [69,96,232,354]. A different approach is to measure the overall average change in the scene compared to the expected background and use this to update the model [108,408]. If no real-time requirements are present, both past and future values can be used to update the background [106]. In general, for a good model update only pixels classified as unchanged background should be updated.

Rapid changes in the scene are accommodated by adding a new mode to the model. For the MoG model a new mode is a new Gaussian distribution, which is initiated whenever a non-background pixel is detected. The more pixels (over time) that support this distribution the more weight it will have. A similar approach is seen in [108,196] where the background model, denoted a codebook, for each pixel is represented by a number of codewords (cylinders [196] or cones [108] in RGB-space). During run-time each foreground pixel creates a new codeword. A codeword not having any pixels assigned to it for a certain number of frames is eliminated. A similar idea can be found in [140,141].

3.1.4. Background initialization

A background model needs to be learned during an initialization phase. Earlier approaches assumed that no moving objects are present in a number of consecutive frames and then learn the model parameters in this period. However, in real scenarios this assumption will be invalid and recent methods have therefore focused on initialization in the presence of moving objects.

In the MoG representation moving objects can to some extend be accepted during initialization since each foreground object will be represented by its own distribution which is likely to have a low weight. However, this erroneous distribution is likely to produce false positives in the classification process. A different approach is to find only pixels that are true background pixels and then only apply these for initialization. This can be done using a temporal median filter if less than 50% of the values belong to foreground objects [101,119,138]. Eng et al. [101] combine this with a skin detector to find and remove humans from the training images.

Recent alternatives first divide the pixels in the initialization phase into temporal subintervals with similar values. Second, the ''best'' subinterval belonging to the background is found as the subinterval with the minimum average motion (measured by optical flow) [130] or the subinterval with the maximum ratio between the number of samples in the subinterval and their variance [385,386]. The codeword method mentioned above uses a temporal filter after the initialization phase to eliminate any codeword that has not recurred for a long period of time [196]. A similar approach has used in [140,141].

For comparative studies among some of the different background subtraction methods see [55,61,385,386].

3.2. Motion-based segmentation

Motion-based figure-ground segmentation is based on the notion that differences in consecutive images arise from moving humans, i.e., by finding the motion you find the human. The motion is measured using either flow or image differencing.

Sidenbladh [331] calculates optical flow for a large number of image windows each containing a walking human. A support vector machine (SVM) is used to detect walking humans in video. Optical flow can be noisy and instead image flow can be measured using higher level entities. For example, Gonzalez et al. [122] track KLT-features to obtain flow vectors, Sangi et al. [320] extract flow vectors from displacements of pixel-blocks, and Bradski and Davis [40] find flow vectors as gradients in motion history images (MHI) [80].

Image differencing adapts quickly to changes in the scene, but pixels from a human that has not moved or are similar to their neighbors are not detected. Therefore, an improved version is to use three consecutive images [66,138,185]. A different type of image differencing is used by Viola et al. [382]. They apply the principle of their novel face detector [380], where simple features are combined in a cascade of progressively more advanced classifiers. A rectangle of pixels in the current image is compared to the corresponding rectangle in the previous image. This is done by shifting the rectangle in the current image up, down, left, and right. Image differencing is then preformed and the lower the energy in the output the higher the probability that the human has actually moved (shifted) in this direction. The output of these operations is used to build a person detector, which is trained using AdaBoost.

3.3. Appearance-based segmentation

Segmentation based on the appearance of the human is built on the idea that (1) the appearance of human and background is different and (2) the appearance of individuals are different. The approaches work by building an appearance model of each human and then either building appearance models of the segmented foreground objects in the current image and comparing them with the predicted models, or by directly segmenting the pixels in the current image that belong to each model. Some of these methods are independent on the temporal context, meaning that the methods apply a general appearance model of a human, as opposed to methods where the appearance model of the human is learned/updated based on previous images in the current sequence.

3.3.1. Temporal context-free

Temporal context-free methods are used to detect humans in a still image [254], to detect humans entering a scene [269], or to index images in databases [275]. Advances are mostly on using massive amount of training data for learning good classifiers. For example, Okuma et al. [269]

use 6000 images to train an Adaboost-based classifier. Other examples are using DCT coefficients [275], using partial-occlusion handling body-part detectors [254], (see also Section 4.1.1), or the block-based method by Utsumi and Tetsutani [374]. In [374] the image is divided into a number of blocks and the mean and covariance matrix of the intensities are calculated for each block. A distance matrix is constructed where an entry represents the generalized Mahalanobis distance between two blocks. The detection is now based on the fact that for non-human images the distances between blocks in the proximity will be larger than for images containing a human.

Common for these methods is that the human is detected as a box (normally a bounding box) and clutter in the background will therefore have an effect on the results. Furthermore, as the methods usually represent the human as one entity, as opposed to a number of sub-entities, occlusion will in general effect the methods strongly. Drastic illumination changes will also effect the methods since the models are general and do not adapt to the current scene.

3.3.2. Temporal context

Temporal context refers to methods where a model which is learned and updated in previous images is used to either detect foreground pixels or to classify foreground pixels to a particular human being tracked. The methods either operate at pixel level or region level. At pixel level the likelihood of each (foreground) pixel belonging to a human model is calculated. The region level is when a region in the image, such as a bounding box, is compared to an appearance model of the humans that are predicted to be present in the current frame, i.e., the probability that a region in an image corresponds to a particular human model. Color-based appearance models have recently received attention leading to advances allowing tracking in outdoor scenes with partial occlusion. This has led to a need for models that can represent the differences between individuals even during partial occlusion.

In many systems the color of a human is represented as either a color histogram [67,155,232,269,401,421] or a MoG [187,194,316,404].[3] Color histograms are normally compared using the Bhattacharyya distance, which can be improved by weighting pixels close to the center of the human higher than those close to the border [67,421]. In Zhao [421] the similarity is combined with the dissimilarity with respect to the color histogram of the background. MoG representations are normally compared using the Mahalanobis distance, which can be evaluated efficiently by using only one Gaussian [187] and assuming independence between color channels [70]. Alternatively, only the mean can be used [404].

Representing the entire human by just one color model is often too coarse a representation even though the model contains multiple modes. Recent advances are therefore on including spatial information. For example using a Correlogram, which is a co-occurrence matrix that expresses the probability of two different colored pixels being found at a certain distance form each other [52,162]. Another way of adding spatial information is to divide the human into a number of sub-regions and represent each sub-region with either a color histogram or a MoG [244,269,316,404]. Hu et al. [155] use an adaptive approach to obtain three sub-regions representing the head, torso, and legs. A more general approach is to model the human as a number of blobs where each blob is a connected group of pixels having a similar color [194,282]. Grouping the blobs together temporally and spatially into an entire human requires some bookkeeping, but a rough human model can assists as seen in [282].

As mentioned in Section 2—Model initialization—appearance-based models able to handle changes over time remains an open issue. On one hand a model should adapt quickly to changes, but on the other hand long term temporal consistency is required, e.g., to handle occlusions. The KLT-tracker [329] to some degree handles this dilemma by only updating the model by data from the previous image as long as it is not too different from the initial model. A more general framework is suggested by Jepson et al. [178]. They update each pixel in their appearance model by a weighted combination of a slowly changing model, a fast changing model, and a noise model. The weights are updated in accordance with the support of the different models in the current image.

3.4. Shape-based segmentation

The shape of a human is often very different from the shape of other objects in a scene. Shape-based detection of humans can therefore be a powerful cue. As opposed to the appearance-based models, the shapes of individuals are often very similar. Hence, shape-based methods applied to tracking only involves simple correspondences. The advances are first of all to allow human detection and tracking in uncontrolled environments. Due to the recent advances in background subtraction reliable silhouette outlines can describe the shape of the humans in the image sequence. Furthermore, advances in representations and segmentation methods of humans in still images have also been reported. As was done for the appearance-based methods, we divide the shape-based methods into those not using the temporal context and those using the context.

3.4.1. Temporal context-free

Zhao and Thorpe [417] use depth data to extract the silhouettes of individuals in the image. A neural network is trained on upright humans and used to verify whether the extracted silhouettes actually originate from humans or not. To make the method more robust the gradients

[3] According to McKenna et al. [232] MoG is preferred with small sample sets and many possible colors, whereas a color histogram is preferred when many color samples are present in a coarsely quantified color space.

of the outline of a silhouette are used to represent the shape of the human. Leibe et al. [215] learn the outlines of walking humans and store them as a number of templates. Each of these are matched with an edge version of the input image over different scales using Chamfer matching. The results are combined with the probability of a person being present, which is measured by comparing small learned image patches of the appearance of humans and their occurrence distribution. Wu and Yu [399] learn a prior shape model for human edges and represent it as a Boltzmann distribution in a Markov Field. The detector searches for different locations, scales, and rotations and is implemented using a Particle Filter. Dalal and Triggs [75] use an SVM to detect humans in a window of pixels. The input is a set of features encoding the shape of a human. The features come from using a spatially arranged set of HOG (histogram of oriented gradients) descriptors. The HOG descriptor operates by dividing an image region into a number of cells. For each cell a 1D histogram of gradient directions over the pixels in the cell is calculated. In [76] the work is extended by including motion histograms. This allows for detecting humans even when the camera and/or background is moving. HOGs are related to Shape Contexts [30] and SIFT (scale invariant feature transformation) [224]. Zhao and Davis [416] learn a hierarchy of silhouette templates for the upper body. The outline of the silhouettes in the templates is used to detect sitting humans in a frame. This is done using Chamfer matching at different scales together with a color-based detector that is updated iteratively.

3.4.2. Temporal context

When the temporal context is taken into consideration shape-based methods can be applied to track individuals over time. In case of temporal smoothness the shape in the previous frame can be used to find the human in the current frame. Haritaoglu et al. [138] perform a binary edge correlation between the outlines of the silhouettes in the last frame and the immediate surroundings in the current image. Davis et al. [84] use a point distribution model (PDM) to represent the outline of the human. The most likely configurations of the outline from the previous frame are used to predict the location in the current frame using a particle filter. Predictions are evaluated by comparing the edges of the outline with those in the image. A similar approach is seen in [198] where the active shape model is applied to find a fit in the current frame. Atsushi et al. [21] model the pose of the human in the previous frame by an ellipse and predict nine possible poses of the human in to the current frame. Each of these is correlated with the silhouettes in the current image in order to define the current pose of the human. Krüger et al. [204] correlate the extracted silhouette with a learned hierarchy of silhouettes of walking persons. At run-time a Bayesian tracking framework concurrently estimates the translation, scale, and type of silhouette.

In situations of partial occlusion the shape-based methods just described often fail due to lack of global shape information. Advances therefore include detection of humans based on only a few parts of the overall shape. In the work by Wu and Nevatia [396] four different (body) parts are detected: full-body, head-shoulder, torso, and legs. For each part a detector is trained using a boosting classifier together with edgelets (small connected chains of edge pixels) which are quantified into different orientations, see also Section 4.1.1. When people group together the occlusion often becomes severe and the only reliably shape information is the head or head-shoulder profile. While this work is limited to frontal/rear views, extended work also handles side views [395].

In [138,156,406] the head candidates are found by analyzing the silhouette boundary and the vertical projected histogram of the silhouette. A similar approach is seen in [419] except that also an edge-based method to find the head-shoulder profile inside silhouettes is applied.

3.5. Depth-based segmentation

Figure-ground segmentation using depth data are based on the idea that the human stands out in a 3D environment. Methods are either-based directly on estimated 3D data for the scene [135,139,171,222,407] or indirectly by combining different camera views after features have been extracted [172,243,244,405]. Advances are mainly due to faster computers allowing for handling multiple camera inputs.

Background subtraction can be sensitive to lighting changes. Therefore a depth-based approach can be taken where the background is modeled as a depth model and compared to estimated depth data for each incoming frame in order to segment the foreground. A real-time dense stereo algorithm is, however, still problematic unless special hardware is applied [222]. An approach to circumvent this is the work by Ivanov et al. [171] where an online depth map is not required. Instead the mapping between pixels in two cameras is learnt. This allows for an online comparison between associated pixels (defined by the mapping) in the two cameras. Detection is now performed based on the assumption that the color and intensity are similar for the pixels if and only if they depict the background. In [222] the merits and drawbacks of this approach are studied in detail.

Other advances in human detection based on depth data include the work by Haritaoglu et al. [135] where depth data produced by ceiling-mounted cameras are projected to the ground-plane. Here humans are located by looking for a 3D head-shoulder profile. Similar approaches are seen in [139,407] except for the camera placement and that [139] apply voxels as opposed to 3D points.

Mittal and Davis [243,244] detect humans using an appearance-based method in each camera view. The center of each detected human is combined with those found in another image using region-based stereo constrained by the epipolar geometry. The resulting 3D points are projected to the ground-plane and represented probabilistically using Gaussian kernels and an occlusion likelihood. In Yang et al. [405] silhouettes from different cameras are combined into the visual hull. The incorrect interpretations are pruned using a size criterion as well as the temporal history. Iwase and Saito [172] apply multiple cameras to detect and track multiple people. In each camera the feet of each person are detected using background subtraction and knowledge of the environment. For each camera all detected feet are mapped to a virtual ground-plane where an iterative procedure resolves ambiguities. A similar approach can be found in [195].

3.6. Temporal correspondences

One of the primary tasks of a tracking algorithm is to find the temporal correspondences. That is, given the state of N persons in the previous frame(s) and the current input frame(s), what are the states of the same persons in the current frame(s). Here the state is mainly the image position of a person, but can contain other attributes, e.g., 3D position, color, and shape.

Previously tracking algorithms were mostly tested in controlled environments and with only a few people present in the scene. Recently, algorithms have addressed more natural outdoor scenarios where multiple people and occlusions are present. One problem is to have better figure-ground segmentation as discussed above. Another equally important problem is how to handle multiple people that might occlude each other. In this section we discuss advances related to temporal correspondences *before and after occlusion* and temporal correspondences *during occlusion*.

3.6.1. Temporal correspondences before and after occlusion

A model of each individual must be constructed before any tracking can commence. Recent methods are aiming at doing this automatically. One way is to look for (new) large foreground objects possible near the boundaries[4] [18,19,138,232,316]. Alternatively, a new person can be defined as a foreground object detected far from any predictions [52]. Khan and Shah [194] fit 1D Gaussians to the foreground pixels projected to the horizontal axis. If the number of good fits is higher than the predicted number of people in the scene then a new person has entered the scene.

When the tracking has commenced the problem is to find the temporal correspondences between predicted and measured states. This has recently been approached using a correspondence matrix, which has the predicted objects in one direction and the measured objects in the other direction. For each entry in the matrix a distance between predicted and measured object is calculated. This gives the likelihood that a predicted and measured object are the same. By analyzing the columns and rows the following situations can be hypothesized: new object, object lost, object match, split situation, and merge situation. In case of for example merge and split situations the matrix can not be resolved directly and ad hoc methods are applied. For example by analyzing the motion vectors and the area (change) of each foreground object [52,70,129,232,395,401,408].

Alternatively, global optimizations can also be applied. Polat et al. [290] use a Multiple Hypothesis Tracker to construct different hypotheses which each explains all the predictions and measurements, and chooses the hypothesis which is most likely. To prune the combinatorial number of different hypotheses smoothness constraints on the motion trajectories are introduced. If the total number of people in the scene is known in advance the pruning becomes less difficult [29,155]. Another global optimization can be seen in [345,421] where a Particle Filter [169] is applied and where each state is a multi-object configuration (hypothesis). Objects are allowed to enter and exit the scene meaning that the number of elements in the state vector can change. To handle this the particle filter is enhanced by a trans-dimensional Markov chain Monte Carlo approach [126]. This allows new objects to enter and other objects to leave the scene, i.e., the dimensionality of the state space may change. In the work by Li et al. [219] a tree-based global optimization for correspondence between multiple objects across multiple views is presented. This approach is used for real-time tracking of hand, head, and feet for whole-body pose estimation. Antonini et al. [19] learn behavioral models for pedestrians' preferences regarding acceleration and direction. These models are used to find globally coherent trajectories.

3.6.2. Temporal correspondences during occlusion

Tracking during occlusion was not addressed in previous work, instead the track of the group was used to update the states of the individuals. However, this makes it impossible to update the models of the individuals, which can result in unreliable tracking after the group splits up. Furthermore, interactions between humans during occlusions is difficult to analyze when they are represented as one foreground object. Therefore, the problem of finding the correspondences during occlusion has been investigated recently.

In some recent systems the first task is to actually detect that an occlusion is present. This can be done using the corresponding matrix mentioned above or as in [52,194,316]. Khan and Shah [194] detect a non-occlusion situation as a situation when the detected foreground objects are far from each other. Capellades et al. [52] define a merge as a situation where the total number of foreground objects has decreased and where two or more foreground objects from the previous frame overlap with one foreground object in the current frame. In the work by Roth et al.

[4] Similar approaches can be used to detect when people are leaving the scene, see e.g., [18,129].

[316] a merge is detected as one of eight different types of occlusion based on the depth ordering and the layout of the bounding boxes. This allows for only using the reliable parts of the bounding box to update the position of the human.

Different approaches for assigning pixels to individuals during occlusion have been reported in recent publications. A local approach is to assign each pixel to the most likely predicted model using a probabilistic method [194,282]. A local approach allows for bypassing the occlusion problem but it is also sensitive to noise and therefore often combined with some post-processing to reassign wrongly classified pixels. Global approaches try to classify pixels based on for example the assumption that people in a group are standing side by side with respect to the camera. This assumption allows for defining vertical dividers between the individuals based on the positions of their heads. Foreground pixels are then assigned to individuals based on these dividers [137,401,406]. When a certain depth ordering is present in the group the assumption fails.

In the work by McKenna et al. [232] the depth ordering is found explicitly. During occlusion the likelihood of each pixel in the foreground object belonging to a person is calculated using Bayes rule. The posteriors for each person are added to obtain an overall probability of each person. These probabilities are then used to define the fraction of each person that is visible. This is denoted a visibility index and can be applied to find the depth ordering. In [316] the depth ordering is based on assuming a planar floor. This will result in the closest object to the camera having the highest vertically coordinate. Xu and Puig [401] generalize this idea by using projective geometry to find the line in the image that corresponds to the "horizon line" in the 3D scene. The object closest to the camera is found as the object closest to this horizontal line.

3.7. Discussion of advances human tracking

Advances in figure-ground segmentation have to a large extent been motivated by the increased focus on surveillance applications. For example, in order to have fully autonomous systems operating in uncontrolled environments the segmentation methods have to be adaptive. This has to some extent been achieved within background subtraction where analysis of video sequences of several hours has been reported [108]. However, for 24 h operation special cameras (and algorithms) are required. Work in this direction has started [66,82] but no one has so far been able to report a truly autonomous system. Furthermore, in most surveillance applications multiple cameras are required to cover the scene of interest at an acceptable resolution. Systems for self-calibrating and tracking across different cameras are being investigated [21,187,193,369], but again, no fully autonomous system has been reported.

Another advance in segmentation is to apply spatial information in the color-based appearance models, for example by dividing each foreground object into a number of regions each having a color representation [155,194,244,269,282,316,404] or by correlograms [52,162]. This has allowed for relatively reliable detection and tracking of people even when multiple people are present with occlusion. Even an accurate appearance model might fail when the lighting changes are significant.

The recent focus on natural scenes has also led to advances within methods for temporal correspondence, especially handling the occlusion problem. Advances are mainly due to the use of probabilistic methods, for example to segment pixels to individuals during occlusion [194,232,282,285] and also to handle multiple hypotheses and uncertainties using stochastic sampling methods [155,269,290,345,404,421]. In fact, concurrent segmentation and tracking can be handled by stochastic sampling methods. It is expected that future work will be based on this framework since it unifies segmentation and tracking *and* the associated uncertainties.

The use of common benchmark data has begun to underpin progress. As has been seen in the speech community for many years and lately in the face recognition community, widely acceptable benchmark data can help to focus research. Within human detection a few recent benchmark data sets have been reported [75,254]. Within tracking in general the PETS and VS-PETS data sets [5] have been applied in many systems.

4. Pose estimation

Pose estimation refers to the process of estimating the configuration of the underlying kinematic or skeletal articulation structure of a person. This process may be an integral part of the tracking process as in model-based analysis-by-synthesis approaches or may be performed directly from observations on a per-frame basis. The previous survey [247] separated pose estimation algorithms into three categories based on their use of a prior human model:

Model-free. This class covers methods where there is no explicit a priori model. Previous methods in this class take a bottom up approach to tracking and labeling of body parts in 2D [394] or direct mapping from 2D sequences of image observations to 3D pose [41].

Indirect model use. In this class methods use an a priori model in pose estimation as a reference or look-up table to guide the interpretation of measured data. Previous examples include human body part labeling using aspect ratios between limbs [49] or pose recognition [136].

Direct model use. This class uses an explicit 3D geometric representation of human shape and kinematic structure to reconstruct pose. The majority of approaches employ an analysis-by-synthesis methodology to optimize the similarity between the model projection and observed images [148,383].

In this section we identify recent contributions and advances in each category of pose estimation algorithms.

A number of trends can be identified from the literature. Three research directions which have each received considerable attention are: the introduction of probabilistic approaches to detect body parts and assemble part configurations in the model-free category; the incorporation of learnt motion models in pose estimation to constrain the recovered 3D human motion; and the use of stochastic sampling techniques in model-based analysis-by-synthesis to improve robustness of 3D pose estimation.

Two important distinctions relating to the difficulty of the pose estimation problem are identified in this analysis: pose estimation from single vs. multiple view images; and 2D pose estimation in the image plane vs. full 3D pose reconstruction. The most difficult and ill-posed problem is the recovery of full 3D pose from single view images towards which initial steps have been made. There has also been substantial research addressing the problems of 2D pose estimation from single view and 3D pose estimation from multiple views. For example recent advances have demonstrated 2D pose estimation in complex natural scenes such as film footage.

4.1. Model free

A recent trend to overcome limitations of tracking over long sequences has been the investigation of direct pose detection on individual image frames. Two approaches have been investigated which fall into this model-free pose estimation category: *probabilistic assemblies of parts* where individual body parts are first detected and then assembled to estimate the 2D pose; and *example-based methods* which directly learn the mapping from 2D image space to 3D model space.

4.1.1. Probabilistic assemblies of parts

Probabilistic assemblies of parts have been introduced for direct bottom-up 2D pose estimation by first detecting likely locations of body parts and then assembling these to obtain the configuration which best matches the observations. A potential advantage of detection over tracking is that the pose can be estimated independently at each frame, allowing pose estimation for rapid movements. Temporal information may be incorporated to estimate consistent pose configurations over sequences. Forsythe and Fleck [110] introduced the notion of 'body plans' to represent people or animals as a structured assembly of parts learnt from images. Following this direction [104,167,168] used pictorial structures to estimate 2D body part configurations from image sequences. Combinations of body part detectors have recently been used to address the related problem of locating multiple people in cluttered scenes with partial occlusion [254,396], see Section 3.

Probabilistic assemblies of body part detectors (face, hands, arms, legs, and torso) have been investigated for bottom up estimation of whole-body 2D pose in individual frames or sequences [235,296,310,314]. Individual body parts are detected using 2D shape [310], SVM classifiers [314], AdaBoost [235], and locally initialized appearance models [296]. Mikolajczyk et al. [240] introduced probabilistic assemblies of robust AdaBoost body part detectors to locate people in images providing a coarse 2D localization. The probabilistic assembly of parts models the joint likelihood of a body part configuration. In [235] this approach is extended to whole-body 2D pose estimation in frontal images using RANSAC to assemble body part configurations with prior pose constraints. Ramanan et al. [296] present a related approach where lateral views of a 'scissor-leg' pose for a person walking or running are detected from film footage. Detected poses are then used as keyframes to initialize a local appearance model for body part detection and 2D pose estimation at intermediate frames.

Recent work has also introduced approaches for 2D pose estimation from single images. Ren et al. [302] use pairwise constraints between body parts to assemble body part detections into 2D pose configurations. Ramanan et al. [297] learn a global body part configuration model based on conditional random fields to simultaneously detect all body parts. Pairwise constraints include aspect ratio, scale, appearance, orientation, and connectivity. Hua et al. [158] present an approach to 2D pose estimation from a single image using bottom-up feature cues together with a Markov network to model part configurations. Both of these approaches demonstrate impressive results for pose estimation in cluttered scenes such as sports images.

An important contribution of approaches based on the probabilistic assembly of parts is 2D pose estimation in cluttered natural scenes from a single view. This overcomes limitations of many previous pose estimation methods which require structured scenes, accurate prior models or multiple views.

4.1.2. Example-based methods

A number of example-based methods for human pose estimation have been proposed which compare the observed image with a database of samples. Brand [41] used a hidden Markov model (HMM) to represent the mapping from 2D silhouette sequences in image space to skeletal motion in 3D pose space. In this work the mapping for specific motion sequences was learnt using rendered silhouette images of a humanoid model. The HMM was used to estimate the most likely 3D pose sequence from an observed 2D silhouette sequence for a specific view. Similarly, Rosales et al. [294,315] learn a mapping from visual features of a segmented person to static pose using neural networks. This representation allows 3D pose estimation invariant to speed and direction of movement. Viewpoint invariant representation of the mapping from image to pose is investigated in [272].

To overcome limitations of tracking researchers have investigated example-based approaches which directly lookup the mapping from silhouettes to 3D pose [6,152,326,340]. Howe [152] uses a direct silhouette lookup using Chamfer distance to select candidate poses together with a Markov chain for temporal propagation for 3D pose estimation of walking and dancing. Shakhnarovich et al. [326] present an example-based approach for viewpoint invariant pose estimation of upper-body 3D pose from a single image. Parameter-sensitive hashing is used to represent the mapping between observed segmented images from multiple views and the corresponding 3D pose. Grauman et al. [125] learn a probabilistic representation of the mapping from multiple view silhouette contours to whole-body 3D joint locations. Pose reconstruction is demonstrated for close-up images of a walking person from multiple or single views. Similarly, Elgammal and Lee [98] learn multiple view-dependent mapping from silhouettes to 3D pose for walking actions. Agarwal and Triggs [6,8] presented an example-based approach for 3D pose estimation from single view image sequences. Nonlinear regression is used to learn the mapping from silhouette shape descriptors to 3D pose. Results demonstrate reconstruction of long sequences of walking motions with turns from monocular video.

Example-based approaches represent the mapping between image and pose space providing a powerful mechanism for directly estimating 3D pose. Commonly these approaches exploit rendering of motion capture data to provide training examples with known 3D pose. A limitation of current example-based approaches is the restriction to the poses or motions used in training. Extension to a wider vocabulary of movements may introduce ambiguities in the mapping.

4.2. Indirect model use

A number of researchers have investigated direct reconstruction of both model shape and motion from the visual-hull [59,237,238] without a prior model. Mikic et al. [237,238] present an integrated system for automated recovery of both a human body model and motion from multiple view image sequences. Model acquisition is based on a hierarchical rule-based approach to body part localization and labelling. Prior knowledge of body part shape, relative size, and configuration is used to segment the visual-hull. An extended Kalman filter is then used for human motion reconstruction between frames. A voxel labelling procedure is used to allow large inter-frame movements. Cheung et al. [59] first reconstruct a model of the kinematic structure, shape, and appearance of a person and then use this to estimate the 3D movement. Tracking is performed by hierarchically matching the approximate body model to the visual-hull using color matching along the silhouette boundary edge.

An alternative approach based on full 3D-to-3D non-rigid surface matching using spherical mapping is presented in [353]. Alignment of a skeletal model with the first frame allows the 3D motion to be recovered from the non-rigid surface motion. Results of these approaches demonstrate 3D human pose estimation for rapid movement of subjects wearing tight clothing.

These approaches exploit scene reconstruction from multiple views to directly recover both shape and motion. This approach is suitable for multiple camera studio-based systems allowing estimation of complex human movements.

4.3. Direct model use

The use of an explicit model of a person's kinematics, shape, and appearance in an analysis-by-synthesis framework is the most widely investigated approach to human pose estimation from video. In the previous survey [247] fifty papers (40% of those surveyed) were in this category starting with some of the earliest work in human pose estimation [148]. Model-based analysis-by-synthesis has continued to be a dominant methodology for human pose estimation.

The main novel research directions are: the introduction of stochastic sampling techniques based on sequential Monte Carlo; and the introduction of constraints on the model in particular learnt models of human motion. In this section we review key papers contributing to these advances in multiple and single view model-based pose estimation.

4.3.1. Multiple view 3D pose estimation

Up to 2000 the majority of approaches to human pose estimation employed deterministic gradient descent techniques to iteratively estimate changes in pose [86,289]. The extended Kalman filter was widely applied to human tracking with low-order dynamics used to predict change in pose [384]. Recent work using model-based analysis-by-synthesis has extended deterministic gradient descent-based approach to more complex motions. For example Plänkers and Fua [289] demonstrated upper body tracking of arm movements with self-occlusion using stereo and silhouette cues. A common limitation of gradient descent approaches is the use of a single pose or state estimate which is updated at each time step. In practice if there is a rapid movement or visual ambiguities pose estimation may fail catastrophically. To achieve more robust tracking, techniques which employ a deterministic or stochastic search of the pose state space have been investigated.

Stochastic tracking techniques, such as the *particle filter*, were introduced for robust visual tracking of objects where sudden changes in movement or cluttered scenes can result in failure. The principal difficulty with their application to human pose estimation is the dimensionality of the state space. The number of samples or particles required increases exponentially with dimensionality. Typically whole-body human models use more than 20 degrees-of-freedom making direct application of particle filters computationally prohibitive. MacCormick and Isard [230] proposed partitioned sampling of the state space for efficient 2D pose estimation of articulated objects such as the hand. However, this approach does not extend directly to the dimensionality required for whole-body pose estimation. Deutscher et al. [90] introduced the *annealed particle filter* which combines a deterministic annealing approach with stochastic sampling to reduce the number of samples required. At each time step the particle set is refined through a series of annealing cycles with decreasing temperature to approximate the local maxima in the fitness function. Results [85,90] demonstrate reconstruction of complex motion such as a hand-stand. A hierarchal stochastic sampling scheme to efficiently estimate the 3D pose for complex movements or multiple people is presented in [242]. This approach initially estimates the torso pose for each person and propagates samples with high fitness to estimate the pose of adjacent body parts.

Recent work has combined deterministic or stochastic search with gradient descent for local pose refinement to recover complex whole-body motion. Carranza et al. [53] demonstrate whole-body human motion estimation from multiple views combining a deterministic grid search with gradient descent. Pose estimation is performed hierarchically starting with the torso. For each body part a grid search first finds the set of valid poses for which the joint positions project inside the observed silhouettes. A fitness function is then evaluated for all valid poses to determine the best pose estimate. Finally gradient descent optimization is performed to refine the estimated pose. This search procedure is made feasible by the use of graphics hardware to evaluate the fitness function which is based on the overlap between the projected model and observed silhouette across all views. In related work Kehl et al. [191] propose *stochastic meta descent* for whole-body pose estimation with 24 degrees-of-freedom from multiple views. Stochastic meta descent combines a stochastic sampling of the set of model points used at each iteration of a gradient descent algorithm. This introduces a stochastic search element to the optimization which allows the approach to avoid convergence to local minima. The use of a small number of samples (5) per body part together with adaptive step size allows efficient performance. Results of these approaches demonstrate reconstruction of complex movements such as kicking and dancing.

In summary, the introduction of stochastic sampling and search techniques has achieved whole-body pose estimation of complex movements from multiple views. Current approaches are limited to gross-body pose estimation of torso, arms, and legs and do not capture detailed movement such as hand-orientation or axial arm rotation. Multiple hypothesis sampling achieves robust tracking but does not provide a single temporally consistent motion estimate resulting in jitter which must be smoothed to obtain visually acceptable results. There remains a substantial gulf between the accuracy of commercial marker-based and markerless video-based human motion reconstruction.

4.3.2. Monocular 3D pose estimation

Reconstruction of human pose from a single view image sequence is considerably more difficult than either the problem of 2D pose estimation or 3D pose estimation from multiple views. To resolve the inherent ambiguity in monocular human motion reconstruction additional constraints on kinematics and movement are typically employed [43,384]. Wachter and Nagel [384] used the extended Kalman filter together with kinematic joint constraints to estimate the 3D motion of a person walking parallel to the image plane. As discussed in the previous section the use of a single hypothesis tracking scheme is prone to failure for complex motions. Loy et al. [225] employ a manual key-frame approach to 3D pose estimation of complex motion in sports sequences.

Sminchisescu and Triggs [343] have investigated the application of stochastic sampling to estimation of 3D pose from monocular image sequences. They observe that alternative 3D poses which give good correspondence to the observations are most likely to occur in the direction of greatest uncertainty. This motivated the introduction of *covariance scaled sampling* an extension of particle filters which increases the covariance in the direction of maximum uncertainty by approximately an order of magnitude to increases the probability of generating samples close to local minima in the fitness function. Samples are then optimized to find the local minima using a gradient descent approach. Results demonstrate monocular tracking and 3D reconstruction of human movements with moderate complexity including walking with changes in direction. Further research [344] has explicitly enumerated the potential kinematic minima which cause visual ambiguities. Incorporating this in the sampling process increases efficiency and robustness allowing reconstruction of more complex human motion from monocular video sequences.

Probabilistic approaches using assemblies of parts together with higher level knowledge of human kinematics and shape have also been investigated for single view 3D pose estimation. Lee and Cohen [211] combine a probabilistic proposal map representing the estimated likelihood of body parts in different 3D locations with an explicit 3D model to recover the 3D pose from single image frames. A data driven Markov chain Monte Carlo (MCMC) is used to search the high-dimensional pose space. The proposal map for each body part represents the likelihood of the projected 3D pose. Proposal distributions are used to efficiently sample the pose space during MCMC search. Results demonstrate 3D pose estimation from single images of sports players in a variety of complex poses. Moeslund and Granum [246,252] apply a data driven sequential Monte Carlo approach to pose estimation of a human arm. A part detector provides likely locations of the hand in the image and their uncertainties. This information is applied to correct the prediction lowering the number of particles required.

Navaratnam et al. [265] combine a hierarchical kinematic model with a bottom up part detection to recover the 3D upper-body pose. The use of part detection allows individual body parts to be independently located at each frame. Kinematic constraints between body parts are represented hierarchically to recover the 3D pose from a single view. Unlike previous model free probabilistic assembly of parts this approach enables recovery of full 3D pose at each frame. Temporal information is also integrated using a HMM framework to reconstruct temporally coherent movement sequences.

Monocular reconstruction of complex 3D human movement remains an open problem. Recent research has investigated the use of learnt motion models to provide strong priors to constrain the search.

4.3.3. Learnt motion models

There has been increasing interest in the use of learnt models of human pose and motion to constrain vision-based reconstruction of human movement from single or multiple views. The availability of marker-based human motion capture data [1,2,4] has led to the use of learnt models of human motion for both animation synthesis in computer graphics and vision-based human motion synthesis.

Learnt models have been developed in computer animation to allow synthesis of natural motions with user specified constraints from a motion capture database [20,199,209,255]. This use of learnt models in computer graphics is relevant to the problem of vision-based reconstruction of human movement in developing methods to predict and constrain human pose and motion estimation. Inverse kinematics of human motion based on learnt models has recently been introduced in computer graphics [128,271]. Ong et al. [271] use a learnt model of whole-body configurations to constrain the pose given a set of end effector positions for a motion sequence. Grochow et al. [128] use scaled Guassian process latent variable models (SGPLVM) to model the probability distribution over all possible whole-body poses to constrain both character pose in animation and pose reconstruction from images.

Sidenbladh et al. [332,334,335] combine stochastic sampling with a strong learned prior of walking motion for tracking. An exemplar-based approach is used in [335] similar to work in motion synthesis [20,199,255] where a database of motion capture examples is indexed to obtain possible movement directions. Statistical priors on human appearance and image motion are used [333] to model the likelihood of observing various image cues for a given movement. These are incorporated in an analysis-by-synthesis approach to human motion reconstruction. Similarly, a hierarchical PCA model of human dynamics learnt from motion capture using a Gaussian mixture and HMM to represent dynamics is proposed for monocular tracking in [189]. Agarwal and Triggs [7] use a learned model of local second order dynamics for 2D tracking of more general motions walking and running with transitions and turns in monocular image sequences. Their work demonstrates that strong priors on human dynamics allows 2D pose estimation for fast movements in cluttered scenes.

Subsequent research has investigated the use of learnt motion models for 3D motion reconstruction primarily from monocular image sequences to overcome the inherent visual ambiguity. In [154] learnt models from short motion sequences are used to infer 3D pose from tracked image features of simple movements. Sigal et al. [336] combine body part detectors with a learned motion model to infer 3D human pose from monocular images of walking with automatic initialization. Their approach uses belief propagation via stochastic sampling over a loopy graph of loosely attached body parts. Urtasun and Fua [373] introduce the use of temporal motion models learnt from sequences of motion capture data to reconstruct human motion using a deterministic gradient descent optimization. Principal component analysis (PCA) is performed on multiple examples of concatenated joint angle sequences for walking and running to provide a low-dimensional parametrization. The parametric motion model is then used to constrain the movement of a 3D humanoid model for walking and running movements with variable speed from stereo [373] and golf swings from a single view [370]. Urtasun et al. [372] advocate an alternative approach to representation of human motion using SGPLVM to learn a low-dimensional embedding of the pose state space for specific movements such as golf-swings or walking from monocular image sequences. Gaussian process models which incorporate dynamics [259,371] have been introduced to ensure continuous embedding of motion in the latent space for robust tracking. Further research following the methodology of using learnt motion models has addressed the problem of viewpoint invariance in tracking human movement [8,272].

Research introducing the use of learnt statistical models of human motion since 2000 has demonstrated that using strong motion priors facilitates reconstruction of 3D pose sequences from monocular images. To date the generality of these approaches has been limited to specific motion models with relatively small variation in motion and fixed transitions. A challenge for future research is to build more general motion models or methods of transitioning between models, to allow the reconstruction of unconstrained human movement.

4.4. Discussion of advances in human pose estimation

As identified in this section research in automatic estimation of human pose has been an active area over the past 5 years with significant advances being made. A number of novel methodologies have been proposed towards the objective of human pose estimation from monocular image sequences in natural scenes. The introduction of methods based on 2D pose estimation as a probabilistic assembly of parts have achieved significant advances for cluttered natural scenes such as film footage or sports [104,158,168,235,296,302,310,314]. These approaches are based on detection of body parts such as the face, hands or limbs independently for each image frame.

Similarly there have been significant advances in the use of example-based methods to learn the mapping from 2D image features such as silhouettes to 3D pose [6,41,152,326,340]. These methods commonly exploit databases of human motion capture data to render images of a model in multiple poses providing known 2D image to 3D pose correspondence. Currently example-based methods are limited to the fixed classes of movement and range of viewpoints used in training. A future challenge is to extend these methods to viewpoint invariant 3D pose estimation for general movement. There is also the possibility of combining learnt 2D to 3D mappings with 2D pose detection to achieve 3D pose detection in cluttered scenes from monocular image sequences or single image frames.

Model-based pose estimation using an analysis-by-synthesis methodology to estimate 3D pose from multiple view images has focused on reliable recovery of complex movements [53,90,191]. Significant advances in the complexity of movement that can be reconstructed have been achieved through the use of stochastic sampling and search techniques in pose estimation from multiple views. Similarly research in 3D pose estimation from monocular image sequences using stochastic sampling [343] has achieved reconstruction in cluttered scenes. Monocular reconstruction of complex 3D human movement remains an open problem. Learnt models of human motion have been applied extensively to constrain the monocular reconstruction problem by providing strong priors on motion [7,334,336,372]. Currently learnt motion models are limited to specific classes of motion. The extension of learnt models to reconstruction of general human movement remains an open problem.

Over the past 5 years there have been significant advances in the range of human motion which can be reconstructed from either monocular or multiple view image sequences. A limitation of existing research which should be addressed in future is the comparison of different approaches on common data sets and performance evaluation of accuracy against ground-truth.

5. Recognition

The field of action and activity representation and recognition is relatively old, yet still immature. This area is presently subject to intense investigation which is also reflected by the large number of different ideas and approaches. The approaches depend on the goal of the researcher and applications for activity recognition are interesting for surveillance, medical studies and rehabilitation, robotics, video indexing, and animation for film and games. For example, in scene interpretation the knowledge is often represented statistically and is meant to distinguish "regular" from "irregular" activities.

In scene interpretation, the representations should be independent from the objects causing the activity and thus are usually not meant to distinguish explicitly, e.g., cars from humans. On the other hand, some surveillance applications focus explicitly on human activities and the interactions between humans. Here, one finds both, holistic approaches, that take into account the entire human body without considering particular body parts, and local approaches. Most holistic approaches attempt to identify "holistic" information such as gender, identity, or simple actions like walking or running. Researchers using local approaches appear often to be interested in more subtle actions or attempt to model actions by looking for action primitives with which the complex actions can be modeled.

We have structured this review according to a visual abstraction hierarchy yielding the following: *scene interpretation* where the entire image is interpreted without identifying particular objects or humans, *holistic recognition* where either the entire human body or individual body parts are applied for recognition, and *action primitives and grammers* where an action hierarchy gives rise to a semantic description of a scene. Before going into these topics we first look closer at the definition of the action hierarchy used in this survey since it has influence on the remaining categories.

5.1. Action hierarchies

Terms like *actions*, *activities*, *complex actions*, *simple actions*, and *behaviors* are often used interchangingly by the different authors. However, in order to be able to describe and compare the different publications we see the need for a common terminology. In a pioneering work [264], Nagel suggested to use a hierarchy of *change*, *event*, *verb*, *episode*, and *history*. An alternative hierarchy (reflecting the computational aspects) is proposed by Bobick [37] who suggests to use *movement, activity*, and *action* as different levels of abstraction (see also [12]). Others suggest to also include *situations* [121] or use a hierarchy of *Action primitives* and *Parent Behaviors* [175].

In this survey we will use the following action hierarchy: *action/motor primitives*, *actions*, and *activities*. *Action primitives* or *motor primitives* will be used for atomic entities out of which actions are built. *Actions* are, in turn, decomposed into *activities*. The granularity of the primitives often depends on the application. For example, in robotics, *motor primitives* are often understood as sets of motor control commands that are used to generate an action by the robot (see Section 5.5).

As an example, in tennis *action primitives* could be, e.g., "forehand", "backhand", "run left", and "run right". The term *action* is used for a sequence of action primitives needed to return a ball. The choice of a particular action depends on whether a forehand, backhand, lob, or volley, etc., is required in order to be able to return the ball successfully. Most of the research discussed below falls into this category. The *activity* then is in this example "playing tennis". *Activities* are larger scale events that typically depend on the context of the environment, objects, or interacting humans.

A good overview of activity recognition is given by Aggarwal and Park [12]. They aim at higher-level understanding of activities and interactions and discuss different aspect such as level of detail, different human models, recognition approaches and high-level recognition schemes. Veeraraghavan et al. [379] discuss the structure of an action and activity space.

5.2. Scene interpretation

Many approaches consider the camera view as a whole and attempt to learn and recognize activities simply by observing the motion of objects without necessarily knowing their identity. This is reasonable in situations where the objects are small enough to be represented as points on a 2D plane.

Stauffer et al. [355] present a full scene interpretation system which allows detection of unusual situations. The system extracts features such as 2D position and speed, size and binary silhouettes. Vector quantization is applied to generate a codebook of K prototypes. Instead of taking the explicit temporal relationship between the symbols into account, Stauffer and Grimson use co-occurrence statistics. Then, they define a binary tree structure by recursively defining two probability mass functions across the prototypes of the code book that best explain the co-occurrence matrix. The leaf nodes of the binary tree are probability distributions of co-occurrences across the prototypes and at a higher tree depth define simple scene activities like pedestrian and car movement. These can then be used for scene interpretation. In Eng et al. [101] a swimming pool surveillance system is presented. From each of the detected and tracked objects features such as speed, posture, submersion index, an activity index, and a splash index, are extracted. These features are fed into a multivariate polynomial network in order to detect water crisis events. Boiman and Irani [39] approach the problem of detection irregularities in a scene as a problem of composing newly observed data using spatio-temporal patches extracted from previously seen visual examples. They extract small image and video patches which are used as local descriptors. In an inference process, they search for patches with a similar geometric configuration and appearance properties, while allowing for small local misalignments in their relative geometric arrangement. This way, they are able to quickly and efficiently infer subtle but important local changes in behavior. Junejo et al. [182] describe an approach to focusses on dynamic information for scene interpretation. Their method can distinguish between objects traversing spatially dissimilar paths or objects traversing spatially proximal paths but with different spatio-temporal characteristics. For this, they learn the paths in a training phase where graph-cuts are used for clustering the trajectories. For matching, they use spatial similarity, velocity characteristics and curvature features.

In [64,375] activity trajectories are modeled using non-rigid shapes and a dynamic model that characterizes the variations in the shape structure. Vaswani et al. [375] uses Kendall's statistical shape theory [192]. Nonlinear dynamical models are used to characterize the shape variation over time. An activity is recognized if it agrees with the learned parameters of the shape and associated dynamics. Chowdhury et al. [63] use a subspace method to model activities as a linear combination of 3D basis shapes. The work is based on the factorization theorem [365]. Deviations from the learned normal activity shapes can be used to identify abnormal ones.

A similar complex task is approached by Xiang and Gong [400]. They present a unified bottom-up and top-down approach to model complex activities of multiple objects in cluttered scenes. Their approach is object-independent and they use a dynamically multi-linked hidden Markov models (HMMs) on conjunction with Schwarz's Bayesian information criterion [324] to interlink between multiple temporal processes corresponding to multiple event classes. Liu and Chua [223] present an HMM-based approach for recognizing multi-agent activities.

5.3. Holistic recognition approaches

The recognition of the identity of a human, based on his/her global body structure and the global body dynamics is discussed in many publications. Of particular interest for identity recognition has been the human gait. Other approaches using global body structure and dynamics are concerned with the recognition of simple actions such as running and walking. Almost all methods are silhouette or contour-based. Subsequent techniques are mostly holistic, e.g., the entire silhouette or contour is being taken into account without detecting individual body parts.

5.3.1. Human body-based recognition of identity

In Wang et al. [390] the silhouette of a human is computed and then unwrapped by evenly sampling the contour. Next, the distance between each contour point and its center of gravity is computed. The unwrapped contour is then processed by PCA. To compute the spatio-temporal correlation they compare trajectories in eigenspace by first applying appropriate time warping to minimize the distance between the probe and the gallery trajectories. On outdoor data and in spite of its simplicity, it gives good results while being computationally efficient. BenAbdelkader et al. [32] use a variation of co-occurrence techniques. After applying a suitable time-warping and normalization with respect to scale a self-similarity plot is computed where silhouette images of the sequences are pairwise correlated. PCA is applied to reduce the dimensionality of these plots and a k-nearest neighbor classifier is applied in eigenspace for recognition.

Foster et al. [111] extract, embox, and normalize silhouettes. Then, a set of binary masks are defined and the area of the silhouette within the mask is computed to give a dynamic signature of the observed person for each mask. A frame rate of 30 fps results in a 30D vector for each signature giving a $n \times 30$ matrix, where n denotes the number of area masks used. To remove the information about the static shape of the silhouette, the average value of each signature can be subtracted. Fisher analysis is applied and the k-nearest neighbor classifier is used for classification. Kale et al. [183,184] define a HMM to model the dynamics of individual gait. A HMM is trained for each individual in the database. Five representative binary silhouette are used as the hidden states for which transition probabilities and observation likelihoods are trained. During the recognition phase, the HMM with the largest probability identifies the individual. Yam et al. [402] investigate the relationship between walking and running. They define a gait signature based on a frequency analysis of thigh and lower leg rotations. Phase and magnitude of the Fourier descriptions are multiplied to give the phase-weighted magnitude (PWM). It appears that the signatures for walking and running for an individual is related by a phase modulation. The additional individual relationship between walking and running is used to derive improved gait-recognition which can recognize both, walking and running patterns.

5.3.2. Human body-based recognition

While a large number of papers recognize individuals based on their dynamics, the dynamics can also be used to recognize *what* the individual is doing. The approaches discussed in this subsection are again based on holistic body information where no attempt is made to identify individual body parts.

A pioneering work in this context has been presented by Efros et al. [94]. They attempt to recognize simple actions of people whose images in the video are only 30 pixels tall and where the video quality is poor. They use a set of features that are based on blurred optic flow (blurred motion channels). First, the person is tracked so that the image is stabilized in the middle of a tracking window. The blurred motion channels are computed on the residual motion that is due to the motion of the body parts. Spatio-temporal cross-correlation is used for matching with a database. Roh et al. [312] base their action recognition task on curvature scale space templates of a player's silhouette.

Of further interest is the enhancement where complex actions can be dynamically composed out of the set of simple actions. Robertson and Reid [311] attempt to *understand* actions by building a hierarchical system that is based on reasoning with belief networks and HMMs on the highest level and on the lowest level with features such as position and velocity as action descriptors. Their action descriptor is based on the work by Efros et al. [94]. The system is able to output qualitative information such as *walking—left-to-right—on the sidewalk*.

A large number of publications work with space-time volumes. One of the main approaches is to use spatio-temporal *XT*-slices from an image volume *XYT* [304,305] where articulated motions of a human can be associated with a typical trajectory pattern. Ricquebourg and Bouthemy [304] demonstrate how *XT*-slices can facilitate tracking and reconstruction of 2D motion trajectories. The reconstructed trajectory allows a simple classification between pedestrians and vehicles. Ritscher et al. [305] discuss the recognition in more detail by a closer investigation of the *XT*-slices. Quantifying the braided pattern in the slices of the spatio-temporal cube gives rise to a set of features (one for each slice) and their distribution is used to classify the actions.

Bobick and Davis pioneered the idea of temporal templates [37,38]. They propose a representation and recognition theory [37,38] that is based on *motion energy images* (MEI) and *motion history images* (MHI). The MEI is a binary cumulative motion image. The MHI is an enhancement of the MEI where the pixel intensities are a function of the motion history at that pixel. Matching temporal templates is based on Hu moments. Bradski et al. [40] pick up the idea of MHI and develop timed MHI (tMHI) for motion segmentation. tMHI allow determination of the normal optical flow. Motion is segmented relative to object boundaries and the motion orientation. Hu moments are applied to the binary silhouette to recognize the pose. A work conceptually related to [38] is by Masound and Papanikolopoulos [231]. Here, motion information for each video frame is represented by a feature image. However, unlike [38], an action is represented by several feature images. PCA is applied for dimensionality reduction and each action is then represented by a manifold in PCA space.

Yi et al. [409] present the idea of a pixel change ratio map (PCRM) which is conceptually similar to the MHI. However, further processing is based on motion histograms which are computed from the PCRM. Weinberg et al. [393] suggest replacing the motion history image by a 4D motion history volume. For this, they first compute the visual hull from multiple cameras. Then, they consider the variations around the central vertical axes and use cylindric coordinates to compute alignments and comparisons. Motion history images can also be used to detect and interpret actions in compressed video data. Babu and Ramakrishnan [23] compute a flow history (MFH) from the motion data available in compressed video. In addition to MFH, they also use motion history images to classify activities.

As the search of activities in large databases gains importance, a full, hierarchical human detection system is presented by Ozer and Wolf [275]. They approach the tracking, pose estimation and action recognition problem in an integrated manner. They apply a number of well-known techniques on (un)compressed video data.

Another approach is that of "Actions Sketches" or "Space-Time Shapes" in the 3D *XYT* volume. Yilmaz and Shah [410] propose to use spatio-temporal volumes (STV) for action recognition: the 3D contour of a person gives rise to a 2D projection. Considering this projection over time defines the STV. Yilmaz and Shah extract information such as speed, direction and shape by analyzing the differential geometric properties of the STV. They approach action recognition as an object matching task by interpreting the STV as rigid 3D objects. Blank et al. [36] also analyze the STV. They generalize techniques for the analysis of 2D shapes [123] for the use on the STV. Blank et al. argue that the time domain introduces properties that do not exist in the *xy*-domain and needs thus a different treatment. For the analysis of the STV they utilize properties of the solution of the Poisson equation [123].

This gives rise to local and global descriptors that are used for recognizing simple actions.

Instead of using spatio-temporal volumes, a large number of papers choose the more classical approach of considering sequences of silhouettes. Yu et al. [412] extract silhouettes and their contours are unwrapped and processed by PCA. A three-layer feed forward network is used to distinguish "walking", "running"and "other" based on the trajectories in eigenspace. In another PCA-based approach, Rahman and Robles-Kelly [295] suggest to use a tuned eigenspace technique. Their tuned eigenspaces allow to treat the action problem as a nearest-neighborhood problem in eigenspace. Jiang et al. [179] attempt to match a given sequence of poses to a novel video. They treat this problem as an optimal matching problem by changing the usually highly non-convex problem into a convex one.

Elgammal and Lee [13] use optic flow in addition to the shape features and a HMM is used to model the dynamics. In [98,97], Elgammal and Lee use local linear embedding (LLE) [317,362] in order to find a linear embedding of human silhouettes. In conjunction with a generalized radial basis function interpolation, they are able to separate style and content of the performed actions [97] as well as to infer 3D body pose from 2D silhouettes [98]. Sato and Aggarwal [321] are concerned with the detection of interaction between two individuals. This is done by grouping foreground pixels according to similar velocities. A subsequent tracker tracks the velocity blobs. The distance between two people, the slope of relative distance and the slope of each person's position are the features used for interaction detection and classification. In Cheng et al. [58], walking is distinguished from running based on sport event video data. The data comes from real-life programs. They compute a dense motion field and foreground segmentation is performed based on color and motion. Within the foreground region, the mean motion magnitude between frames is computed over time followed by an analysis in frequency space to compute a characteristic frequency. A Gaussian classifier is used for classification. Gao et al. [113] consider a smart room application. A dining room activity analysis is performed by combining motion segmentation with tracking. They use motion segmentation based on optical flow and RANSAC. Then, they combine the motion segmentation with a tracking approach which is sensitive to subtle motion. In order to identify activities, they identify predominant directions of relative movements.

In a number of publications, recognition is based on HMMs and dynamic Bayes networks (DBNs). Elgammal et al. [99] propose a variant of semi-continuous HMMs for learning gesture dynamics. They represent the observation function of the HMM as non-parametric distributions to be able to relate a large number of exemplars to a small set of states. Luo et al. [227] present a scheme for video analysis and interpretation where the higher-level knowledge and the spatio-temporal semantics of objects are encoded with DBNs. The DBNs are based on key-frames and are defined for video objects. Shi et al. [330] present an approach for semi-supervised learning of the HMM or DBN states to incorporate prior knowledge. Leo et al. [217] attempt to classify actions at an archaeological site. They present a system that uses binary patches and an unsupervised clustering algorithm to detect human body postures. A discrete HMM is used to classify the sequences of poses into a set of four different actions.

Smith et al. [347] suggest to use multiple levels of zoom for activity analysis to combine both detailed and coarse views of a scene. They find feature correspondencies across different zoom levels using epipolar, spatial, trajectory, and appearance constraints.

A totally different approach is presented by Wang et al. [392] where the aim is at classifying actions in *still* images. Unsupervised learning is used to generate action classes out of a large training set. These action classes are then used to label test images. The approach uses a technique for deformable matching of edges of image pairs, based on linear programming relaxation techniques.

5.4. Recognition based on body parts

Many authors are concerned with the recognition of actions based on the dynamics and settings of individual body parts. Some approaches, e.g., [83], start out with silhouettes and detect the body parts using a method inspired by the W4-system [138]. Others use 3D-model based body tracking approaches (see Section 4) where the recognition of (often periodic) action is used as a loop-back to support pose estimation. Other approaches circumvent the vision problem by using a motion capture system in order to be able to focus on the action issues [81,277,279].

In a work related to [390], Wang et al. [389] present an approach where contours are extracted and a mean contour is computed to represent the static contour information. Dynamic information is extracted by using a detailed model composed of 14 rigid body parts, each one represented by a truncated cone. Particle filtering is used to compute the likelihood of a pose given an input image. For classification, a nearest neighbor classifier (NN) was used.

Davis and Taylor [83] present an approach to distinguish walking from non-walking. A method based on the W4-system is used to detect body parts from silhouettes. Based on the feet locations four motion properties are extracted of which three (cycle time, stance/swing ratio, and double support time) reflect dynamic features and one (extension angle) reflects a structural feature. The walking category is defined by three pairs of the dynamic features and the structural feature. In a similar approach Ren and Xu [300] use as input a binary silhouette from which they detect the head, torso, hands, and elbow angles. Then, a primitive-based coupled HMM is used to recognize natural complex and predefined actions. They extend their work in [301] by introducing primitive-based DBNs. Parameswaran and Chellappa [277,279] consider the problem of view-invariant action recognition based on point-light displays by investigating 2D and 3D invariant theory. As no general, non-trivial 3D–2D invariants exist, Parameswaran and Chellappa employ a convenient 2D invariant representation by decomposing and combining the patches of a 3D scene. For example, key poses can be identifies where joints in the different poses are aligned. In the 3D case, six-tuples corresponding to six joints give rise to 3D invariant values and it is suggested to use the progression of these invariants over time for action representation. A similar issue is discussed in the work by Yilmaz and Shah [411] where joint trajectories from several uncalibrated moving cameras are considered. They propose an extension to the standard epipolar geometry-based approach by introducing a temporal fundamental matrix that models the effects of the camera motion. The recognition problem is then approached in terms of the quality of the recovered scene geometry. Gritai et al. [127] address the invariant recognition of human actions, and investigate the use of anthropometry to provide constraints on matching. Gritai et al. use the constraints to measure the similarity between poses and pose sequences. Their work is based on a point-light display like representation where a pose is presented through a set of points in 3D space. Sheikh et al. [328] pick up these results of [127,411] and discuss that the three most important sources of variability in the task of recognizing actions come from variations in viewpoint, execution rate, and anthropometry of the actors. Then, they argue that the variability associated with the execution of an action can be closely approximated by a linear combination of action bases in joint spatio-temporal space. Davis' and Gao's [79,81] aim is to recognize properties from visual target cues, e.g., the sex of an individual or the weight of a carried object is estimated from how the individuals move. Davis and Gao [81] recognize the gender of a person based on the gait. Labeled 2D trajectories from motion capture devices of humans are factored using three-mode PCA into components interpreted as *posture*, *time*, and *gender*. An importance weight for each of the trajectories is learned automatically. Davis et al. [79] use the three-mode PCA framework to recognize human action efforts. Here, the three modes *pose*, *time*, and *effort* are used. In order to detect particular body parts Fanti et al. [103] give the structure of a human as model knowledge. To find the most likely model alignment with input data they exploit appearance information which remains approximately invariant within the same setting. Expectation maximization is used for unsupervised learning of the parameters and structure of the model for a particular action and unlabeled input data. Action is then recognized by maximum likelihood estimation. Ning et al. [267] use a parabola to model the shoulders of a human. Fisher discriminant analysis (FDA) on the parabola parameters are used to detect shrugs.

5.5. Action primitives and grammars

There is strong neurobiological evidence that human actions and activities are directly connected to the motor control of the human body [117,307,308]. When viewing other agents performing an action, the human visual system seems to relate the visual input to a sequence of motor primitives. The neurobiological representation for visually perceived, learned, and recognized actions appears to be the same as the one used to drive the motor control of the body. These findings have gained considerable attention from the robotics community [77,322]. In *imitation learning* the goal is to develop a robot system that is able to relate perceived actions to its own motor control in order to learn and to later recognize and perform the demonstrated actions. Consequently, it is ongoing research to identify a set of motor primitives that allow (a) representation of the visually perceived action and (b) motor control for imitation. In addition, this gives rise to the idea of interpreting and recognizing activities in a video scene through a hierarchy of primitives, simple actions and activities. Most of the following researchers attempt to learn the motor or action primitives by defining a "suitable" representation and then learning the primitives from demonstrations. The representations used to describe the primitives vary a lot across the literature and are subject to ongoing research. Most of the subsequently mentioned work is based on motion capture data.

Jenkins et al. [176,177] suggest to apply a spatio-temporal non-linear dimension reduction technique on manually segmented human motion capture data. Similar segments are clustered into primitive units which are generalized into parameterized primitives by interpolating between them. In the same manner, they define action units ("behavior units") which can be generalized into actions. Ijspeert et al. [165] approach the problem of defining motor primitives from the motor side. They define a set of nonlinear differential equations that form a control policy (CP) and quantify how well different trajectories can be fitted with these CPs. The parameters of a CP for a primitive movement are learned in a training phase. These parameters are also used to compute similarities between movements. Billard and Calinon [34,50,51] use an HMM-based approach to learn characteristic features of repetitively demonstrated movements. They suggest to use the HMM to synthesize joint trajectories of a robot. For each joint, one HMM is used. Calinon et al. [51] use an additional HMM to model end-effector movement. In these approaches, the HMM structure is heavily constrained to assure convergence to a model that can be used for synthesizing joint trajectories.

A number of publications attempt to decouple actions into action primitives and to interpret actions as a composition on the alphabet of these action primitives, however, without the constraints of having to drive a motor controller with the same representation. Vecchio and Perona [376] employ techniques from the dynamical systems framework to approach segmentation and classification. System identification techniques are used to derive analytical error analysis and performance estimates. Once, the primitives are detected an iterative approach is used to find the sequence of primitives for a novel action. Another approach in this context is presented by Bissacco [35]. They extract some temporal statistics from the images and use them to build a dynamical system that models contact forces explicitly. Then, they explicitly factor out exogenous inputs that are not unique to an individual.

Lu et al. [226] also approach the problem from a system theoretic point of view. Their goal is to segment and represent repetitive movements. For this, they model the joint data over time with a second order auto-regressive (AR) model and the segmentation problem is approached by detection significant changes of the dynamical parameters. Then, for each motion segment and for each joint, they model the motion with a damped harmonic model. In order to compare actions, a metric based on the dynamic model parameters is defined. A different problem is studied by Wang et al. [387] addressing what kind of cost function should be used to assure smooth transitions between primitives.

While most scientists concentrate on the action representation by circumventing the vision problem, Rao et al. [298] take a vision-based approach. They propose a view-invariant representation of action based on *dynamic instants* and *intervals*. Dynamic instants are used as primitives of actions which are computed from discontinuities of 2D hand trajectories. An interval represents the time period between two dynamic instants (key poses). A similar approach of using meaningful instants in time is proposed by Reng et al. [303] where key poses are found based on the curvature and covariance of the normalized trajectories. Cuntoor et al. [72] find key poses through evaluation of anti-eigenvalues.

Gonzàlez et al. [121] employ the point distribution model [68] to model the variability of joint angle settings of a stick figure model. An action spaces, *aSpace*, is trained by giving a set of joint angle settings coming from different individuals but showing the same action. *aSpaces* are then used for synthesis and recognition of known actions. Modeling of activities on a semantic level has been attempted by Park and Aggarwal[281]. The system they describe has 3 abstraction levels. At the first level, human body parts are detected using a Bayesian network. At the second level, DBNs are used to model the actions of a single person. At the highest level, the results from the second level are used to identify the interactions between individuals. Ivanov and Bobick [170] suggest using stochastic parsing for a semantic representation of an action. They discuss that for some activities, where it comes to semantic or temporal ambiguities or insufficient data, stochastic approaches may be insufficient to model complex actions and activities. They suggest decoupling actions into primitive components and using a stochastic parser for recognition. In [170] they pick up a work by Stolcke [356] on syntactic parsing in speech recognition and enhance this work for activity recognition in video data. Yamamoto et al. [403] present an application where a stochastic context free grammar is used for action recognition. A somewhat different approach is taken by Yu and Yang [413]. They use neural networks to find primitives. They apply self-organizing maps (SOMs, Kohonen's feature maps [197]) which cluster the training images based on shape feature data. After training the SOMs generated a label for each input image which converts an input image sequence into a sequence of labels. A subsequent clustering algorithm allows to find repeatedly appearing substructures in these label sequences. These substructures are then interpreted as motion primitives. A very interesting approach is presented by Lv and Nevatia in [229] where the authors are interested in recognizing and segmenting full-body human action. Lv and Nevatia decompose the large joint space into a set feature spaces where each feature corresponds to a single joint or combinations of related joints. They use then HMMs to recognize each action class based on the features and an AdaBoost scheme to detect and recognize the features.

5.6. Discussion of advances in human action recognition

The field of recognizing human actions has received a considerable increase of attention in the last few years. It is apparent from the published works, that the major interest lies in the field of surveillance and the related action understanding problems. While in some publications, the actions are interpreted without explicitly considering humans, others discuss the dynamics of humans, explicitly. In the latter, a large attention is devoted to rather simple actions such as walking, running, and sitting. Here, only a small body of literature goes beyond these simple actions into motion interpretation where scene context and the interaction with other humans is considered, e.g., [170,281,311,400,403]. Much more work is expected to appear in this context and the approaches will be interesting as they are likely to bridge the traditional vision field with the field of artificial intelligence.

On the other hand, a good understanding of these simple actions is necessary before they can be combined into more complex ones. The issues lie, e.g., in the invariances with respect to viewing angle, speed, and variations between individuals [98,328,379].

Another significant part of the discussed articles draw some of their motivations from neuroscientific studies [117,307,308] and deal explicitly with action primitives, action grammars [170,281,403], and the close relationship between action recognition and action synthesis [34,50,51,77,322]. As these works also build on action primitives a better understanding of action primitives is necessary also in this context, e.g., in order to generalize the HMMs as proposed by Billard and Calinon [34,50,51].

6. Conclusion

Over the past 5 years vision-based human motion estimation and analysis has continued to be a thriving area of research. This survey has identified over three-hundred related publications over the period 2000–06 in major conferences and journals. Increased activity in this research area has been driven by both the scientific challenge of automatic scene interpretation and the demands of potential mass-market applications in surveillance, entertainment production and indexing visual media.

During this period there has been substantial progress towards automatic human motion tracking and reconstruction. Recognition of human motion has also become a central focus of research interest. Key advances identified in this review include:

Initialization. Automatic initialization of model shape, appearance and pose has been addressed in recent work [59,238]. A major advance is the introduction of methods for pose detection from static images [158,302,315,326] which potentially provide automatic initialization for human motion reconstruction.

Tracking. Surveillance applications have motivated research advances towards reliable tracking of multiple people in unstructured outdoor scenes. Advances in especially the use of appearance, shape and motion for figure-ground segmentation have increased reliability of detecting and tracking people with partial occlusion [155,194,244,269,280,316,404]. Probabilistic classification methods [194,232,280,285] and stochastic sampling [155,269,290,345,404,421] have been introduced to improve the reliability of temporal correspondence during occlusion. Systems for self-calibrating and tracking across multiple cameras have been investigated [21,187,193,369]. There remains a gap between the state-of-the-art and robust tracking of people for surveillance in outdoor scenes.

Human motion reconstruction from multiple views. Significant progress has been made towards the goal of automatic reconstruction of human movement from video. The model-based analysis-by-synthesis methodology, pioneered in early work [148], has been extended with the introduction of techniques to efficiently search the space of possible pose configurations for robust reconstruction from multiple view video acquisition [53,90,191,238]. Current approaches capture gross body movement but do not accurately reconstruct fine detail such as hand movements or axial rotations.

Monocular human motion reconstruction. Progress has also been made towards human motion capture from single views with stochastic sampling techniques [211,265,332,343]. An increasing trend in monocular tracking has been the use of learnt motion models to constrain reconstruction based on movement [7,8,332,334,373,372]. Research has demonstrated that the use of strong a priori models enables improved monocular tracking of specific movements.

Pose estimation in natural scenes. A recent trend to overcome limitations of monocular tracking in video of unstructured scenes has been direct pose detection on individual frames. Probabilistic assemblies of parts based on robust body part detection has achieved 2D pose estimation in challenging cluttered scenes such as film footage [158,235,240,296,302,314]. Example-based methods which learn a mapping from image to 3D pose space have been presented for reconstruction of specific movements [8,315,326].

Recognition. Understanding behavior and action has recently seen an explosion of research interest. Considerable steps have been made to advance surveillance applications towards automatic detection of unusual activities. Progress can also be seen for the recognition of simple actions and the description of action grammars. Relatively few papers have so far dealt with higher abstraction levels in action grammars which touch the border of semantics and AI. Association of actions and activities with affordances of objects will also bring a new perspective to object recognition.

Future research in visual analysis of human movement must address a number of open problems to satisfy the common requirements of potential applications for reliable automatic tracking, reconstruction and recognition. Body part detectors which are invariant to viewpoint, body shape, and clothing are required to achieve reliable tracking and pose estimation in cluttered natural scenes. The use of learnt models of pose and motion are currently restricted to specific movements. More general models are required to provide constraints for capturing a wide range of human movement. Whilst there has been substantial advances in human motion reconstruction the visual understanding of human behavior and action remains immature despite a surge of recent interest. Progress in this area requires fundamental advances in behavior representation for dynamic scenes, viewpoint invariant relationships for movement and higher level reasoning for interpretation of actions [325].

Industrial applications also require specific advances: human motion capture for entertainment production requires accurate multiple view reconstruction; surveillance applications require both reliable detection of people and recognition of movement and behavior from relatively low quality imagery; human-computer interfaces require low-latency real-time recognition of gestures, actions, and natural behaviors. The potential of these applications will continue to inspire the advances required to realize reliable visual capture and analysis of moving people.

Acknowledgments

The authors thank the following people for providing valuable comments to the paper: the anonymous reviewers, Prof. Larry S. Davis, Dr. Jordi González, Dr. Hedvig Kjellström (formerly Sidenbladh), Prof. Hans-Hellmut Nagel, Prof. Ramakant Nevatia, and Prof. Mubarak Shah. Thomas B. Moeslund is supported by the Danish National Research Councils and HERMES (FP6 IST-027110). Adrian Hilton is supported by EPSRC GR/S13576 Visual Media Platform Grant. Volker Krüger is supported by PACO-PLUS (FP6 IST-IP-027657).

References

[1] http://www.charactermotion.com/products/powermoves/megamocap/.
[2] http://mocap.cs.cmu.edu/.
[3] http://www.eyetoy.com.
[4] http://www.ict.usc.edu/graphics/animWeb.
[5] http://www.visualsurveillance.org.
[6] A. Agarwal, B. Triggs, 3D human pose from silhouettes by relevance vector regression, in: Computer Vision and Pattern Recognition, Washington DC, USA, June 2004.
[7] A. Agarwal, B. Triggs, Tracking articulated motion with piecewise learned dynamic models, in: European Conference on Computer Vision, Prague, Czech Republic, May 11–14, 2004.
[8] A. Agarwal, B. Triggs, Recovering 3D human pose from monocular images, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (1) (2006) 44–58.
[9] J.K. Aggarwal, Q. Cai, Human motion analysis: a review, Computer Vision and Image Understanding 73 (3) (1999) 428–440.
[10] J.K. Aggarwal, Q. Cai, W. Liao, B. Sabata, Articulated and elastic non-rigid motion: a review, in: Workshop on Motion of Non-Rigid and Articulated Objects, Austin, Texas, Nov 1994.
[11] J.K. Aggarwal, Q. Cai, W. Liao, B. Sabata, Nonrigid motion analysis: articulated and elastic motion, Computer Vision and Image Understanding 70 (2) (1998) 142–156.
[12] J.K. Aggarwal, S. Park, Human motion: modeling and recognition of actions and interactions, in: Second International Symposium on 3D Data Processing, Visualization and Transmission, Thessaloniki, Greece, Sep 6–9, 2004.
[13] M. Ahmad, S. Lee, Human action recognition using multi-view image sequence features, in: International Conference on Automatic Face and Gesture Recognition, Southampton, UK, April 10–12, 2006.
[14] B. Allen, B. Curless, Z. Popovic, Articulated body deformation from range scan data, in: ACM SIGGRAPH, 2002, pp. 612—619.
[15] B. Allen, B. Curless, Z. Popovic, The space of human body shapes: reconstruction and parameterization from range images, in: ACM SIGGRAPH, 2003, pp. 587—594.
[16] J. Ambrosio, J. Abrantes, G. Lopes, Spatial reconstruction of human motion by means of a single camera and a biomechanical model, Journal of Human Movement Science 20 (2001) 829–851.
[17] J. Ambrosio, G. Lopes, J. Costa, J. Abrantes, Spatial reconstruction of the human motion based on images of a single camera, Journal of Biomechanics 34 (2001) 1217–1221.
[18] P.F. Andersen, R. Corlin, Tracking of Interacting People and Their Body Parts for Outdoor Surveillance, Master's thesis, Laboratory of Computer Vision and Media Technology, Aalborg University, Denmark, 2005.
[19] G. Antonini, S.V. Martinez, M. Bierlaire, J.P. Thiran, Behavioral priors for detection and tracking of pedestrians in video sequences, International Journal of Computer Vision 69 (2) (2006) 159–180.
[20] O. Arikan, D.A. Forsyth, Synthesizing constrained motions from examples, in: ACM SIGGRAPH, 2002, pp. 483—490.
[21] N. Atsushi, K. Hirokazu, H. Shinsaku, I. Siji, Tracking multiple people using distributed vision systems, in: International Conference on Robotics and Automation, Washington DC, USA, May 2002.
[22] Y. Azoz, L. Devi, M. Yeasin, Tracking the human arm using constraint fusion and multiple-cue localization, Machine Vision and Applications 13 (5–6) (2003) 286–302.
[23] R.V. Babu, K.R. Ramakrishnan, Compressed domain human motion recognition using motion history information, in: International Conference on Acoustics, Speech and Signal Processing, Hong Kong, China, April 6–10, 2003.
[24] A.O. Balan, M.J. Black, An adaptive appearance model approach for model-based articulated object tracking, in: Computer Vision and Pattern Recognition, New York City, New York, USA, June 17–22, 2006.
[25] A.O. Balan, L. Sigal, M.J. Black, A quantitative evaluation of video-based 3D person tracking, in: Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, Beijing, China, Oct 15–21, 2005.
[26] C. Barron, I.A. Kakadiaris, Estimating anthropometry and pose from a single image, in: Computer Vision and Pattern Recognition, Hilton Head Island, South Carolina, June 13–15, 2000.
[27] C. Barron, I.A. Kakadiaris, Estimating anthropometry and pose from a single uncalibrated image, Computer Vision and Image Understanding 81 (3) (2001) 269–284.
[28] C. Barron, I.A. Kakadiaris, On the improvement of anthropometry and pose estimation from a single uncalibrated image, Machine Vision and Applications 14 (4) (2003) 229–236.
[29] C. Beleznai, B. Fruhstuck, H. Bischof, Tracking multiple humans using fast mean shift mode seeking, in: Workshop on Performance Evaluation of Tracking and Surveillance, Breckenridge, Colorado, Jan 2005.
[30] S. Belongie, J. Malik, J. Puzicha, Matching shapes, in: International Conference on Computer Vision, Vancouver, Canada, July 9–12, 2001.
[31] J. Ben-Arie, Z. Wang, P. Pandit, S. Rajaram, Human activity recognition using multidimensional indexing, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (8) (2002) 1091–1104.
[32] C. BenAbdelkader, R. Cutler, L. Davis, Motion-based recognition of people in EigenGait space, in: International Conference on Automatic Face and Gesture Recognition, Washington DC, USA, May 20–21, 2002.
[33] J. Berclaz, F. Fleuret, P. Fua, Robust people tracking with global trajectory optimization, in: Computer Vision and Pattern Recognition, New York City, New York, USA, June 17–22, 2006.
[34] A. Billard, Y. Epars, S. Calinon, S. Schaal, G. Cheng, Discovering optimal imitation strategies, Robotics and Autonomous Systems 47 (2004) 69–77.
[35] A. Bissacco, S. Soatto, Classifying human dynamics without contact forces, in: Computer Vision and Pattern Recognition, New York City, New York, USA, June 17–22, 2006.
[36] B. Blank, L. Gorelick, E. Shechtman, M. Irani, R. Basri, Actions as space-time shapes, in: Internatinal Conference on Computer Vision, Beijing, China, Oct 15–21, 2005.
[37] A. Bobick, Movement, activity, and action: the role of knowledge in the perception of motion, Philosophical Transactions of the Royal Society of London 352 (1997) 1257–1265.
[38] A. Bobick, J. Davis, The recognition of human movement using temporal templates, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (3) (2001) 257–267.
[39] O. Boiman, M. Irani, Detecting irregularities in images and in video, in: Internatinal Conference on Computer Vision, Beijing, China, Oct 15–21, 2005.
[40] G.R. Bradski, J.W. Davis, Motion segmentation and pose recognition with motion history gradients, Machine Vision and Applications 13 (3) (2002) 174–184.
[41] M. Brand, Shadow puppetry, in: International Conference on Computer Vision, Corfu, Greece, Sep 1999.
[42] M. Bray, P. Kohli, P.H.S. Torr, PoseCut: simultaneous segmentation and pose estimation of humans using dynamic graph-cuts, in: European Conference on Computer Vision, Graz, Austria, May 7–13, 2006.
[43] C. Bregler, J. Malik, K. Pullen, Twist based acquisition and tracking of animal and human kinematics, International Journal of Computer Vision 56 (3) (2004) 179–194.
[44] G.J. Brostow, I. Essa, D. Steedly, V. Kwatra, Novel skeletal representation for articulated creatures, in: European Conference on Computer Vision, Prague, Czech Republic, May 11–14, 2004.
[45] J.M. Buades, R. Mas, F.J. Perales, Matching a human walking sequence with a VRML synthetic model, in: Workshop on Articulated Motion and Deformable Objects, LNCS 1899, Palma de Mallorca, Spain, Sep 2000.

[46] J.M. Buades, F.J. Perales, Upper body tracking for interactive applications, in: Conference on Articulated Motion and Deformable Objects (AMDO), Andratx, Mallorca, Spain, July 11–14, 2006.
[47] D. Bullock, J. Zelek, Towards real-time 3D monocular visual tracking of human limbs in unconstrained environments, Real-Time Imaging 11 (2005) 323–353.
[48] H. Buxton, Learning and understanding dynamic scene activity: a review, Image and Vision Computing 21 (1) (2003) 125–136.
[49] Q. Cai, A. Mitiche, J.K. Aggarwal, Tracking human motion in an indoor environment, in: International Conference on Image Processing, Washington DC, USA, Oct 23–26, 1995.
[50] S. Calinon, A. Billard, Stochastic gesture production and recognition model for a humanoid robot, in: International Conference on Intelligent Robots and Systems, Alberta, Canada, Aug 2–6, 2005.
[51] S. Calinon, F. Guenter, A. Billard, Goal-directed imitation in a humanoid robot, in: International Conference on Robotics and Automation, Barcelona, Spain, April 18–22, 2005.
[52] M.B. Capellades, D. Doermann, D. DeMenthon, R. Chellappa, An appearance based approach for human and object tracking, in: International Conference on Image Processing, Barcelona, Spain, Sep 14–17, 2003.
[53] J. Carranza, C. Theobalt, M. Magnor, H.-.P. Seidel, Free-viewpoint video of human actors, in: ACM SIGGRAPH, 2003, pp. 565—577.
[54] C. Cedras, M. Shah, Motion-based recognition: a survey, Image and Vision Computing 13 (2) (1995) 129–155.
[55] T.H. Chalidabhongse, K. Kim, D. Harwood, L. Davis, A perturbation method for evaluating background subtraction algorithms, in: Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, Beijing, China, Oct 15–16, 2005.
[56] I.C. Chang, C.L. Huang, The model-based human motion analysis system, Image and Vision Computing 18 (2000) 1067–1083.
[57] M. Chen, G. Ma, S. Kee, Pixels classification for moving object extraction, in: IEEE Workshop on Motion and Video Computing (MOTION'05), Breckenridge, Colorado, Jan 2005.
[58] F. Cheng, W.J. Christmas, J. Kittler, Recognising human running behaviour in sports video sequences, in: International Conference on Pattern Recognition, Quebec, Canada, Aug 11–15, 2002.
[59] G. Cheung, S. Baker, T. Kanade, Shape-from-silhouette for articulated objects and its use for human body kinematics estimation and motion capture, in: Computer Vision and Pattern Recognition, Madison, Wisconsin, USA, June 16–22, 2003.
[60] K. Cheung, S. Baker, T. Kanade, Shape-from-silhouette across time part II: applications to human modeling and markerless motion tracking, International Journal of Computer Vision 63 (3) (2005) 225–245.
[61] S.S. Cheung, C. Kamath, Robust techniques for background subtraction in Urban traffic video, in: Video Communications and Image Processing. SPIE Electronic Imaging, vol. 5308, San Jose, California, Jan 2004.
[62] K. Choo, D.J. Fleet, People tracking using hybrid Monte Carlo filtering, in: International Conference on Computer Vision, Vancouver, Canada, July 9–12, 2001.
[63] A.R. Chowdhury, R. Chellappa, A factorization approach for event recognition, in: CVPR Event Mining Workshop, Madison, Wisconsin, USA, June 16–22, 2003.
[64] A.R. Chowdhury, R. Chellappa, A factorization approach for activity recognition, in: Computer Vision and Pattern Recognition, Washington DC, USA, June 2004.
[65] C.-W. Chu, O.C. Jenkins, M.J. Mataric, Markerless kinematic model and motion capture from volume sequences, in: Computer Vision and Pattern Recognition, Madison, Wisconsin, USA, June 16–22, 2003.
[66] Collins, Lipton, Kanade, Fujiyoshi, Duggins, Tsin, Tolliver, Enomoto, Hasegawa, A system for video surveillance and monitoring: VSAM final report, Technical report CMU-RI-TR-00-12, Robotics Institute, Carnegie Mellon University, 2000.
[67] D. Comaniciu, V. Ramesh, P. Meer, Kernel-based object tracking, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (5) (2003) 564–575.
[68] T.F. Cootes, C.J. Taylor, D.H. Hopper, J. Graham, Active shape models—their training and application, Computer Vision and Image Understanding 61 (1) (1995) 39–59.
[69] R. Cucchiara, C. Grana, M. Piccardi, A. Prati, Detecting moving objects, ghosts, and shadows in video streams, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (10) (2003) 1337–1342.
[70] R. Cucchiara, C. Grana, G. Tardini, R. Vezzani, Probabilistic people tracking for occlusion handling, in: International Conference on Pattern Recognition, Cambridge, UK, Aug 23–26, 2004.
[71] R. Cucchiara, R. Vezzani, Assessing temporal coherence for posture classification with large occlusions, in: Workshop on Motion and Video Computing (MOTION'05), Breckenridge, Colorado, Jan 2005.
[72] N. Cuntoor, R. Chellappa, Key frame-based activity representation using anti-eigenvalues, in: Asian Conference on Computer Vision, vol. 3852 of LNCS, Hyderabad, India, Jan 13–16, 2006.
[73] C. Curio, M.A. Giese, Combining view-based and model-based tracking of articulated human movements, in: MOTION, Breckenridge, Colorado, USA, Jan 5–7, 2005.
[74] M. Dahmane, J. Meunier, Real-time video surveillance with self-organizing maps, in: Canadian Conference on Computer and Robot Vision, Victoria, British Columbia, Canada, May 9–11, 2005.
[75] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Computer Vision and Pattern Recognition, San Diego, CA, June 20–25, 2005.
[76] N. Dalal, B. Triggs, C. Schmid, Human detection using oriented histograms of flow and appearance, in: European Conference on Computer Vision, Graz, Austria, May 7–13, 2006.
[77] B. Dariush, Human motion analysis for biomechanics and biomedicine, Machine Vision and Applications 14 (2003) 202–205.
[78] N. Date, H. Yoshimoto, D. Arita, R. Taniguchi, Real-time human motion sensing based on vision-based inverse kinematics for interactive applications, in: International Conference of Pattern Recognition, Cambridge, UK, Aug 23–26, 2004.
[79] J. Davis, H. Gao, Recognizing human action efforts: an adaptive three-mode PCA framework, in: International Conference on Computer Vision, Nice, France, Oct 13–16, 2003.
[80] J.W. Davis, A. Bobick, The representation and recognition of action using temporal templates, in: Conference on Computer Vision and Pattern Recognition, San Juan, Puerto Rico, June 1997.
[81] J.W. Davis, H. Gao, Gender recognition from walking movements using adaptive three-mode PCA, in: Computer Vision and Pattern Recognition, Washington DC, USA, June 2004.
[82] J.W. Davis, V. Sharma, Robust detection of people in thermal imagery, in: International Conference on Pattern Recognition, Cambridge, UK, Aug 23–26, 2004.
[83] J.W. Davis, S.R. Taylor, Analysis and recognition of walking movements, in: International Conference on Pattern Recognition, Quebec, Canada, Aug 11–15, 2002.
[84] L. Davis, V. Philomin, R. Duraiswami, Tracking humans from a moving platform, in: International Conference on Pattern Recognition, Barcelona, Spain, Sep 3–8, 2000.
[85] A.J. Davison, J. Deutscher, I.D. Reid, Markerless motion capture of complex full-body movement for character animation, in: Eurographics Workshop on Computer Animation and Simulation, Manchester, UK, Sep 2001.
[86] Q. Delamarre, O. Faugeras, 3D articulated models and multi-view tracking with physical forces, Computer Vision and Image Understanding 81 (3) (2001) 328–357.
[87] D. Demirdjian, Enforcing constraints for human body tracking, in: Workshop on Multiple Object Tracking, Madison, Wisconsin, USA, June 16–22, 2003.
[88] D. Demirdjian, Combining geometric- and view-based approaches for articulated pose estimation, in: European Conference on Computer Vision, Prague, Czech Republic, May 11–14, 2004.

[89] D. Demirdjian, T. Ko, T. Darrell, Constraining human body tracking, in: International Conference on Computer Vision, Beijing, China, Oct 15–21, 2003.

[90] J. Deutscher, A. Blake, I. Reid, Articulated body motion capture by annealed particle filtering, in: Computer Vision and Pattern Recognition, Hilton Head Island, South Carolina, June 13–15, 2000.

[91] J. Deutscher, A. Davison, I. Reid, Automatic partitioning of high dimensional search spaces associated with articulated body motion capture, in: Computer Vision and Pattern Recognition, Kauai Marriott, Hawaii, Dec 9–14, 2001.

[92] J. Deutscher, I. Reid, Articulated body motion capture by stochastic search, International Journal of Computer Vision 61 (2) (2005) 185–205.

[93] M. Dimitrijevic, V. Lepetit, P. Fua, Human body pose recognition using spatio-temporal templates, in: Workshop on Modeling People and Human Interaction, Beijing, China, Oct 15–21, 2005.

[94] A.A. Efros, A.C. Berg, G. Mori, J. Malik, Recognizing action at a distance, in: International Conference on Computer Vision, Nice, France, Oct 13–16, 2003.

[95] A. Elgammal, R. Duraiswami, L.S. Davis, Probabilistic tracking in joint feature-spatial spaces, in: Computer Vision and Pattern Recognition, Madison, Wisconsin, June 16–22, 2003.

[96] A. Elgammal, D. Harwood, L. Davis, Non-parametric model for background subtraction, in: European Conference on Computer Vision, Dublin, Ireland, June 2000.

[97] A. Elgammal, C. Lee, Separating style and content on a nonlinear manifold, in: Computer Vision and Pattern Recognition, Washington DC, USA, June 2004.

[98] A. Elgammal, C.S. Lee, Inferring 3D body pose from silhouettes using activity manifold learning, in: Computer Vision and Pattern Recognition, Washington DC, USA, June 2004.

[99] A. Elgammal, V. Shet, Y. Yacoob, L. Davis, Learning dynamics for exemplar-based gesture recognition, in: Computer Vision and Pattern Recognition, Madison, Wisconsin, USA, June 16–22, 2003.

[100] A.M. Elgammal, L.S. Davis, Probabilistic framework for segmenting people under occlusion, in: International Conference on Computer Vision, Vancouver, Canada, July 9–12, 2001.

[101] H.L. Eng, K.A. Toh, A.H. Kam, J. Wang, W.Y. Yau, An automatic drowning detection surveillance system for challenging outdoor pool environments, in: International Conference on Computer Vision, Nice, France, Oct 13–16, 2003.

[102] H.L. Eng, J. Wang, A.H.K.S. Wah, W.Y. Yau, Robust human detection within a highly dynamic aquatic environment in real time, IEEE Transactions on Image Processing 15 (6) (2006) 1583–1600.

[103] C. Fanti, L. Zelnik-Manor, P. Perona, Hybrid models for human motion recognition, in: Computer Vision and Pattern Recognition, San Diego, California, USA, June 20–25, 2005.

[104] P.F. Felzenszwalb, D.P. Huttenlocher, Efficient matching of pictorial structures, in: Computer Vision and Pattern Recognition, Hilton Head Island, South Carolina, USA, June 13–15, 2000.

[105] P. Figueroa, N. Leite, R.M.L. Barros, Tracking soccer players using the graph representation, in: International Conference on Pattern Recognition, Cambridge, UK, Aug 2004.

[106] P. Figueroa, N. Leite, R.M.L. Barros, Background recovering in outdoor image sequences: an example of soccer players segmentation, Image and Vision Computing 24 (4) (2006) 363–374.

[107] P. Figueroa, N. Leite, R.M.L. Barros, Tracking soccer players aiming their kinematical motion analysis, Computer Vision and Image Understanding 101 (2006) 122–135.

[108] P. Fihl, R. Corlin, S. Park, T.B. Moeslund, M.M. Trivedi, Tracking of individuals in very long video sequences, in: International Symposium on Visual Computing (ISVC), Stateline, Lake Tahoe, Nevada, USA, Nov 6–8, 2006.

[109] P. Fihl, M.B. Holte, T.B. Moeslund, L. Reng, Action recognition using motion primitives and probabilistic edit distance, in: Conference on Articulated Motion and Deformable Objects (AMDO), Andratx, Mallorca, Spain, July 11–14, 2006.

[110] D.A. Forsythe, M.M. Fleck, Body plans, in: Computer Vision and Pattern Recognition, Puerto Rico, June 17–19, 1997.

[111] J.P. Foster, M.S. Nixon, A. Prngel-Bennett, Automatic gait recognition using area-based metrics, Pattern Recognition Letters 24 (2003) 2489–2497.

[112] P. Fua, A. Gruen, N. D'Apuzzo, R. Plänkers, Markerless full body shape and motion capture from video sequences, in: Symposium on Close Range Imaging, International Society for Photogrammetry and Remote Sensing, Corfu, Greece, Sep 2002.

[113] J. Gao, A.G. Hauptmann, H.D. Wactlar, Combining motion segmentation with tracking for activity analysis, in: International Conference on Automatic Face and Gesture Recognition, Seoul, Korea, May 17–19, 2004.

[114] D.M. Gavrila, The visual analysis of human movement: a survey, Computer Vision and Image Understanding 73 (1) (1999) 82–98.

[115] P. Gerard, A. Gagalowicz, Human body tracking using a 3D generic model applied to golf swing analysis, in: Conference on Model-based Imaging, Rendering, image Analysis and Graphical special Effects, INRIA Rocquencourt, France, March 10–11, 2003.

[116] J. Giebel, D.M. Gavrila, C. Schnvrr, A Bayesian framework for multi-cue 3D object tracking, in: European Conference on Computer Vision, Prague, Czech Republic, May 11–14, 2004.

[117] M. Giese, T. Poggio, Neural mechanisms for the recognition of biological movements, Nature Reviews 4 (2003) 179–192.

[118] M. Gleicher, Evaluating video-based motion capture, in: Conference on Computer Animation, Geneva, Switzerland, June 19–21, 2002.

[119] B. Gloyer, H.K. Aghajan, K.Y.S. Siu, T. Kailath, Video-based freeway monitoring system using recursive vehicle tracking, in: IS&T-SPIE Symposium on Electronic Imaging: Image and Video Processing, 1995.

[120] J. Gonzàlez, Human Sequence Evaluation: the Key-frame Approach, Ph.D. thesis, Univertitat Autonoma de Barcelona, Spain, 2004.

[121] J. Gonzàlez, J. Varona, F.X. Roca, J.J. Villanueva, aSpaces: action spaces for recognition and synthesis of human actions, in: International Workshop on Articulated Motion and Deformable Objects, Palma de Mallorca, Spain, Nov 21–23, 2002.

[122] J.J. Gonzalez, I.S. Lim, P. Fua, D. Thalmann, Robust tracking and segmentation of human motion in an image sequence, in: International Conference on Acoustics, Speech, and Signal Processing, Hong Kong, April 2003.

[123] L. Gorelick, M. Galun, E. Sharon, A. Brandt, R. Basri, Shape representation and recognition using the poisson equation, in: Computer Vision and Pattern Recognition, Washington DC, USA, June 2003.

[124] N. Grammalidis, G. Goussis, G. Troufakos, M.G. Strintzis, 3D human body tracking from depth images using analysis by synthesis, in: International Conference on Image Processing, Thessaloniki, Greece, Oct 7–10, 2001.

[125] K. Grauman, G. Shakhnarovich, T. Darrell, Inferring 3D structure with a statistical image-based shape model, in: International Conference on Computer Vision, Nice, France, Oct 13–16, 2003.

[126] P.J. Green, Highly Structured Stochastic Systems, Chapter Trans-dimensional Markov Chain Monte Carlo, Oxford University Press, 2003.

[127] A. Gritai, Y. Sheikh, M. Shah, On the use of anthropometry in the invariant analysis of human actions, in: International Conference on Pattern Recognition, Cambridge, UK, Aug 23–26, 2004.

[128] K. Grochow, S.L. Martin, A. Hertzmann, Z. Popovic, Style-based inverse kinematics, in: ACM Transactions on Graphics (SIGGRAPH), 2004.

[129] P. Guha, A. Mukerjee, K.S. Venkatesh, Efficient occlusion handling for multiple agent tracking by reasoning with surveillance event primitives, in: Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, Beijing, China, Oct 15–16, 2005.

[130] D. Gutchess, M. Trajkovic, E.C. Solal, D. Lyons, A. Jain, A background model initialization algorithm for video surveillance, in:

International Conference on Computer Vision, Vancouver, Canada, July 9–12, 2001.
[131] A.K. Halvorsen, Model-based Methods in Motion Capture, Ph.D. thesis, Uppsala University, Sweden, 2002.
[132] T.X. Han, H. Ning, T.S. Huang, Efficient nonparametric belief propagation with application to articulated body tracking, in: Computer Vision and Pattern Recognition, New York City, New York, USA, June 17–22, 2006.
[133] M. Hariadi, A. Harada, T. Aoki, T. Higuchi, An LVQ-based technique for human motion segmentation, in: Asia-Pacific Conference on Circuits and Systems, Bali, Indonesia, Oct 28–31, 2002.
[134] I. Hariatoglu, R. Cutler, D. Harwood, L.S. Davis, Backpack: detection of people carrying objects using silhouettes, Computer Vision and Image Understanding 81 (3) (2001) 385–397.
[135] I. Haritaoglu, M. Flickner, D. Beymer, Ghost3D: detecting body posture and parts using stereo, in: Workshop on Motion and Video Computing, Orlando, Florida, Nov 7, 2002.
[136] I. Haritaoglu, D. Harwood, L.S. Davis, Ghost: a human body part labeling system using silhouettes, in: International Conference on Pattern Recognition, Queensland, Australia, Aug 17–20, 1998.
[137] I. Haritaoglu, D. Harwood, L.S. Davis, W4: Who? When? Where? What?—A real time system for detecting and tracking people, in: International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 1998.
[138] I. Haritaoglu, D. Harwood, L.S. Davis, W4: real-time surveillance of people and their activities, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (8) (2000) 809–830.
[139] K. Hayashi, M. Hashimoto, K. Sumi, K. Sasakawa, Multiple-person tracker with a fixed slanting stereo camera, in: International Conference on Automatic Face and Gesture Recognition, Seoul, Korea, May 17–19, 2004.
[140] M. Heikkila, M. Pietikainen, J. Heikkila, A texture-based method for detecting moving objects, in: British Machine Vision Conference, London, UK, Sep 7–9, 2004.
[141] M. Heikkila, M. Pietikainen, A texture-based method for modeling the background and detecting moving objects, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (4) (2006) 657–662.
[142] L. Herda, Using Biomechanical Constraints to Improve Video-Based Motion Capture, Ph.D. thesis, Computer Vision Lab, EPFL, Lausanne, Switzerland, 2003.
[143] L. Herda, P. Fua, R. Plänkers, R. Boulic, D. Thalmann, Using skeleton-based tracking to increase the reliability of optical motion capture, Human Movement Science 20 (3) (2001).
[144] L. Herda, R. Urtasun, P. Fua, Hierarchical implicit surface joint limits to constrain video-based motion capture, in: European Conference on Computer Vision, Prague, Czech Republic, May 11–14, 2004.
[145] L. Herda, R. Urtasun, P. Fua, Hierarchical implicit surface joint limits for human body tracking, Computer Vision and Image Understanding 99 (2) (2005) 189–209.
[146] L. Herda, R. Urtasun, P. Fua, A. Hanson, An automatic method for determining quaternion field boundaries for ball-and-socket joint limits, in: International Conference on Automatic Face and Gesture Recognition, Washington DC, USA, May 20–21, 2002.
[147] A. Hilton, D. Beresford, T. Gentils, R. Smith, W. Sun, Virtual people: capturing human models to populate virtual worlds, in: International Conference on Computer Animation, May 1999.
[148] D. Hogg, Model-based vision: a program to see a walking person, Image and Vision Computing 1 (1) (1983) 5–20.
[149] T. Horprasert, D. Harwood, L.S. Davis, A statistical approach for real-time robust background subtraction and shadow detection, in: IEEE ICCV'99 Frame-Rate Workshop, Corfu, Greece, Sep 1999.
[150] R. Hoshino, S. Yonenmoto, D. Arita, R.I. Taniguchi, Real-time analysis of human motion using multi-view silhouette contours, in: The 12th Scandinavian Conference on Image Analysis, Bergen, Norway, 2001.
[151] N. Howe, Flow lookup and biological motion perception, in: International Conference on Image Processing, Genova, Italy, Sep 11–14, 2005.
[152] N.R. Howe, Silhouette lookup for automatic pose tracking, in: Workshop on Articulated and Non-Rigid Motion, Washington DC, USA, June 2004.
[153] N.R. Howe, Boundary fragment matching and articulated, in: Conference on Articulated Motion and Deformable Objects (AMDO), Andratx, Mallorca, Spain, July 11–14, 2006.
[154] N.R. Howe, M.E. Leventon, W.T. Freeman, Bayesian reconstruction of 3D human motion from single-camera videoAdvances in Neural Information Processing Systems, vol. 12, MIT Press, Cambridge, MA, 2000.
[155] M. Hu, W. Hu, T. Tan, Tracking people through occlusion, in: International Conference on Pattern Recognition, Cambridge, UK, Aug 2004.
[156] W. Hu, M. Hu, X. Zhou, T. Tan, J. Lou, S. Maybank, Principal axis-based correspondence between multiple cameras for people tracking, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (4) (2006) 663–671.
[157] W. Hu, T. Tan, L. Wang, S. Maybank, A survey on visual surveillance of object motion and behaviors, Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews 34 (3) (2004) 334–352.
[158] G. Hua, M.-H. Yang, Y. Wu, Learning to estimate human pose with data driven belief propagation, in: Computer Vision and Pattern Recognition, San Diego, California, USA, June 20–25, 2005.
[159] C.L. Huang, C.C. Lin, Model-based human body motion analysis for MPEG IV video encoding, in: International Conference on Information Technology: Coding and Computing, Las Vegas, Nevada, April 2–4, 2001.
[160] C.L. Huang, T.H. Tseng, H.C. Shih, A model-based articulated human motion tracking system, in: Asian Conference on Computer Vision, Jeju, Korea, Jan 27–30, 2004.
[161] F. Huang, H. Di, G. Xu, Viewpoint insensitive posture representation for action recognition, in: International Conference on Articulated Motion and Deformable Objects, Springer LNCS 4069, Andratx, Mallorca, Spain, July 11–14, 2006.
[162] J. Huang, S.R. Kumar, M. Mitra, W. Zhu, Spatial color indexing and applications, International Journal of Computer Vision 35 (3) (1999) 91–101.
[163] Y. Huang, T.S. Huang, Model-based human body tracking, in: International Conference on Pattern Recognition, Quebec, Canada, Aug 11–15, 2002.
[164] I. Huerta, D. Rowe, J. Gonzàlez, J.J. Villanueva, Efficient incorporation of motionless foreground objects for adaptive background segmentation, in: Conference on Articulated Motion and Deformable Objects (AMDO), Andratx, Mallorca, Spain, July 11–14, 2006.
[165] A.J. Ijspeert, J. Nakanishi, S. Schaal, Movement imitation with nonlinear dynamical systems in humanoid robots, in: International Conference on Robotics and Automation, Washington DC, USA, May 2002.
[166] S.S. Intille, A.F. Bobick, Recognizing planned multiperson action, Computer Vision and Image Understanding 81 (3) (2001) 414–445.
[167] S. Ioffe, D. Forsyth, Finding people by sampling, in: International Conference on Computer Vision, Korfu, Greece, 1999.
[168] S. Ioffe, D. Forsyth, Human tracking with mixtures of trees, in: International Conference on Computer Vision, Vancouver, Canada, July 9–12, 2001.
[169] M. Isard, A. Blake, Condensation—Conditional density propagation for visual tracking, International Journal on Computer Vision 29 (1) (1998) 5–28.
[170] Y. Ivanov, A. Bobick, Recognition of visual activities and interactions by stochastic parsing, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (8) (2000) 852–872.
[171] Y.A. Ivanov, A.F. Bobick, J. Liu, Fast lighting independent background subtraction, International Journal of Computer Vision 37 (2) (2000) 199–207.

[172] S. Iwase, H. Saito, Parallel tracking of all soccer players by integrating detected positions in multiple view images, in: International Conference on Pattern Recognition, Cambridge, UK, Aug 2004.

[173] D.S. Jang, S.W. Jang, H.I. Choi, 2D human body tracking with structural Kalman filter, Pattern Recognition 35 (2002) 2041–2049.

[174] T. Jeaggli, E.K. Meier, L.V. Gool, Monocular tracking with a mixture of view-dependent learned models, in: Conference on Articulated Motion and Deformable Objects (AMDO), Andratx, Mallorca, Spain, July 11–14, 2006.

[175] O.C. Jenkins, M. Mataric, Automated modularization of human motion into actions and behaviors, Technical Report CRES-02-002, Center for Robotics and Embedded Systems, University of S. California, 2002.

[176] O.C. Jenkins, M. Mataric, Deriving action and behavior primitives from human motion capture data, in: International Conference on Robotics and Automation, Washington DC, USA, May 2002.

[177] O.C. Jenkins, M.J. Mataric, Deriving action and behavior primitives from human motion data, in: International Conference on Intelligent Robots and Systems, Lausanne, Switzerland, Sep 30–Oct 4, 2002, pp. 2551–2556.

[178] A.D. Jepson, D.J. Fleet, T.F. El-Maraghi, Robust online appearance models for visual tracking, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (10) (2003) 1296–1311.

[179] H. Jiang, M. Drew, Z. Li, Successive convex matching for action detection, in: Computer Vision and Pattern Recognition, New York City, New York, USA, June 17–22, 2006, pp. 1646–1653.

[180] H. Jiang, M.S. Drew, Z.-N. Li, Successive convex matching for action detection, in: Computer Vision and Pattern Recognition, New York City, New York, USA, June 17–22, 2006.

[181] S. Ju, Human Motion Estimation and Recognition (Depth Oral Report), Technical report, University of Toronto, 1996.

[182] I.N. Junejo, O. Javed, M. Shah, Multi feature path modeling for video surveillance, in: International Conference on Pattern Recognition, Cambridge, UK, Aug 2004.

[183] A. Kale, A. Rajagopalan, N. Cuntoor, V. Krueger, Human identification using gait, in: International Conference on Automatic Face and Gesture Recognition, Washington DC, USA, May 21–22, 2002.

[184] A. Kale, A. Sundaresan, A.N. Rjagopalan, N. Cuntoor, A.R. Chowdhury, V. Krnger, R. Chellappa, Identification of humans using gait, IEEE Transactions on Image Processing 9 (2004) 1163–1173.

[185] Y. Kameda, M. Minoh, A human motion estimation method using 3-successive video frames, in: International Conference on Virtual Systems and Multimedia, Gifu City, Japan, Sep 1996.

[186] D.W. Kang, Y. Onuma, J. Ohya, Estimating complicated and overlapped human body postures by wearing a multiple-colored suit using color information processing, in: International Conference on Automatic Face and Gesture Recognition, Seoul, Korea, May 17–19, 2004.

[187] J. Kang, I. Cohen, G. Medioni, Persistent objects tracking across multiple non-overlapping cameras, in: IEEE Workshop on Motion and Video Computing (MOTION'05), Breckenridge, Colorado, Jan 2005.

[188] J. Kang, I. Cohen, G. Medioni, C. Yuan, Detection and tracking of moving objects from a moving platform in presence of strong parallax, in: International Conference on Computer Vision, Beijing, China, Oct 15–21, 2005.

[189] I.A. Karaulova, P.M. Hall, A.D. Marshall, A hierarchical models of dynamics for tracking people with a single video camera, in: British Machine Vision Conference, Bristol, UK, Sep 11–14, 2000.

[190] Y. Ke, R. Sukthankar, M. Hebert, Efficient visual event detection using volumetric features, in: International Conference on Computer Vision, Beijing, China, Oct 15–21, 2005.

[191] R. Kehl, M. Bray, L. VanGool, Full body tracking from multiple views using stochastic sampling, in: Computer Vision and Pattern Recognition, San Diego, California, USA, June 20–25, 2005.

[192] D.G. Kendall, D. Barden, T.K. Carne, H. Le, Shape and Shape Theory, Wiley, New York, 1999.

[193] S. Khan, O. Javed, Z. Rasheed, M. Shah, Human tracking in multiple cameras, in: International Conference on Computer Vision, Vancouver, Canada, July 9–12, 2001.

[194] S. Khan, M. Shah, Tracking people in presence of occlusion, in: Asian Conference on Computer Vision, Taipei, Taiwan, Jan 8–11, 2000.

[195] S.M. Khan, M. Shah, A multiview approach to tracking people in crowded scenes using a planar homography constraint, in: European Conference on Computer Vision, Graz, Austria, May 7–13, 2006.

[196] K. Kim, T.H. Chalidabhongse, D. Harwood, L. Davis, Real-time foreground-background segmentation using codebook model, Real-Time Imaging 11 (3) (2005) 172–185.

[197] T. Kohonen, Self-organized formation of topologically correct feature maps, Biological Cybernetics 43 (1982) 59–69.

[198] A. Koschan, S. Kang, J. Paik, B. Abidi, M. Abidi, Color active shape models for tracking non-rigid objects, Pattern Recognition Letters 24 (2003) 1751–1765.

[199] L. Kovar, M. Gleicher, F. Pighin, Motion graphs, in: ACM SIGGRAPH, 2002, pp. 473–482.

[200] N. Krahnstoever, R. Sharma, Articulated models from video, in: Computer Vision and Pattern Recognition, Washington DC, USA, June 2004.

[201] N. Krahnstoever, M. Yeasin, R. Sharma, Automatic acquisition and initialization of articulated models, Machine Vision and Applications 14 (4) (2003) 218–228.

[202] F. Kristensen, P. Nilsson, V. Öwall, Background segmentation beyond RGB, in: Asian Conference on Computer Vision, LNCS 3852, Hyderabad, India, Jan 13–16, 2006.

[203] T. Krosshaug, R. Bahr, A model-based image-matching technique for three-dimensional reconstruction of human motion from uncalibrated video sequences, Journal of Biomechanics 38 (4) (2005) 919–929.

[204] V. Krüger, J. Anderson, T. Prehn, Probabilistic model-based background subtraction, in: Scandinavian Conference on Image Analysis, Joensuu, Finland, Jun 19–22, 2005.

[205] M.P. Kumar, P.H.S. Torr, A. Zisserman, Learning Layered motion segmentations of video, in: International Conference on Computer Vision, Beijing, China, Oct 15–21, 2005.

[206] C.S. Lee, A. Elgammal, Carrying object detection using pose preserving dynamic shape models, in: Conference on Articulated Motion and Deformable Objects (AMDO), Andratx, Mallorca, Spain, July 11–14, 2006.

[207] C.S. Lee, A. Elgammal, Human motion synthesis by motion manifold learning and motion primitive segmentation, in: Conference on Articulated Motion and Deformable Objects (AMDO), Andratx, Mallorca, Spain, July 11–14, 2006.

[208] D.S. Lee, Effective Gaussian mixture learning for video background substraction, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (5) (2005) 827–832.

[209] J. Lee, J. Chai, P.S.A. Reitsma, J.K. Hodgins, N.S. Pollard, Interactive control of avatars animated with human motion data, in: ACM SIGGRAPH, 2002, pp. 491–500.

[210] M.W. Lee, I. Cohen, Human upper body pose estimation in static images, in: European Conference on Computer Vision, Prague, Czech Republic, May 11–14, 2004.

[211] M.W. Lee, I. Cohen, Proposal maps driven MCMC for estimating human body pose in static images, in: Computer Vision and Pattern Recognition, Washington DC, USA, June 2004.

[212] M.W. Lee, I. Cohen, S.K. Jung, Particle filter with analytical inference for human body tracking, in: Workshop on Motion and Video Computing, Orlando, Florida, Nov 7, 2002.

[213] M.W. Lee, R. Navatia, Human pose tracking using multi-level structured models, in: European Conference on Computer Vision, Graz, Austria, May 7–13, 2006.

[214] M.W. Lee, R. Nevatia, Dynamic human pose estimation using Markov chain Monte Carlo approach, in: Workshop on Motion and

Video Computing (MOTION'05), Breckenridge, Colorado, Jan 2005.
[215] B. Leibe, E. Seemann, B. Schiele, Pedestrian detection in crowded scenes, in: Computer Vision and Pattern Recognition, San Diego, CA, June 20–25, 2005.
[216] I. Leichter, M. Lindenbaum, E. Rivlin, A general framework for combining visual trackers—-the black boxes approach, International Journal of Computer Vision 67 (3) (2006) 343–363.
[217] M. Leo, T. D'Orazio, I. Gnoni, P. Spagnolo, A. Distante, Complex human activity recognition for monitoring wide outdoor environments, in: International Conference on Pattern Recognition, Cambridge, UK, Aug 23–26, 2004.
[218] B. Li, R. Chellappa, H. Moon, Monte Carlo simulation techniques for probabilistic tracking, in: Thirty-Fifth Asilomar Conference on Signals, Systems and Computers, Pacific Grove, California, Nov 4–7, 2001.
[219] Y. Li, A. Hilton, J. Illingworth, A relaxation algorithm for real-time multiple view 3D-tracking, Image and Vision Computing 20 (12) (2002) 841–859.
[220] D. Liebowitz, S. Carlsson, Uncalibrated motion capture exploiting articulated structure constraints, International Journal of Computer Vision 51 (3) (2003) 171–187.
[221] H. Lim, V.I. Morariu, O.I. Camps, M. Sznaier, Dynamic apperance modeling for human tracking, in: Computer Vision and Pattern Recognition, New York City, New York, USA, June 17–22, 2006.
[222] S.N. Lim, A. Mittal, L.S. Davis, N. Paragios, Fast illumination-invariant background subtraction using two views: error analysis, sensor placement and applications, in: Computer Vision and Pattern Recognition, San Diego, CA, June 20–25, 2005.
[223] X. Liu, C. Chua, Multi-agent activity recognition using observation decomposed hidden Markov models, Image and Vision Computing 24 (2006) 166–175.
[224] D.G. Lowe, Distinctive Image features from scale-invariant keypoints, International Journal on Computer Vision 60 (2) (2004) 91–110.
[225] G. Loy, M. Eriksson, J. Sullivan, Monocular 3D reconstruction of human motion in long action sequences, in: European Conference on Computer Vision, Prague, Czech Republic, May 11–14, 2004.
[226] C. Lu, N. Ferrier, Repetitive motion analysis: segmentation and event classification, IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (2) (2004) 258–263.
[227] Y. Luo, T.-W. Wu, J.-N. Hwang, Object-based analysis and interpretation of human motion in sports video sequences by dynamic Bayesian networks, Computer Vision and Image Understanding 92 (2003) 196–216.
[228] F. Lv, J. Kang, R. Nevatia, I. Cohen, G. Medioni, Automatic tracking and labeling of human activities in a video sequence, in: 6th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, Prague, May 11–14, 2004.
[229] F. Lv, R. Navatia, Recognition and segmentation of 3D human action using HMM and multi-class AdaBoost, in: European Conference on Computer Vision, Graz, Austria, May 7–13, 2006.
[230] J. MacCormick, M. Isard, Partitioned sampling, articulated objects, and interface-quality hand tracking, in: European Conference on Computer Vision, Dublin, Ireland, June 2000.
[231] O. Masound, N. Papanikolopoulos, A method for human action recognitoin, Image and Vision Computing 21 (2003) 729–743.
[232] S.J. McKenna, S. Jabri, Z. Duric, H. Wechsler, Tracking Interacting People, in: International Conference on Automatic Face and Gesture Recognition, Grenoble, France, March 2000.
[233] C. Menier, E. Boyer, B. Raffin, 3D skeleton-based body pose recovery, in: International Symposium on 3D Data Processing, Visualisation and Transmission, 2006.
[234] D. Metaxas, From visual input to modeling humans, in: Conference on Computer Animation, Geneva, Switzerland, June 19–21, 2002.
[235] A. Micilotta, E. Ong, R. Bowden, Detection and tracking of humans by probabilistic body part assembly, in: British Machine Vision Conference, Oxford, UK, Sep 2005.
[236] A.S. Micilotta, E.-J. Ong, R. Bowden, Real-time upper body detection and 3D pose estimation in monoscopic images, in: European Conference on Computer Vision, Graz, Austria, May 7–13, 2006.
[237] I. Mikić, M. Trivedi, E. Hunter, P. Cosman, Human Body model acquisition and motion capture using voxel data, in: F.J. Perales, E.R. Hancock (Eds.), AMDO 2002, LNCS, vol. 2492, Springer, New York, 2002.
[238] I. Mikić, M. Trivedi, E. Hunter, P. Cosman, Human body model acquisition and tracking using voxel data, International Journal of Computer Vision 53 (3) (2003) 199–223.
[239] I. Mikić, M.M. Trivedi, E. Hunter, P. Cosman, Articulated body posture estimation from multi-camera voxel data, in: Computer Vision and Pattern Recognition, Kauai Marriott, Hawaii, Dec 9–14, 2001.
[240] K. Mikolajczyk, D. Schmid, A. Zisserman, Human detection based on a probabilistic assembly of robust part detectors, in: European Conference on Computer Vision, Prague, Czech Republic, May 11–14, 2004.
[241] J. Mitchelson, A. Hilton, From visual tracking to animation using hierarchical sampling, in: Conference on Model-based Imaging, Rendering, image Analysis and Graphical special Effects, Rocquencourt, France, March 10–11, 2003.
[242] J. Mitchelson, A. Hilton, Hierarchical tracking of multiple people, in: British Machine Vision Conference, Norwich, UK, Sep 2003.
[243] A. Mittal, L.S. Davis, M2Tracker: a multi-view approach to segmenting and tracking people in a cluttered scene using region-based stereo, in: European Conference on Computer Vision, Copenhagen, Denmark, 2002.
[244] A. Mittal, L.S. Davis, M2Tracker: a multi-view approach to segmenting and tracking people in a cluttered scene, International Journal of Computer Vision 51 (3) (2003) 189–203.
[245] T.B. Moeslund, Computer Vision-Based Motion Capture of Human Body Language, Ph.D. thesis, Lab of Computer Vision and Media Technology, Aalborg University, Denmark, 2003.
[246] T.B. Moeslund, Pose Estimating the Human Arm using Kinematics and the Sequential Monte Carlo Framework, chapter 4 in part IX of Cutting Edge Robotics. Pro literatur Verlag, ISBN:3-86611-038-3, 2005.
[247] T.B. Moeslund, E. Granum, A survey of computer vision-based human motion capture, Computer Vision and Image Understanding 81 (3) (2001) 231–268.
[248] T.B. Moeslund, E. Granum, Pose estimation of a human arm using kinematic constraints, in: The 12th Scandinavian Conference on Image Analysis, Bergen, Norway, 2001.
[249] T.B. Moeslund, E. Granum, Modelling and estimating the pose of a human arm, Machine Vision and Applications 14 (4) (2003) 237–247.
[250] T.B. Moeslund, E. Granum, Sequential Monte Carlo tracking of body parameters in a sub-space, in: International Workshop on Analysis and Modeling of Faces and Gestures, Nice, France, Oct 2003.
[251] T.B. Moeslund, E. Granum, Motion capture of articulated chains by applying auxiliary information to the sequential Monte Carlo algortihm, in: International Conference on Visualization, Imaging, and Image Processing, Marbella, Spain, Sep 2004.
[252] T.B. Moeslund, C.B. Madsen, E. Granum, Modelling the 3D pose of a human arm and the shoulder complex utilising only two parameters, International Journal on Integrated Computer-Aided Engineering 12 (2) (2005) 159–177.
[253] T.B. Moeslund, M. Vittrup, K.S. Pedersen, M.K. Laursen, M.K.D. Sørensen, H. Uhrenfeldt, E. Granum, Estimating the 3D shoulder position using monocular vision, in: International Conference on Imaging Science, Systems, and Technology, Las Vegas, Nevada, June 24–27, 2002.
[254] A. Mohan, C. Papageorgiou, T. Poggio, Example-based object detection in images by components, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (4) (2001) 349–361.

[255] L. Molina, A. Hilton, Sythesis of novel movements from a database of motion capture data, in: International Conference on Human Motion Analysis, Dec 2000.
[256] A. Monnet, A. Mittal, N. Paragios, V. Ramesh, Background modeling and subtraction of dynamic scenes, in: International Conference on Computer Vision, Nice, France, Oct 13–16, 2003.
[257] M. Montemerlo, S. Thrun, W. Whittaker, Conditional particle filter for simultaneous mobil robot localization and people-tracking, in: International Conference on Robotics and Automation, Washington DC, USA, May 2002.
[258] H. Moon, R. Chellappa, A. Rosenfeld, Tracking of human activities using shape-encoded particle propagation, in: International Conference on Image Processing, Thessaloniki, Greece, Oct 7–10, 2001.
[259] K. Moon, V. Pavlovic, Impact of dynamics on subspace embedding and tracking of sequences, in: Computer Vision and Pattern Recognition, New York City, New York, USA, June 17–22, 2006.
[260] G. Mori, J. Malik, Recovery of 3D human body configurations using shape contexts, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (7) (2006) 1052–1062.
[261] G. Mori, X. Ren, A. Efros, J. Malik, Recovering human body configurations: combining segmentation and recognition, in: Computer Vision and Pattern Recognition, Washington DC, USA, June 2004.
[262] J. Mulligan, Upper body pose estimation from stereo and hand-face tracking, in: Canadian Conference on Computer and Robot Vision, Victoria, British Columbia, Canada, May 9–11, 2005.
[263] T. Murakita, T. Ikeda, H. Ishiguro, Multisensor human tracker based on the Markov chain Monte Carlo method, in: International Conference on Pattern Recognition, Cambridge, UK, Aug 2004.
[264] H.-H. Nagel, From image sequences towards conceptual descriptions, Image and Vision Computing 6 (2) (1988) 59–74.
[265] R. Navaratnam, A. Thayananthan, P.H.S. Torr, R. Cipolla, Hierarchical part-based human body pose estimation, in: British Machine Vision Conference, Oxford, UK, Sep 2005.
[266] P. Nillius, J. Sullivan, S. Carlsson, Multi-target tracking—linking identities using baysian network inference, in: Computer Vision and Pattern Recognition, New York City, New York, USA, June 17–22, 2006.
[267] H. Ning, T. Han, Y. Hu, Z. Zhang, U. Fu, T. Huang, A realtime shrug detector, in: International Conference on Automatic Face and Gesture Recognition, Southampton, UK, April 10–12, 2006.
[268] K. Ogaki, Y. Iwai, M. Yachida, Posture estimation based on motion and structure models, Systems and Computers in Japan 32 (4) (2001).
[269] K. Okuma, A. Taleghani, N.D. Freitas, J.J. Little, David G. Lowe, A boosted particle filter: multitarget detection and tracking, in: European Conference on Computer Vision, Prague, Czech Republic, May 11–14, 2004.
[270] N. Oliver, B. Rosario, A. Pentland, A Bayesian computer vision system for modeling human interactions, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (8) (2000) 831–843.
[271] E.-J. Ong, A. Hilton, Learnt inverse kinematics for animation synthesis, in: Conference on Vision, Video and Graphics, Edinburgh, UK, June 2005.
[272] E.-J. Ong, A. Hilton, A.S. Micilotta, Viewpoint invariant exemplar-based 3D human tracking, in: First IEEE Workshop on Modeling People and Human Interaction (PHI'05), Beijing, China, Oct 15, 2005.
[273] D. Ormoneit, M.J. Black, T. Hastie, H. Kjellström, Representing cyclic human motion using functional analysis, Image and Vision Computing 23 (14) (2005) 1264–1276.
[274] D. Ormoneit, H. Sidenbladh, M.J. Black, T. Hastie, Learning and tracking cyclic human motion, in: Workshop on Human Modeling, Analysis and Synthesis at CVPR, Hilton Head Island, South Carolina, June 13–15, 2000.
[275] I.B. Ozer, W.H. Wolf, A hierarchical human detection system in (Un) compressed domains, IEEE Transactions on Multimedia 4 (2) (2002) 283–300.
[276] C. Pan, S. Ma, Parametric tracking of human contour by exploiting intelligent edge, in: International Conference on Automatic Face and Gesture Recognition, Seoul, Korea, May 17–19, 2004.
[277] V. Parameswaran, R. Chellappa, View invariants for human action recognition, in: Computer Vision and Pattern Recognition, Madison, Wisconsin, June 16–22, 2003.
[278] V. Parameswaran, R. Chellappa, View independent human body pose estimation from a single perspective image, in: Computer Vision and Pattern Recognition, Washington DC, USA, June 2004.
[279] V. Parameswaran, R. Chellappa, View invariance for human action recognition, International Journal of Computer Vision 66 (1) (2006) 83–101.
[280] S. Park, J.K. Aggarwal, Segmentation and tracking of interacting human body parts under occlusion and shadowing, in: Workshop on Motion and Video Computing, Orlando, Florida, Nov 7, 2002.
[281] S. Park, J.K. Aggarwal, Semantic-level understanding of human actions and interactions using event hierarchy, in: CVPR Workshop on Articulated and Non-Rigid Motion, Washington DC, USA, June 2004.
[282] S. Park, J.K. Aggarwal, Simultaneous tracking of multiple body parts of interacting persons, Computer Vision and Image Understanding 102 (1) (2006) 1–21.
[283] S. Park, M. Trivedi, A Track-based human movement analysis and privacy protection system adaptive to environmental contexts, in: Asian Conference on Computer Vision, LNCS 3852, Hyderabad, India, Jan 13–16, 2006.
[284] A.E.C. Pece, Tracking og non-Gaussian clusters in the PETS2001 image sequences, in: Workshop on Performance Evaluation of Tracking and Surveillance, Kauai, Hawaii, Dec 9, 2001.
[285] A.E.C. Pece, From cluster tracking to people counting, in: Workshop on Performance Evaluation of Tracking and Surveillance, Copenhagen, Denmark, June 1, 2002.
[286] J. Pers, M. Bon, S. Kovacic, M. Sibila, B. Dezman, Observation and analysis of large-scale human motion, Human Movement Science 21 (2002) 295–311.
[287] R. Plänkers, P. Fua, Tracking and modeling people in video sequences, Computer Vision and Image Understanding 81 (3) (2001) 285–303.
[288] R. Plänkers, P. Fua, Model-based silhouette extraction for accurate people tracking, in: European Conference on Computer Vision, Copenhagen, Denmark, June 2002.
[289] R. Plänkers, P. Fua, Articulated soft objects for multiview shape and motion capture, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (9) (2003) 1182–1187.
[290] E. Polat, M. Yeasin, R. Sharma, Robust tracking of human body parts for collaborative human computer interaction, Computer Vision and Image Understanding 89 (2003) 44–69.
[291] F. Porikli, Trajectory distance metric using hidden Markov model based representation, in: 6th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, Prague, May 11–14, 2004.
[292] A. Prati, I. Mikić, R. Cucchiara, M.M. Trivedi, Analysis and detection of shadows in video streams: a comparative evaluation, in: Computer Vision and Pattern Recognition Conference, Hawaii, USA, Dec 2001.
[293] A. Prati, I. Mikić, M.M. Trivedi, R. Cucchiara, Detecting moving shadows: algorithms and evaluation, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (7) (2003) 918–923.
[294] R. Rosales, M. Siddiqui, J. Alon, S. Sclaroff, Estimating 3D Body pose using uncalibrated cameras, in: Computer Vision and Pattern Recognition, Kauai Marriott, Hawaii, Dec 9–14, 2001.
[295] M.M. Rahman, A. Robles-Kelly, A tuned eigenspace technique for articulated motion recognition, in: European Conference on Computer Vision, Graz, Austria, May 7–13, 2006.
[296] D. Ramanan, D.A. Forsyth, A. Zisserman, Strike a pose: tracking people by finding stylized poses, in: Computer Vision and Pattern Recognition, San Diego, California, USA, June 20–25, 2005.

[297] D. Ramanan, C. Sminchisescu, Tranining deformable models for localization, in: Computer Vision and Pattern Recognition, New York City, New York, USA, June 17–22, 2006.
[298] C. Rao, A. Yilmaz, M. Shah, View-invariant representation and recognition of actions, Journal of Computer Vision 50 (2) (2002) 203–226.
[299] F. Remondino, 3-D reconstruction of static human body shape from image sequence, Computer Vision and Image Understanding 93 (2004) 65–85.
[300] H. Ren, G. Xu, Human action recognition with primitive-based coupled-HMM, in: International Conference on Pattern Recognition, Quebec, Canada, Aug 11–15, 2002.
[301] H. Ren, G. Xu, S. Kee, Subject-independent natural action recognition, in: International Conference on Automatic Face and Gesture Recognition, Seoul, Korea, May 17–19, 2004.
[302] X. Ren, A.C. Berg, J. Malik, Recovering human body configurations using pairwise constraints between parts, in: International Conference on Computer Vision, Beijing, China, Oct 15–21, 2005.
[303] L. Reng, T.B. Moeslund, E. Granum, Finding motion primitives in human body gestures, in: S. Gibet, N. Courty, J.-F. Kamps (Eds.), GW 2005, LNAI, vol. 3881, Springer, Berlin, Heidelberg, 2006, pp. 133–144.
[304] Y. Ricquebourg, P. Bouthemy, Real-time tracking of moving persons by exploiting spatio-temporal image slices, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (8) (2000) 797–808.
[305] J. Rittscher, A. Blake, S.J. Roberts, Towards the automatic analysis of complex human body motions, Image and Vision Computing 20 (2002) 905–916.
[306] I. Rius, J. Varona, X. Roca, J. Gonzàlez, Posture constraints for bayesian human motion tracking, in: Conference on Articulated Motion and Deformable Objects (AMDO), Andratx, Mallorca, Spain, July 11–14, 2006.
[307] G. Rizzolatti, L. Fogassi, V. Gallese, Parietal cortex: from sight to action, Current Opinion in Neurobiology 7 (1997) 562–567.
[308] G. Rizzolatti, L. Fogassi, V. Gallese, Neurophysiological mechanisms underlying the understanding and imitation of action, Nature Reviews 2 (2001) 661–670.
[309] T.J. Roberts, S.J. McKenna, I.W. Ricketts, Adaptive learning of statistical appearance models for 3D human tracking, in: British Machine Vision Conference, Cardiff, UK, 2002.
[310] T.J. Roberts, S.J. McKenna, I.W. Ricketts, Human pose estimation using learnt probabilistic region similarities and partial configurations, in: European Conference on Computer Vision, Prague, Czech Republic, May 11–14, 2004.
[311] N. Robertson, I. Reid, Behaviour understanding in video: a combined method, in: Internatinal Conference on Computer Vision, Beijing, China, Oct 15–21, 2005.
[312] M. Roh, B. Christmas, J. Kittler, S. Lee, Robust player gesture spotting and recognition in low-resolution sports video, in: European Conference on Computer Vision, Graz, Austria, May 7–13, 2006.
[313] M.-C. Roh, B. Christmas, J. Kittler, S.-W. Lee, Robust player gesture spotting and recognition in low-resolution sports video, in: European Conference on Computer Vision, Graz, Austria, May 7–13, 2006.
[314] R. Ronfard, C. Schmid, B. Triggs, Learning to parse pictures of people, in: European Conference on Computer Vision, Copenhagen, Denmark, June 27–31, 2002.
[315] R. Rosales, S. Sclaroff, Learning and synthesizing human body motion and posture, in: The fourth International Conference on Automatic Face and Gesture Recognition, Grenoble, France, March 2000.
[316] D. Roth, P. Doubek, L.V. Gool, Bayesian pixel classification for human tracking, in: IEEE Workshop on Motion and Video Computing (MOTION'05), Breckenridge, Colorado, Jan 2005.
[317] S. Roweis, L. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (2000) 2323–2327.
[318] M.S. Ryoo, J.K. Aggarwal, Recognition of composite human activities through context-free grammer based representation, in: Computer Vision and Pattern Recognition, New York City, New York, USA, June 17–22, 2006.
[319] A. Sanfeliu, J.J. Villanueva, An approach of visual motion analysis, Pattern Recognition Letters 26 (3) (2005) 355–368.
[320] P. Sangi, J. Heikkilä, O. Silven, Extracting motion components from image sequences using particle filters, in: The 12th Scandinavian Conference on Image Analysis, Bergen, Norway, 2001.
[321] K. Sato, J.K. Aggarwal, Tracking and recognizing two-person interactions in outdoor image sequences, in: Workshop on Multi-Object Tracking, Vancouver, Canada, July 8, 2001.
[322] S. Schaal, Is imitation learning the route to humanoid robots? Trends in Cognitive Sciences 3 (6) (1999) 233–242.
[323] K. Schindler, H. Wang, Smooth foreground-background segmentation for video processing, in: Asian Conference on Computer Vision, LNCS 3852, Hyderabad, India, Jan 13–16, 2006.
[324] G. Schwarz, Estimating the dimension of a model, Annals of Statistics 6 (1978) 461–464.
[325] M. Shah, Understanding human behavior from motion imagery, Machine Vision and Applications 14 (4) (2003) 210–214.
[326] G. Shakhnarovich, P. Viola, T. Darrell, Fast pose estimation with parameter-sensitive hashing, in: International Conference on Computer Vision, Nice, France, Oct 13–16, 2003.
[327] Y. Sheikh, M. Shah, Bayesian modelling of dynamic scenes for object detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (11) (2005) 1778–1792.
[328] Y. Sheikh, M. Sheikh, M. Shah, Exploring the space of human action, in: Internatinal Conference on Computer Vision, Beijing, China, Oct 15–21, 2005.
[329] J. Shi, C. Tomasi, Good features to track, in: Computer Vision and Pattern Recognition, Seattle, Washington, June 1994.
[330] Y. Shi, A. Bobick, I. Essa, Learning temporal sequence model from partially labeled data, in: Computer Vision and Pattern Recognition, New York City, New York, USA, June 17–22, 2006.
[331] H. Sidenbladh, Detecting human motion with support vector machines, in: International Conference on Pattern Recognition, Cambridge, UK, Aug 2004.
[332] H. Sidenbladh, M.J. Black, Learning image statistics for Bayesian tracking, in: International Conference on Computer Vision, Vancouver, Canada, July 9–12, 2001.
[333] H. Sidenbladh, M.J. Black, Learning the statistics of people in images and video, International Journal of Computer Vision 54 (1/2/3) (2003) 183–209.
[334] H. Sidenbladh, M.J. Black, D.J. Fleet, Stochastic tracking of 3D human figures using 2D image motion, in: European Conference on Computer Vision, Dublin, Ireland, June 2000.
[335] H. Sidenbladh, M.J. Black, L. Sigal, Implicit probabilistic models of human motion for synthesis and tracking, in: European Conference on Computer Vision, Copenhagen, Denmark, 2002.
[336] L. Sigal, S. Bhatia, S. Roth, M.J. Black, M. Isard, Tracking loose-limbed people, in: Computer Vision and Pattern Recognition, Washington DC, USA, June 2004.
[337] L. Sigal, M.J. Black, Measure locally, reason globally: occlusion-sensitive articulated pose estimation, in: Computer Vision and Pattern Recognition, New York City, New York, USA, June 17–22, 2006.
[338] L. Sigal, M.J. Black, Predicting 3D people from 2D pictures, in: International Conference on Articulated Motion and Deformable Objects, Springer LNCS 4069, Andratx, Mallorca, Spain, July 11–14, 2006.
[339] C. Sminchisescu, Consistency and coupling in human model likelihoods, in: International Conference on Automatic Face and Gesture Recognition, Washington DC, USA, May 20–21, 2002.
[340] C. Sminchisescu, A. Kanaujia, Z. Li, D. Metaxas, Discriminative density propagation for 3D human motion estimation, in: Computer Vision and Pattern Recognition, San Diego, California, USA, June 20–25, 2005.

[341] C. Sminchisescu, A. Kanaujia, D. Metaxas, Learning joint top-down and bottom-up processes for 3D visual inference, in: Computer Vision and Pattern Recognition, New York City, New York, USA, June 17–22, 2006.

[342] C. Sminchisescu, B. Triggs, Covarinace scaled sampling for monocular 3D body tracking, in: Computer Vision and Pattern Recognition, Kauai Marriott, Hawaii, Dec 9–14, 2001.

[343] C. Sminchisescu, B. Triggs, Estimating articulated human motion with covariance scaled sampling, International Journal of Robotics Research 22 (6) (2003) 371–391.

[344] C. Sminchisescu, B. Triggs, Kinematic jump processes for monocular 3D human tracking, in: Computer Vision and Pattern Recognition, Madison, Wisconsin, USA, June 16–22, 2003.

[345] K. Smith, D.G. Perez, J.M. Odobez, Using particles to track varying numbers of interacting people, in: Computer Vision and Pattern Recognition, San Diego, CA, June 20–25, 2005.

[346] P. Smith, N. da Vitoria Lombo, M. Shah, Temporal boost for event recognition, in: Internatinal Conference on Computer Vision, Beijing, China, Oct 15–21, 2005.

[347] P. Smith, M. Shah, N. da Vitoria Lobo, Integrating mutiple levels of zoom to enable activity analysis, Computer Vision and Image Understanding 103 (2006) 33–51.

[348] Y. Song, L. Goncalves, E.D. Bernardo, Monocular perception of biological motion in Johansson display, Computer Vision and Image Understanding 81 (3) (2001) 303–327.

[349] Y. Song, L. Goncalves, P. Perona, Learning probabilistic structure for human motion detection, in: Computer Vision and Pattern Recognition, Kauai Marriott, Hawaii, Dec 9–14, 2001.

[350] Y. Song, L. Goncalves, P. Perona, Unsupervised learning of human motion, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (7) (2003) 814–827.

[351] J. Starck, G. Collins, R. Smith, A. Hilton, J. Illingworth, Animated statues, Journal of Machine Vision and Applications 14 (4) (2003) 248–259.

[352] J. Starck, A. Hilton, Model-based multiple view reconstruction of people, in: International Conference on Computer Vision, Nice, France, Oct 13–16, 2003.

[353] J. Starck, A. Hilton, Spherical matching for temporal correspondence of non-rigid surfaces, in: ICCV, Beijing, China, Oct 15–21, 2005.

[354] C. Stauffer, W.E.L. Grimson, Adaptive background mixture models for real-time tracking, in: Computer Vision and Pattern Recognition, Santa Barbara, CA, June 1998.

[355] C. Stauffer, W.E.L. Grimson, Learning patterns of activity using real-time tracking, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (8) (2000) 747–757.

[356] A. Stolcke, An efficient probabilistic context-free parsing algorithm that computes prefix probabilities, Computational Linguistics 21 (2) (1995) 165–201.

[357] M. Störring, T. Kocka, H.J. Andersen, E. Granum, Tracking regions of human skin through illumination changes, Pattern Recognition Letters 24 (2003) 1715–1723.

[358] A. Sundaresan, R. Chellappa, Acquisition of articulated human body models using multiple cameras, in: International Conference on Articulated Motion and Deformable Objects, Andratx, Mallorca, Spain, Springer Lecturer Notes in Computer Science 4069, July 11–14, 2006.

[359] K. Takahashi, T. Sakaguchi, J. Ohya, Remarks on a real-time, noncontact, nonwear, 3D human body posture estimation method, Systems and Computers in Japan 31 (14) (2000) 1–10.

[360] L. Taycher, G. Shakhnarovich, D. Demirdjian, T. Darrell, Conditional random people: tracking humans with CRFs and grid filters, in: Computer Vision and Pattern Recognition, New York City, New York, USA, June 17–22, 2006.

[361] C.J. Taylor, Reconstruction of articulated objects from point correspondences in a single image, Computer Vision and Image Understanding 80 (3) (2000) 349–363.

[362] J. Tenenbaum, V. de Silva, J. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (2000) 2319–2323.

[363] M.N. Thalmann, H. Seo, Data-driven approaches to digital human modeling, in: International Symposium on 3D Data Processing, Visualization, and Transmission, Thessalonica, Greece, Sep 2004.

[364] C. Theobalt, M. Magnor, P. Schuler, H.-P. Seidel, Combining 2D feature tracking and volume reconstruction for online video-based human motion capture, in: Pacific Conference on Computer Graphics and Applications, Tsinghua University, Beijing, China, Oct 8–11, 2002.

[365] C. Tomasi, T. Kanade, Shape and motion from image streams under orthography: a factorization method, International Journal of Computer Vision 9 (1992) 137–154.

[366] S. Toyosawa, T. Kawai, Crowdedness estimation in public pedestrian space for pedestrian guidance, in: Conference on Intelligent Transportation Systems, Vienna, Austria, Sep 13–16, 2005.

[367] M. Trivedi, K. Huang, I. Mikić, Intelligent environments and active camera networks, in: Conference on System, Man and Cybernetics, Nashville, Tennessee, Oct 8–11, 2000.

[368] M.M. Trivedi, I. Mikić, S.K. Bhonsle, Active camera networks and semantic event databases for intelligent environments, in: Workshop on Human Modeling, Analysis and Synthesis at CVPR, Hilton Head Island, South Carolina, June 13–15, 2000.

[369] N. Ukita, T. Matsuyama, Real-time cooperative multi-target tracking by communicating active vision agents, Computer Vision and Image Understanding 97 (2) (2005) 137–179.

[370] R. Urtasun, D.J. Fleet, P. Fua, Monocular 3-D tracking of the golf swing, in: Computer Vision and Pattern Recognition, San Diego, California, USA, June 20–25, 2005.

[371] R. Urtasun, D.J. Fleet, P. Fua, 3D people tracking with gaussian process dynamical models, in: Computer Vision and Pattern Recognition, New York City, New York, USA, June 17–22, 2006.

[372] R. Urtasun, D.J. Fleet, A. Hertzmann, P. Fua, Priors for people tracking from small training sets, in: International Conference on Computer Vision, Beijing, China, Oct 15–21, 2005.

[373] R. Urtasun, P. Fua, 3D human body tracking using deterministic temporal motion models, in: European Conference on Computer Vision, Prague, Czech Republic, May 11–14, 2004.

[374] A. Utsumi, N. Tetsutani, Human detection using geometrical pixel value structures, in: International Conference on Automatic Face and Gesture Recognition, Washington DC, USA, May 20–21, 2002.

[375] N. Vasvani, A. Roy Chowdhury, R. Chellappa, Activity recognition using the dynamics of the configuration of interacting objects, in: Computer Vision and Pattern Recognition, Madison, Wisconsin, USA, June 16–22, 2003.

[376] D.D. Vecchio, R.M. Murray, P. Perona, Decomposition of human motion into dynamics-based primitives with application to drawing tasks, Automatica 39 (12) (2003) 2085–2098.

[377] A. Veeraraghavan, R. Chellappa, A.K. Roy-Chowdhury, The function space of an activity, in: Computer Vision and Pattern Recognition, New York City, New York, USA, June 17–22, 2006.

[378] A. Veeraraghavan, A.K. Roy-Chowdhury, R. Chellappa, Matching shape sequences in video with applications in human movement analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (12) (2005) 1896–1909.

[379] A. Veeraranghavan, R. Chellappa, A. Roy-Chowdhury, The function space of an activity, in: Computer Vision and Pattern Recognition, New York City, New York, USA, June 17–22, 2006.

[380] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: Computer Vision and Pattern Recognition, Kauai Marriott, Hawaii, Dec 9–14, 2001.

[381] P. Viola, M.J. Jones, D. Snow, Detecting pedestrians using patterns of motion and appearance, in: International Conference on Computer Vision, Nice, France, Oct 13–16, 2003.

[382] P. Viola, M.J. Jones, D. Snow, Detecting pedestrians using patterns of motion and appearance, International Journal of Computer Vision 63 (2) (2005) 153–161.

[383] S. Wachter, H.-H. Nagel, Tracking of persons in monocular image sequences, in: Workshop on Motion of Non-Rigid and Articulated Objects, Puerto Rico, USA, 1997.
[384] S. Wachter, H.-H. Nagel, Tracking persons in monocular image sequences, Computer Vision and Image Understanding 74 (3) (1999) 174–192.
[385] H. Wang, D. Suter, Background initialization with a new robust statistical approach, in: International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, Beijing, China, Oct 15–16, 2005.
[386] H. Wang, D. Suter, A novel robust statistical method for background initialization and visual surveillance, in: Asian Conference on Computer Vision, Hyderabad, India, Lecturer Notes in Computer Science 3851, Jan 13–16, 2006.
[387] J. Wang, B. Bodenheimer, An evaluation of a cost metric for selecting transitions between motion segments, in: SIGGRAPH Symposium on Computer Animation, 2003.
[388] L. Wang, W. Hu, T. Tan, Recent development in human motion analysis, Pattern Recognition 36 (3) (2003) 585–601.
[389] L. Wang, H. Ning, T. Tan, W. Hu, Fusion of static and dynamic body biometrics for gait recognition, in: International Conference on Computer Vision, Nice, France, Oct 13–16, 2003.
[390] L. Wang, T. Tan, H. Ning, W. Hu, Silhouette analysis-based gait recognition for human identification, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (12) (2003) 1505–1518.
[391] Y. Wang, G. Baciu, Human motion estimation from monocular image sequence based on cross-entropy regularization, Pattern Recognition Letters 24 (1–3) (2003).
[392] Y. Wang, H. Jiang, M. Drew, Z. Li, G. Mori, Unsupervised discovery of action classes, in: Computer Vision and Pattern Recognition, New York City, New York, USA, June 17–22, 2006.
[393] D. Weinberg, R. Ronfard, E. Boyer, Motion history volumes for free viewpoint action recognition, in: Workshop on Modeling People and Human Interaction (PHI'05), Beijing, China, Oct 15, 2005.
[394] C.R. Wren, A. Azarbayejani, T. Darrell, A.P. Pentland, Pfinder: real-time tracking of the human body, Transactions on Pattern Analysis and Machine Intelligence 19 (7) (1997) 780–785.
[395] B. Wu, R. Navatia, Tracking of multiple, partially occluded humans based on static body part detection, in: Computer Vision and Pattern Recognition, New York City, New York, USA, June 17–22, 2006.
[396] B. Wu, R. Nevatia, Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detection, in: International Conference on Computer Vision, Beijing, China, Oct 15–21, 2005.
[397] Q.Z. Wu, H.Y. Cheng, B.S. Jeng, Motion Detection via change-point detection for cumulative histograms of ratio images, Pattern Recognition Letters 26 (5) (2004) 555–563.
[398] Y. Wu, G. Hua, T. Yu, Tracking articulated body by dynamic markov network, in: International Conference on Computer Vision, Nice, France, Oct 13–16, 2003.
[399] Y. Wu, T. Yu, A field model for human detection and tracking, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (5) (2006) 753–765.
[400] T. Xiang, S. Gong, Beyond tracking: modelling action and understanding behavior, International Journal of Computer Vision 67 (1) (2006) 21–51.
[401] L.Q. Xu, P. Puig, A hybrid blob- and appearance-based framework for multi-object tracking through complex occlusions, in: International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, Beijing, China, Oct 15–16, 2005.
[402] C. Yam, M. Nixon, J. Carter, On the relationship of human walking and running: automatic person identification by gait, in: International Conference on Pattern Recognition, Quebec, Canada, Aug 11–15, 2002.
[403] M. Yamamoto, H. Mitomi, F. Fujiwara, T. Sato, Bayesian classification of task-oriented actions based on stochastic context-free grammar, in: International Conference on Automatic Face and Gesture Recognition, Southampton, UK, April 10–12, 2006.
[404] C. Yang, R. Duraiswami, L. Davis, Fast multiple object tracking via a hierarchical particle filter, in: International Conference on Computer Vision, Beijing, China, Oct 15–21, 2005.
[405] D.B. Yang, H.H.G. Banos, L.J. Guibas, Counting people in crowds with a real-time network of simple image sensors, in: International Conference on Computer Vision, Nice, France, Oct 13–16, 2003.
[406] H.D. Yang, S.W. Lee, Multiple pedestrian detection and tracking based on weighted temporal texture features, in: International Conference on Pattern Recognition, Cambridge, UK, Aug 23–26, 2004.
[407] M.T. Yang, Y.C. Shih, S.C. Wang, People tracking by integrating multiple features, in: International Conference on Pattern Recognition, Cambridge, UK, Aug 23–26, 2004.
[408] T. Yang, S.Z. Li, Q. Pan, J. Li, Real-time multiple objects tracking with occlusion handling in dynamic scenes, in: Computer Vision and Pattern Recognition, San Diego, CA, June 20–25, 2005.
[409] H. Yi, D. Rajan, L.-T. Chia, A new motion histogram to index motion content in video segments, Pattern Recognition Letters 26 (2004) 1221–1231.
[410] A. Yilmaz, M. Shah, Actions sketch: a novel action representation, in: Computer Vision and Pattern Recognition, San Diego, California, USA, June 20–25, 2005.
[411] A. Yilmaz, M. Shah, Recognizing human actions in videos acquired by uncalibrated moving cameras, in: Internatinal Conference on Computer Vision, Beijing, China, Oct 15–21, 2005.
[412] H. Yu, G.-M. Sun, W.-X. Song, X. Li, Human motion recognition based on neural networks, in: International Conference on Communications, Circuits and Systems, Hong Kong, China, May 2005.
[413] X. Yu, S.X. Yang, A study of motion recognition from video sequences, Computing and Visualization in Science 8 (2005) 19–25.
[414] J. Zhang, R. Collins, Y. Liu, Bayesian body localization using mixture of nonlinear shape models, in: International Conference on Computer Vision, Beijing, China, Oct 15–21, 2005.
[415] J. Zhao, L. Li, K.C. Keong, 3D posture reconstruction and human animation from 2D feature points, Computer Graphics Forum 24 (4) (2005) 759–771.
[416] L. Zhao, L.S. Davis, Closely coupled object detection and segmentation, in: International Conference on Computer Vision, Beijing, China, Oct 15–21, 2005.
[417] L. Zhao, C.E. Thorpe, Stereo- and neural network-based pedestrian detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 1 (3) (2000) 148–154.
[418] T. Zhao, R. Nevatia, Stochastic human segmentation from a static camera, in: Workshop on Motion and Video Computing, Orlando, Florida, Nov 7, 2002.
[419] T. Zhao, R. Nevatia, Bayesian human segmentation in croweded situations, in: Computer Vision and Pattern Recognition, Madison, Wisconsin, June 16–22, 2003.
[420] T. Zhao, R. Nevatia, Tracking multiple humans in complex situations, IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (9) (2004) 1208–1221.
[421] T. Zhao, R. Nevatia, Tracking multiple humans in crowded environments, in: Computer Vision and Pattern Recognition, Washington DC, USA, June 2004.
[422] T. Zhao, R. Nevatia, F. Lv, Segmentation and tracking of multiple humans in complex situations, in: Computer Vision and Pattern Recognition, Kauai Marriott, Hawaii, Dec 9–14, 2001.
[423] J. Zhong, S. Sclaroff, Segmentating foreground objects from a dynamic textured background via robust kalman filter, in: International Conference on Computer Vision, Nice, France, Oct 13–16, 2003.
[424] Z. Zivkovic, Improved adaptive Gausian mixture model for background subtraction, in: International Conference on Pattern Recognition, Cambridge, UK, Aug 2004.