Pattern Recognition 79 (2018) 55–64

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Articulated motion and deformable objects

Jun Wan[a,*], Sergio Escalera[b], Francisco J. Perales[c], Josef Kittler[d]

[a] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, China
[b] University of Barcelona and Computer Vision Center, Spain
[c] Computer Graphics, Vision and Artificial Intelligence Group, Universitat de les Illes Balears, Spain
[d] Centre for Vision, Speech and Signal Processing, University of Surrey, UK

ARTICLE INFO

Article history:
Available online 1 February 2018

Keywords:
Articulated motion and deformable Objects
Pose estimation
Action recognition
Gesture recognition
Face analysis

ABSTRACT

This guest editorial introduces the twenty two papers accepted for this Special Issue on Articulated Motion and Deformable Objects (AMDO). They are grouped into four main categories within the field of AMDO: human motion analysis (action/gesture), human pose estimation, deformable shape segmentation, and face analysis. For each of the four topics, a survey of the recent developments in the field is presented. The accepted papers are briefly introduced in the context of this survey. They contribute novel methods, algorithms with improved performance as measured on benchmarking datasets, as well as two new datasets for hand action detection and human posture analysis. The special issue should be of high relevance to the reader interested in AMDO recognition and promote future research directions in the field.

1. Introduction

Articulated motion and deformable objects (AMDO) is a challenging research area which focuses on the automatic analysis of complex objects, such as the human body, exhibiting high variabilities both in terms of spatial and temporal dimensions. AMDO is of high interest in the fields of pattern recognition, computer vision, computer graphics, biometrics, machine learning and human-computer interface (HCI), to mention just a few.

In the late 2016, contributions to a special issue on AMDO had been invited for possible publication in the Patter Recognition journal by an open call for papers. The scope of the special issue had been defined so as to cover pattern recognition schemes on any AMDO related topics, including human motion analysis and tracking, human reconstruction, multimodal AMDO, 2D/3D deformable models, and new pattern recognition applications in the field of AMDO. All 48 manuscripts submitted to this SI were subject to the same rigurious review process assessing their overall quality and significance. A total of 22 papers were accepted for publication in this special issue.

The rest of this guest editorial article is organized as follows. Section 2 provides a brief review of the state of the art in four AMDO subtopics, namely, human motion analysis, human pose estimation, deformable segmentation, and face analysis. Section 3 summarises the papers accepted for the Special Issue. We conclude with a brief outlook to the future in Section 4.

2. Articulated motion and deformable objects

We structure the review of the AMDO literature into four main topics, namely, human motion analysis, human pose estimation, deformable shape segmentation, and facial analysis. For each subtopic, we review the commonly used benchmark datasets and the main state-of-the-art methods published in the last two years. The papers published in this special issue are referred to by their unique number, e.g. SI_1, SI_2.

2.1. Human motion analysis

The two main topics of human motion analysis are action and gesture recognition. Recently, the use of multimodal data in the context of human motion analysis has received a lot of attention in the literature. RGB, depth, and skeletal information are the commonly considered modalities for multimodal action and gesture recognition. In this section, first, we provide a brief introduction to action and gesture recognition datasets. Then we review the state-of-the-art methods in this topic area.

2.1.1. Benchmark datasets for action and gesture recognition

Table 1 shows the commonly used datasets for isolated or continuous action and gesture recognition, and summarises the key statistical attributes of these datasets, namely data modality, number of classes, number of subjects, and number of samples. Each

* Corresponding author.
E-mail address: jun.wan@ia.ac.cn (J. Wan).

https://doi.org/10.1016/j.patcog.2018.01.038
0031-3203/

Table 1
Statistics of popular benchmark datasets for human motion analysis in the field of AMDO. All datasets are used at least once by accepted manuscripts in this special issue.

index	Dataset	Year	Modality	#Class	#Subjects	#Samples
1	MSR Action3D, [1]	2010	RGB skeleton	20	10	567
2	MSRC-12, [2]	2012	skeleton	12	30	594
3	MSR Daily Activity 3D, [3]	2012	RGB-D skeleton	16	10	320
4	UTKinect, [4]	2012	RGB-D skeleton	10	10	200
5	SBU Kinect, Interaction [5]	2012	RGB-D skeleton	7	8	300
6	NTU RGB+D, [6]	2016	RGB-D skeleton	60	40	56880
7	DHG-14/28, [7]	2016	RGB-D skeleton	14	20	2800
8	Montalbano V2, [8]	2014	Depth skeleton	20	27	13858
9	MIVIA action, [9]	2014	RGB-D	7	14	500
10	NATOPS gesture, [10]	2011	RGB-D	24	20	9600

dataset is used at least once in the manuscripts accepted for this special issue.

MSR Action3D [1]. It includes 567 sequences , including twenty actions (namely, high arm waving, horizontal arm waving, hammering, hand catching, forward punching, high throwing, drawing x, drawing a tick, drawing a circle, hand clapping, two hand waving, side-boxing, bending, forward kicking, side kicking, jogging, tennis serve, golf swing, pickup and throw), each action being performed three times by ten subjects. The data is recorded from a fixed point of view while the subjects are facing the camera.

MSRC-12 [2]. It comprises thirty subjects performing twelve gestures. These gestures are grouped into two categories: iconic and metaphoric gestures. The iconic gestures directly correspond to real world actions and represent first people shooter (FPS) gaming actions. It contains six FPS gaming actions: crouching, shooting, throwing, using night goggles, changing weapon and kicking.

MSR DailyActivity3D [3]. This dataset contains sixteen daily human activities in a living room: drinking, eating, reading a book, calling a cellphone, writing on a paper, using a laptop, using a vacuum cleaner, cheering up, sitting still, tossing paper, playing a game, laying down on a sofa, walking, playing a guitar, standing up, sitting down. Ten subjects are recorded performing these actions while sitting on the sofa or standing close to the sofa. The camera is fixed in front of the sofa. The dataset also provides depth and skeleton data.

UTKinect [4]. Ten types of human action are recorded twice by ten subjects. The actions include walking, sitting down, standing up, picking up, carrying, throwing, pushing, pulling, waving, clapping hands. The actions were recorded from a variety of views. The dataset is composed of 200 sequences, recording RGB-D data and skeleton joint locations.

SBU Kinect Interaction [5]. It has eight classes which are commonly used in two-person interactions, namely, approaching, departing, pushing, kicking, punching, exchanging objects, hugging, and shaking hands. This dataset is challenging because of the similarity of some actions in terms of motion (e.g., exchanging object and shaking hands). RGB and depth video with 15 frames per second are provided, with an image resolution of 640×480.

NTU RGB + D [6]. It is a large scale dataset for human action recognition, which consists of 56,880 action samples with four different data modalities for each sample: RGB videos, depth map sequences, 3D skeletal data, and infrared videos. It has 60 classes in total, which are divided into three major groups: 40 daily actions (i.e., drinking, eating, etc.), 9 health-related action (i.e., sneezing, staggering, etc.), and 11 interactions (i.e., punching, kicking, etc.).

DHG-14/28 [7]. The dynamic hand gesture dataset DHG-14/28 has fourteen gesture classes. Each one is executed five times by twenty participants in two ways, resulting in 2800 sequences. Gestures are subdivided into the categories of fine and coarse£ºGrab (G, fine), Tap (T, coarse), Expand (E, fine), Pinch (P, fine), Rotation CW (R-CW, fine), Swipe Right (S-R, coarse), Swipe Left (S-L, coarse), Swipe Up (S-U, coarse), Swipe Down (S-D, coarse), Swipe X (S-X, coarse), Swipe V (S-V, coarse), Swipe + (S-+, coarse), Shake (Sh, coarse).

Montalbano V2 [8]. This dataset was released for the ChaLearn Looking at People Challenge 2014. It contains about 14,000 samples from a vocabulary of twenty Italian sign gesture categories in continuous data series. It provides RGB, depth, mask body, and voice information for each sample.

MIVIA Action [9]. It consists of RGB-D videos of seven actions, namely, opening a jar, drinking, sleeping, random motion, stopping, interacting with a table, and sitting, performed by fourteen subjects.

NATOPS gesture [10]. It consists of 24 aircraft handling signals from the Naval Air Training and Operating Procedures Standardization (NATOPS) manual. These gestures are captured using a Kinect sensor at 20 FPS with a resolution of 320×240. The location of the skeletal joints in the upper body along with the hand sign are available with the dataset. These 24 upper body gestures were performed by 20 subjects for 20 times, resulting in 400 observations for each (subject, gesture) pair.

2.1.2. The state-of-the-art in human motion analysis

Table 2 shows a comparison of different state-of-the-art methods on the ten datasets used by the papers on human motion analysis accepted for the SI. Three main tasks are considered, namely, action detection, isolated action/gesture recognition, and continuous gesture recognition. For different tasks, the evaluation metrics vary. They include accuracy for isolated action/gesture recognition, F1-score for action detection and Jaccard index (JI) for continuous gesture recognition. In Table 2, we show the performance of state-of-the-art methods in different datasets and provide the details of the specific evaluation protocols for each cited paper.

The listed methods in Table 2 can be grouped into two categories. The first category includes traditional methods for motion analysis, which consist of a pipeline commencing with feature ex-

Table 2
A Comparison of state-of-the-art methods on the benchmark action/gesture datasets listed in Table 1. All cited papers have been published in the last two years. Protocol A: half of the subjects are used for training (i.e. odd subjects) and the rest for testing (i.e. even subjects); Protocol B: average over all splits; Protocol C: leave-one-subject-out; Protocol D: training with the first 5 subjects, testing with other 10 subjects; acc: accuracy; JI: Jaccard Index.

Dataset	Method	Metrics		Protocol and notes
MSR Action3D [1]	CNN + LSTM [SI_1]	acc(%)	95.7/96.0	A/ B
	DSRF [SI_3]	acc(%)	95.24	no specific mention
	Bag of Gesturelets [SI_8]	acc(%)	96.49	A
	Pose and Kinematics [SI_{19}]	acc(%)	96.77	A
	MIMTL [11], 2017	acc(%)	96.37	A
	trajectorylet + exemplar-SVMs [12], 2017	acc(%)	97.9	A
	DMMs [13], 2016	acc(%)	90.5	A
	Riemannian Manifold [14], 2016	acc(%)	96.97	A
MSRC-12 [2]	DSRF [SI_3]	acc(%)	95.64/95.36	C/ A
	Bag of Gesturelets [SI_8]	F1-score	89.8	C
	Clustered Spatiotemporal Manifolds [SI_9]	F1-score	77.3	C
	Enhanced Skeleton [SI_{11}]	acc(%)	96.62	A
	Pose Lexicon [SI_{18}]	acc(%)	92.03	-
	Pose and Kinematics [SI_{19}]	acc(%)	91.20	C
	View Invariant Information + CNNs [15], 2017	acc(%)	96.62	A
	Trajectorylet + Exemplar-SVMs [12], 2017	acc(%)	94.9/95.1	A/ C
	Encoded Spatial-temporal Information + CNN [16], 2016	acc(%)	94.27	A
	Joint Trajectory Maps + CNNs [17], 2016	acc(%)	93.12	A
MSR Daily Activity 3D [3]	CNN + LSTM [SI_1]	acc(%)	63.1	no specific mention
	DSSCA SSLM [18], 2017	acc(%)	97.5	A
	Unsupervised training [19], 2017	acc(%)	86.9	no specific mention
	MFSK + BOW [20], 2016	acc(%)	95.7	C
UTKinect [4]	CNN + LSTM [SI_1]	acc(%)	99.0	C
	DSRF [SI_3]	acc(%)	97.85	A
	Geometric Feature + LSTM [21], 2017	acc(%)	95.96	A
	VLDA + LMNN + k-NN [22], 2017	acc(%)	98	C
	JSG (top-K RVJRDs) + JSGK [23], 2017	acc(%)	98.3	C
	Triplet motion + LBP [24], 2016	acc(%)	98.0	3-fold cross-validation
Kinect Interaction [5]	Motion Information + CNN [SI_5]	acc(%)	90.98	5-fold cross validation
	Geometric Feature + LSTM [21], 2017	acc(%)	99.02	5-fold cross validation
	SkeletonNet [25], 2017	acc(%)	93.47	5-fold cross validation
	Co-occurence feature + LSTM [26], 2016	acc(%)	90.41	5-fold cross validation
NTU RGB + D [6]	CNN + LSTM [SI_1]	acc(%)	67.5/76.21	cross-subject/cross-view
	Enhanced Skeleton [SI_{11}]	acc(%)	80.03/87.21	cross-subject/cross-view
	DSSCA SSLM [18], 2017	acc(%)	74.9	cross-subject
	Joint distance maps + CNN [27], 2017	acc(%)	76.2/82.3	cross-subject/cross-view
	Part-aware LSTM Network [6], 2016	acc(%)	62.03/70.27	cross-subject/cross-view
DHG-14/28 [7]	CNN + LSTM [SI_1]	acc(%)	85.6/81.1	C; acc of 14/28 gestures
	Geometric shape + fisher vector + SVM [7], 2016	acc(%)	83.07/79.14	C; acc of 14/28 gestures
Montalbano V2 [8]	CNN + LSTM [SI_1]	JI	79.15	no specific mention;
	Moddrop [28], 2016	JI	83	no specific mention;
MIVIA action [9]	Motion Information + CNN [SI_5]	acc(%)	93.37	C
	Key poses + CNN [29], 2017	acc(%)	93.37	C
	String kernel framework [30], 2016	acc(%)	95.4	C
NATOPS gesture [10]	Motion Information + CNN [SI_5]	acc(%)	72.58/86.58	D; top-1/top-2 accuracy
	Random forest [31], 2017	acc(%)	88.1	no specific mention

traction, through feature encoding, to classification. In order to extrtact efficient features from image sequences/videos, some heuristically designed descriptors are proposed, such as the Dual Square-Root Function (DSRF) descriptor in SI_3, Gesturelets in SI_8, to complement trajectorylet [12], mixed features around sparse keypoints (MFSK) [20], co-occurence feature [26], triplet motion and local binary pattern (LBP) [24]. For feature encoding, the representative methods include bag of visual words (BoVW) [20], vector of locally aggregated descriptors (VLAD) [22] and fisher vector (FV) [7]. For the decision-making stage, the popular classifiers applied to the datasets in the Table 2 include KNN [20,22], SVM [7,12] and random forest [31].

The second category comprises the deep learning based methods. It consists of Convolutional neural networks (CNNs), Long Short Term Memory networks (LSTM) and mixed architecture based approaches. CNN-based methods typically encode image sequences or skeletons as dynamic images that capture the spatio-temporal or skeleton-based motion information [15–17,27], and then apply CNN for image-based recognition. LSTM-based methods extract geometric or co-occurence features [21,26] from each frame and train a model which encodes the spatio-temporal information content. Some approaches combine CNN and LSTM to realise an action recognition capability. For example, SI_1 combines CNN and LSTM in a two-stage training strategy designed to optimise the parameters of a CNN+LSTM framework.

For human motion analysis, there are many other CNN-based methods which are not mentioned in this section. For a more detailed survey the reader is referred to [32–36].

2.2. Human Pose Estimation

2.2.1. Benchmark Datasets for Human Pose Estimation

Table 3 summarises the attributes of the human pose estimation dataset used for evaluation by some of the work presented in this special issue.

Table 3
Statistics of the HumanEva dataset used for benchmarking human pose estimation algorithms.

Index	Dataset	Year	Modality	#Class	#Subjects	#Samples
1	HumanEva [37]	2010	RGB,skeleton	6	4	40,000

Table 4
A comparison of the state-of-the-art human pose estimation methods evaluated on the HumanEva [37] dataset. The protocol used to obtain the results in this table involves sequences S1, S2, S3, for training and the validation sequences of all subjects for testing. The evaluation metric is the average joint error in centimeter (cm).

Method	Metrics:jont error (cm)		
	Walking	Jogging	Boxing
Invariants moments [*SI*_4]	Only qualitative results provided		
Marker-less Motion Capture [42], 2017	6.65	–	6.00
Trainable Fusion [41], 2017	2.44	–	–
SMP [40], 2016	3.81	3.68	–
SMPLify [38], 2016	7.72	–	8.28

HumanEva [37]. This dataset contains six classes and about 40K samples recorded by four subjects. It comprises synchronized images and motion capture data and is a standard benchmark for 3D human pose estimation. The output pose is a vector of 15 3D joint coordinates.

2.2.2. The state-of-the-art in human pose estimation

Table 4 shows some methods that use the HumanEva dataset. All the listed papers have been published within the last two years. In SI_4, the authors build virtual humans via a professional free and open-source 3D computer graphics software called Blender[1] and a free software enabling the creation of realistic 3d human makehuman[2] data. These avatars can be animated to simulate realistic actions based on the motion capture data. One of the main advantages is that one can automatically generate ground truth data. The software thus saves a lot of effort by avoiding manual data collection and annotation. However, as the authors said "due to either the lack of motion capture file for importation into the graphics software or the lack of 3D ground truth, this makes a quantitative evaluation and comparison on public datasets difficult". Therefore, only qualitative results on the HumanEva dataset are presented in SI_4.

Table 4 lists the recently published papers based on deep learning. Federica et al. [38] present a SMPLify framework that falls within the classical paradigm of bottom up estimation followed by top down verification (generative model). For the bottom-up estimation, a CNN-based method called DeepCut is used to predict the 2D body joint locations. The role of the top-down strategy is to fit the body shape to the 2D joints via SMPL [39]. The objective function penalizes the error between the projected 3D model joints and detected 2D joints.

Zhou et al. [40] propose a sparseness meets deepness (SMP) algorithm to address the challenge of 3D full-body human pose estimation from a monocular image sequence. It consists of a novel synthesis of a deep learning-based 2D part regressor, a sparsity driven 3D reconstruction approach and a 3D temporal smoothness prior. This joint consideration combines the discriminative power of the state-of-the-art 2D part detectors, the expressiveness of 3D pose models and regularization by way of aggregating information over time.

Bugra et al. [41] propose a trainable fusion scheme to fuse 2D and 3D image cues for monocular body pose estimation. It consists of two streams. The first CNN stream is used to predict the 2D joint locations and the corresponding uncertainties. The second one leverages all 3D image cues by processes applied directly to the input image. The outputs of these two streams are then fused to obtain the final 3D human pose estimation.

The work in [42] presents a robust marker-less human motion capture algorithm that can track articulated joint motion in challenging indoor and outdoor scenes. It combines the strengths of a discriminative image-based joint detection method with a model-based generative motion tracking algorithm through a unified pose optimization energy. The discriminative part-based pose detection method is implemented using CNNs. It estimates unary potentials for each joint of a kinematic skeleton model. These unary potentials serve as the basis of a probabilistic extraction of pose constraints for tracking by using a weighted sampling from a pose posterior that is guided by the model. In the final energy formula, it combines these constraints with an appearance-based model-to-image similarity term.

Table 5
Statistics of the key attributes of the popular benchmarking dataset for deformable shape segmentation in the field of AMDO. The dataset is used at least once by the manuscripts accepted for this special issue.

Dataset	Year	#Classes	#Objects	#Samples
PSB [43]	2004	19	20	380

2.3. Deformable shape segmentation

2.3.1. Benchmark datasets for deformable shape segmentation

Princeton Segmentation Benchmark (PSB) Dataset [43]. This dataset has been intensively used to evaluate 3D shape segmentation and 3D shape retrieval algorithms. It has 19 different object categories with 20 objects for each category, which results in a total of 380 models (see Table 5).

In order to evaluate segmentation methods, some popular metrics are used including rand index, cut discrepancy, hamming distance and consistency error. Rand index measures the similarity between two segmentations of the same shape. From a mathematical point of view, rand Index is related to the accuracy, but is applicable even when class labels are not used. Rand index error is equal to one minus the Rand Index. Cut discrepancy is a boundary-based method evaluating the distance between different cuts. It sums the distances from points along the cuts in the computed segmentation to the closest cuts in the ground truth segmentation, and vice-versa. Hamming Distance is a region-based method which measures the number of substitutions required to change the assignment of one region into another. Hamming Distance is directional, hence it includes underdetection rate (Rm) and false alarm (Rf) distances. Consistency Errors, whether the global version (GCE) or local version (LCE), are used to compute the hierarchical differences and similarities between segmentations. They are based on the theory that the organisation of perceptual information by humans imposes a hierarchical tree structure on perceived objects. For all four metrics, a smaller value indicates a better result. These metrics are shown in Table 6.

2.3.2. State-of-the-art on deformable shape segmentation

Table 6 shows a comparison of different methods on the PSB dataset [43]. Truc et al. [44] present a multi-view RNN (MV-RNN) algorithm for 3D mesh segmentation. It combines CNNs and a

[1] https://www.blender.org/.
[2] http://www.makehuman.org/.

Table 6
A comparison of the state-of-the-art methods on the PSB dataset [43]. The metrics include (refer to [46] for more details): Cut Discrepancy (CD), Hamming Distance (HD), Rand Index (RI) and Consistency Error (CE), Global Consistency Error (GCE), Local Consistency Error (LCE). Hamming Distance is directional, hence it includes underdetection rate (Rm) and false alarm (Rf) distances.

Paper	Metrics					
	0.149	0.090	0.118	0.124	0.065	
Multi-view RNN [44], 2017	CD 0.144	HD 0.075	Hamming-Run 0.061	Hamming-Rf 0.089	GCE 0.060	LCE 0.041
Stacked auto-encoders [45], 2016	RI: 0.118					

Table 7
The statistics of key attributes of popular benchmarking datasets for face analysis in the field of AMDO. All datasets are used at least once by accepted manuscripts in this special issue.

Dataset	Images	Subjects	Age groups	Gender	In the wild
Adience [51]	26,580	2,284	8	Yes	Yes
IoG [52]	5,080	28,231	7	Yes	Yes
MORPH II [53]	55,134	13,000	Accurate ages	Yes	No

two-layer LSTM to yield coherent segmentation of 3D shapes. The image-based CNN effectively generates the edge probability feature map while the LSTM correlates the edge maps across different views and outputs a well-defined per-view edge image. From Table 6, one can see that the performance of SI_7 is comparable to that of MV-RNN for different evaluation metrics.

The work [45] proposes an unsupervised method for 3D shape segmentation. After over-segmenting the shapes into primitive shapes, it generates high-level features from low-level features of each patch by using stacked auto-encoders. High-level features are then used for segmenting a single shape or co-segmenting a group of shapes.

2.4. Face analysis

In the area of computer vision and patter recognition, face analysis [47–50] is a popular and hot research direction. However, in this section, we limit the review techniques and datasets used in the papers contained in the special issue.

2.4.1. Benchmarking datasets for face analysis

There are three popular benchmarking datasets related to age and gender analysis: Adience [51], IoG [52] and MORPH II [53]. Some key statistics of these three datasets are listed in Table 7.

The face images of Adience and IoG datasets are collected in the wild. Both datasets contain age group and gender information. For MORPH II dataset, 50 thousand images have been collected in a controlled environment. Different from Adience and IoG datasets, MORPH II dataset provides accurate age information for each face image.

2.4.2. The state-of-the-art in face analysis

There are many subtopics of face analysis, such as face verification and recognition [61,62], facial expression recognition [63], and face attribute analysis [57] (i.e. age estimation, gender and ethnicity recognition), to mention just a few. Some solutions already achieve very promising performance that in many respects exceeds that of human face perception [61,62]. It is out of the scope of this editorial to provide a comprehensive coverage of the recent advances in the field of face analysis. We only focus on face attribute analysis, such as age estimation, ethnicity and gender recognition. Some recently published methods that were evaluated on the above three datasets are listed in Table 8. Unfortunately, there is a lack of standardisation and different publications often use different protocols for evaluation. For example, on MORPH II dataset, MRNPE [55] and Soft softmax [57] evaluate their models with CBSR protocol while SI_{10} and AgeED [54] use 80–20 protocol.

Tan et al. [54] propose a group-based method for accurate age estimation. First, they propose an age group-n encoding (AGEn) method, where adjacent ages are merged into the same category. Note that ages merged into the same group would be regarded as independent classes in the training stage. On this basis, authors transform the age estimation problem into a series of binary classification sub-problems. Subsequently, deep CNNs realising multiple classifiers are trained for age group classification. For testing, an age decoding stage is proposed to deduce the estimated age from the age group classification result. As shown in Table 8, this method achieves a MAE of 2.52 with the 80–20 protocol.

The work in [57] proposes a soft softmax loss function for age estimation, where each face image is labeled with a Gaussian label distribution rather than a single label value in softmax loss function. Compared with the traditional definitions, the proposed soft softmax loss function considers not only the chronological age but also its adjacent ages. The authors show the effectiveness of their proposal for age estimation achieving a MAE of 3.03 with a shallow network (AlexNet) on the MORPH II dataset.

Chen et al. [55] propose a Multi-Region Network Prediction Ensemble (MRNPE) for high-accuracy age estimation by leveraging both global and local context information. The model includes multiple sub-networks, where each sub-network takes both a global face image and a local region as input, e.g., face + eye, face + mouth and face + nose. Then, the average over the predictions of all sub-networks is reported as the final predicted age. One disadvantage of this work is that it needs an ensemble of networks to achieve high performance.

Rothe et al. [56] propose a DEX (Deep EXpectation) framework for real and apparent age estimation. They regard age estimation as a deep classification problem followed by a softmax expected value refinement. DEX is a very popular method for age estimation. It won the first place in the Chalearn LAP challenge 2015. In this paper, Rothe et al. also introduce the largest public dataset of face images, IMDB-WIKI datatset, which contains age and gender information for each face image. One drawback of this dataset is that its labels are noisy, as a result of being calculated based on the date of birth of the corresponding celebrity and the year when the photo was taken. Thus, this dataset is usually used for pretraining rather than evaluation.

The work in [58] presents a coarse-to-fine framework for age estimation in unconstrained environment. First, age group classification is carried out to obtain a coarse age range, and then a fine-grained refinement and an error-correcting stage follows to obtain a more reliable prediction.

Zhang et al. [59] propose a residual network of residual networks (RoR) for age group classification and gender classification. The proposed RoR architecture shows better optimization ability for age group and gender classification than alternative CNN architectures. The authors evaluate their model on Adience and IoG datasets, achieving an impressive performance of 90.59% and 90.73%, respectively.

Table 8
The key attribute statistics of popular benchmarking datasets used for age, gender and ethnicity analysis in the field of AMDO.

Methods	Adience [51]			IoG [52]			Morph II [53]	
	Age	Gender	Protocol	Age	Gender	Protocol	Age	Protocol
Deep Attention [SI_{10}]	0.6108	0.9300	–	0.6	0.8690	–	2.56	80–20[4]
AgeED[1] [54], 2017	–	–	–	–	–	–	2.52	80–20
MRNPE [55], 2017	–	–	–	–	–	–	2.73	CBSR[5]
DEX[2] [56], 2016	0.64	–	–	–	–	–	2.68	–
Soft softmax [57], 2016	–	–	–	–	–	–	3.03	CBSR
Cascaded CNN [58], 2016	0.5288	–	FF-SECV[3]	–	–	–	–	–
RoR [59], 2017	–	0.9059	FF-SECV[3]	–	0.9073	FF-SECV[3]	–	–
OR-CNN [60], 2016,	–	–	–	–	–	–	3.27	–

[1] AgeED: Age Encoding + Decoding; [2] DEX: Deep EXpectation;[3] FF-SECV: five-fold, subject-exclusive cross-validation protocol; [4] 80-20: 80% for training and 20% for testing; [5] http://www.cbsr.ia.ac.cn/users/dyi/agr.html.

Niu et al. [60] define the problem of age estimation as an ordinal regression (OR) problem and propose an OR-CNN framework to address it. In OR-CNN, the ordinal regression problem is transformed into a series of binary classification sub-problems and then a CNN with multiple binary classifiers is proposed to solve those sub-problems, where each binary classifier is trained to predict whether the age is larger than a specific value. The authors evaluate their model on MORPH II dataset and achieve a MAE of 3.27.

The work in [64] proposed a new APPA-REAL dataset. This dataset includes large face images with both real and apparent age annotations. The authors studied the realtionship between real and apparent age, and developed a residual age regression method to further improve the performance.

3. Special issue papers

In this section, we briefly introduce the 20 papers accepted for this special issue. The papers are grouped in the above mentioned four AMDO subtopoics (14 papers on human motion analysis, three papers on pose estimation, one on deformable shape, and two on face analysis).

3.1. Human motion analysis

SI_1: The paper "Convolutional Neural Networks and Long Short-Term Memory for Skeleton-Based Human Activity and Hand Gesture Recognition" by Juan C. Núñez, Raúl Cabido, Juan J. Pantrigo, Antonio S. Montemayor and José F. Vélez, proposes a deep learning-based method for skeleton-based human activity and hand gesture recognition. It combines CNN and Long Short-Term Memory (LSTM) recurrent networks. A two-stage training strategy is applied to update CNN+LSTM framework parameters. An exhaustive experimental evaluation on publicly available data benchmarks (i.e. MSR Action3D, MSR DailyActivity3D, UTKinect, NTU RGB+D, DHG-14/28, and Montalbano V2) is presented, showing the proposed method to be competitive in relation to the state-of-the-art alternatives. It relates to the work in [65], which uses a CNN+LSTM architecture for activity recognition in video sequences, but only using skeleton and achieving competitive results on five datasets.

SI_2: The paper "Hand Action Detection from Ego-centric Depth Sequences with Error-correcting Hough Transform" by Chi Xu, Lakshmi N Govindarajan and Li Cheng, presents an effective and efficient solution for hand action detection from mobile ego-centric depth sequences. It proposes a novel error-correcting mechanism to tackle the issue of incorrect votes generated by the Hough transform. The authors also provide a comprehensive in-house annotated ego-centric hand action dataset. We believe this will open new research directions in ego-centric hand action detection. The proposed method delivers favorable performance in real time (about 112 frame/s) on their proposed real-life dataset. It is related to the work in [66], which uses the concept "snippets" for action recognition, but applied to Ego-centric hand detection. Moreover, the released real-life dataset of this paper is also likely push the state of the art in Ego-centric hand detection research.

SI_3: The paper "A Flexible Trajectory Descriptor for Articulated Human Action Recognition" by Yao Guo, Youfu Li and Zhanpeng Shao, proposes an articulated skeleton representation by modeling the skeleton information as interconnections of multiple rigid bodies for action recognition. In this method, six-dimensional rigid body motion trajectories are represented by the invariant Dual Square-Root Function (DSRF) descriptor. The concept of Virtual Rigid Body (VRB) configuration is introduced to produce compact mid-level features for representing the movement of each body part. The Most Informative Part (MIP) trajectory is then used to select a subset of consistency and activity body parts in the final skeletal representation. The experimental results obtained on three datasets (MSR Action3D, MSRC-12, and UTKinect) show that the proposed method outperforms various existing skeleton-based representations in terms of recognition accuracy. It is related to the Square Root Velocity Function [67] (SRVF), which is usually used in shape analysis, but here it is applied to Articulated Human Action Recognition. The proposed DSRF descriptor includes SRVF of the 3-D point trajectory and 3-D angular trajectory.

SI_5: The paper "Human Action Recognition in RGB-D videos using Motion Sequence Information and Deep Learning" by Earnest Paul Ijjina and Chalavadi Krishna Mohan combines motion sequence information and deep learning to recognize human action from RGB-D data. It proposes a new motion representation, which is computed in various temporal regions in the RGB and depth video streams. The new representation puts emphasis on the key poses associated with each action. The derived motion representation feeds into a CNN to learn discriminative features. The proposed approach, extensively evaluated on various action and gesture datasets, is shown to advance the state of the art. More specifically, it has achieved 93.37% accuracy (evaluation protocol: leave-one-subject-out) on the MIVIA action dataset , and 86.58% top-2 accuracy on the NATOPS gesture dataset (evaluation protocol:training on the first 5 subjects, testing on the other 10 subjects).

SI_6: The paper "A Deep Convolutional Neural Network for Video Sequence Background Subtraction" by Mohammadreza Babaee, Duc Tung Dinh and Gerhard Rigoll, proposes a deep CNN architecture (namely DeepBS) for background subtraction from video sequences. The input frame along with the corresponding background image are patch-wise processed. During training, the hypotheisised foreground segmentation is compared with groundtruth segmentation and cross entropy is adopted as the loss function. In the test phase, after merging the individual patches into a single output frame, the output frame is post-processed,

yielding the final output segmentation. The proposed method is evaluated on different data-sets, and shown to outperform the existing algorithms as measured by the average ranking in terms of different evaluation metrics proposed in CDnet 2014. It is similar to the CNN-based work of [68] which uses a fixed background model. However [68] is defined for a concrete video scenario and will require re-training for other scenes with scene-specific data, while SI_6 can handle various video scenes.

SI_8: The paper "Motion Analysis: Action Detection, Recognition and Evaluation based on Motion Capture Data" by Fotini Patrona, Anargyros Chatzitofis, Dimitrios Zarpalas and Petros Daras, presents a motion analysis framework for real-time action detection, recognition and evaluation of motion capture data based on the pose and kinematics information. First, automatically computed dynamic weighting is applied, controling the joint data significance based on action involvement. Then the bag of gesturelets (BoG) model is employed for data representation of each sample and kinetic energy based descriptor sampling is performed before a codebook construction. The automatically segmented and recognized action samples are subsequently fed into a framework evaluation stage. The experimental results provide evidence that the proposed framework can effectively be used for unsupervised gesture/action training. This work is similar to bag of visual words model [20,69] widely used in video-based recognition, but here being specifically designed for a motion analysis task.

SI_9: The paper "Linear Latent Low Dimensional Space for Online Early Action Recognition and Prediction" by Victoria Bloom, Vasileios Argyriou, and Dimitrios Makris, utilizes joint motion data for recognizing actions in linear latent spaces. It operates online and in real time. It is based on supervised learning and dimensionality reduction techniques, which derive a representation for high dimensional nonlinear actions in a linear laten low dimensional space. The proposed method is evaluated on well-know datasets. Compared to the state of the art methods, the proposed approach exhibits high accuracy and low latency properties. Previous works [70–72] considered early action recognition, online action recognition and action prediction as independent events, while SI_9 tackles these three tasks jointly with the proposed Clustered Spatio-Temporal Manifolds.

SI_{11}: The paper "Enhanced skeleton visualization for view invariant human action recognition" by Mengyuan Liu, Hong Liu, and Chen Chen, proposes a new enhanced skeleton visualization method for action recognition. The authors develop a sequence-based view invariant transform, based on spatio-temporal locations of skeleton joints to eliminate the effect of view variations based on spatio-temporal locations of skeleton joints. The method encodes the spatio-temporal information conveyed by the transformed skeletons to generate a series of color images. Last, a CNN model is adopted to extract robust and discriminative features from the color images and the final predicted results are obtained by decision level fusion of the deep features. The experimental evaluation carried out on challenging datasets demonstrates the superiority of the method. It relates to the works in [17,73] where skeleton sequences are described as color images used by CNNs model for classification. Compared with [17,73], SI_{11} can capture more of the abundant spatio-temporal cues, since the generated color images extensively encode both spatial and temporal cues.

SI_{13}: The paper "Estimating 3D Trajectories from 2D Projections via Disjunctive Factored Four-Way Conditional Restricted Boltzmann Machines" by Decebal Constantin Mocanu, Haitham Bou Ammar, proposes a novel deep learning-based method referred to as disjunctive factored four-way conditional restricted Boltzmann machine (DFFW-CRBM). It introduces a novel tensor factorization capable of driving a fourth order Boltzmann machine for high dimensional time series modelling to considerably lower energy levels. Its evaluation on both simulated and real-world data has shown its effectiveness in predicting and classifying complex ball trajectories and human activities. It is related to Factored Four-Way Conditional Restricted Boltzmann Machines (FFW-CRBMs) [74]. However, FFW-CRBMs require three-dimensional labeled information for accurate predictions which is not typically available.

SI_{14}: The paper "Spatio-Temporal Union of Subspaces for Multi-body Non-rigid Structure-from-Motion" by Suryansh Kumar, Yuchao Dai, and Hongdong Li, proposes a unified framework to jointly segment and reconstruct multiple non-rigid objects. It exploits the structure of the scene along the temporal and spatial directions, modelled in terms of 3D non-rigid deformations. The spatio-temporal representation not only provides competitive 3D reconstruction but also outputs a robust segmentation of multiple non-rigid objects. The resultant optimization problem is solved using the Alternating Direction Method of Multipliers (ADMM). The experimental results show the superiority of the method, compared to the state-of-the-art. Compared with other similar methods [75,76], the proposed method of SI_{14} can learn the affinity matrices to exploit efficient spatio-temporal clustering structures.

SI_{16}: The paper "Adaptive Compressive Tracking based on Locality Sensitive Histograms" by Sixian Chan, Xiaolong Zhou, Junwei Li, and Shenyong Chen, proposes an adaptive compressive tracking algorithm which is locality sensitive, and thus robust to illumination variations. A new update mechanism is used to preserve stable features while avoiding noisy appearance variations during tracking. Furthermore it includes a trajectory rectification method to refine the tracking accuracy. The experimental results conducted on a benchmarking dataset show that the tracker achieves the state-of-the-art performance. It is related to the works [77,78] on compressive tracking with color information. Compared to [77,78], SI_{16} presents an updating mechanism to preserve stable features.

SI_{18}: The paper "Semantic Action Recognition by Learning a Pose Lexicon" by Lijuan Zhou, Wanqing Li, Philip Ogunbona, and Zhengyou Zhang, proposes a semantic representation, exploiting a pose lexicon, for action recognition. Each action is represented by a sequence of semantic poses extracted from an associated textual instruction. A visual pose model, defined as a Gaussian mixture, is learned from training samples to characterize the likelihood of an observed visual frame being generated by a visual pose. A pose lexicon model is learned using an extended Hidden Markov Model (HMM) to encode the probabilistic mapping between hidden visual poses and semantic poses sequences. With the lexicon, action classification is formulated as a problem of finding the sequence of semantic poses that best fits the sequence of visual frames as measured in terms of posterior probability. The efficacy of the proposed method is evaluated on different datasets (i.e. MSRC-12, WorkoutUOW-18, and Combined-17 action datasets) using cross-subject, cross-dataset and zero- shot protocols. SI_{18} is an extension of the work in [79]. Compared with [79], SI_{18} jointly generates visual pose sequences and aligns them to semantic pose sequences.

SI_{19}: The paper "Motion Analysis: Action Detection, Recognition and Evaluation based on motion capture data" by Fotini Patrona, Anargyros Chatzitofis, Dimitrios Zarpalas, and Petros Daras, proposes a new framework for real-time action detection and recognition. Automatic and dynamic weighting, altering the joint data significance based on the involved action, and Kinetic energy-based descriptor sampling, are employed for efficient action segmentation and labeling. The automatically segmented and recognized action instances are subsequently fed to the action evaluation stage of the framework. It compares them with the reference instances, estimating their similarity. The experimental results obtained on MSR-Action3D and MSRC12 datasets, provide evidence that the proposed method outperforms state-of-the-art methods by $0.5-6\%$ in all datasets. SI_{19} is similar to the work in [80].

Compared with [80], automatic feature weighting at the frame level is employed in SI_{19} which also uses all 20 joints.

SI_{20}: The paper "Active garment recognition and target grasping point detection using deep learning" by Enric Corona, Guillem Alenya, Antoni Gabas, and Carme Torras, proposes a new method that first identifies the type of garment and then performs a search for the two grasping points that allow a robot to bring the garment to a known pose. The experiments conducted with real robots show that most of the errors are due to unsuccessful grasps and not to the localization of the grasping points, thus a more robust grasping strategy is required. SI_{20} is similar to the work in [81] which makes use of a physics engine to create a training database. However, SI_{20} aims at avoiding costly re-grasping, which is not considered in [81].

SI_{21}: The paper "Rasabodha: Understanding Indian classical dance by recognizing emotions using deep learning" by Aparna Mohanty and Rajiv R. Sahay, proposes a CNN-based method to decipher the meaning of Navarasas associated with Indian classical dance (ICD). The proposed method is the first to use deep learing for recognizing Navarasas in order to semantically understand videos of ICD. Moreover, to evaluate the proposed method, authors also release RGB-D videos both under controlled laboratory conditions and unconstrained environments.

3.2. Human pose estimation

SI_4: The paper "A Very Simple Framework for 3D Human Pose Estimation Using a Single 2D Image: Comparison of Geometric Moments Descriptors" by Dieudinné Fabrice Atrevi, Damien Vivet, Bruno Emile and Florent Duculty, uses geometric moments to analyze the human silhouette from a single image. The proposed framework extracts the 3D human posture from a single 2D image in real time. The approach makes use of the learned correspondences between silhouettes and skeletons, extracted from synthetic 3D human models. The main contribution of this paper is the proposed technique to estimate 3D human motion via 3D synthesis software, which avoids the labour intensive manual data collection and annotation. Extensive experimental results on both synthetic and real-world datasets demonstrate the superior performance of the proposed framework compared with state-of-the-art methods. SI_4 is similar to the works in [82,83] recovering 3D human pose from single 2D images. However SI_4 uses shape-from-silhouette method to find 3D pose from a single image, being robust even in the case of noisy silhouettes.

SI_{12}: The paper "Generation of Human Depth Images with Body Part Labels for Complex Human Pose Recognition" by K. Nishi and J. Miura, develops a method for generating body-part annotated depth images of various body shapes and poses. The method is guided by a flexible human body model and a motion capture system. Based on the proposed method, the authors constructed a dataset of 10K images with eight body types for various sitting poses. The effectiveness of the generated dataset is verified by solving the part-labeling tasks using a convolutional network (FCN). SI_{12} extends the work in [84] from hand-level activities to finger-level hand activities analysis.

SI_{15}: The paper "A Hybrid Framework for Automatic Joint Detection of Human Poses in Depth Frames" by Longbo Kong, Xiaohui Yuan, and Amar Man Maharjan, proposes a novel framework to detect joints automatically by using depth camera. The proposed method categorizes the joints into implicit or dominant joints, where implicit joints are the torso (i.e., neck and shoulders) and dominant joints are elbows and knees. In this framework, a loose skeleton model is used to locate implicit joints and data-driven method is applied to detect dominant joints. It uses a hierarchy of three CNNs with different levels of specialization, trained both with synthetic and real images. The results demonstrate that the proposed work can deliver stable and accurate detection results of joints. Overall, SI_{15} combines a human body model and geodesic features of the human body together to detect and estimate the position of joints, achieving more accurate joint detection than related works in [85,86].

SI_{22}: The paper "Deep Unsupervised Learning of Visual Similarities" by Artsiom Sanakoyeu, Miguel A. Bautista, and Björn Ommer, proposes a single optimization problem to extract batches of samples with mutually consistent relations and uses weak estimates of local similarities. Learning visual similarities is then framed as a sequence of categorization tasks. The CNN then consolidates transitivity relations within and between groups and learns a single representation for all samples without the need for labels. The proposed unsupervised approach has been shown to achieve competitive performance on detailed posture analysis and object classification challenges. SI_{22} extends the unsupervised feature learning work of [87] with CNNs. However, CNNs for example ar-based learning have been rare [87] due to the limitations of the commonly used cross-entropy loss, the imbalance of data sets with many negative samples, and the unknown relationships between samples. SI_{22} overcomes these shortcomings by updating similarities and CNN parameters.

3.3. Deformable shape segmentation

SI_7: The paper "Scale Space Clustering Evolution for Salient Region Detection on 3D Deformable Shapes" by Xupeng Wang, Ferdous Sohel, Mohammed Bennamoun, Yulan Guo and Hang Lei, detects a salient region, based on clustering of a data set in a scale space generated by an auto diffusion function. The proposed method is called Scale Space Clustering Evolution (SSCE). It consists of three parts: scale field construction, shape segmentation initialization and salient region detection. The auto diffusion function is used to extract shape features at multiple time scales. The initial segmentation of the shape is obtained using persistence-based clustering. The salient regions are detected during the evolution of the scale field. The experimental results obtained on popular datasets show a very promising performance of the proposed framework. SSCE inherits the merits of persistence-based clustering [88] and clustering assessment [89] for the benefit of salient region detection on 3D deformable shapes, and thus improving accuracy.

3.4. Face analysis

SI_{10}: The paper "Age and Gender Recognition in the Wild with Deep Attention" by Pau Rodriguez López, Guillem Cucurull Preixens, Josep M Gonfaus, Francesc Xavier Roca Marvá and Jordi Gonzále Sabaté, proposes a feedforward attention mechanism for age and gender classification. In this paper, a model that consists of an attention network is employed to discover the most informative and reliable patches for age and gender classification. These patches are then further processed in a patch network in higher resolution to improve accuracy. With such attention mechanism, the model is able to discover the most informative and reliable parts in a face image even under deformation and occlusion. Experimental validation on the Adience, IoG and MORPH II dataset bemchmarks show that including attention mechanisms enhances the performance of CNNs in terms of robustness and accuracy. SI_{10} is biologically inspired and benefits from the recent successes of attention mechanisms [90].

SI_{17}: The paper "Gaussian Mixture 3D Morphable Face Model" by Paul Koppen, Zhen-Hua Feng, Josef Kittler, William Christmas, Xiao-Jun Wu, and He-Feng Yin, presents a Gaussian Mixture 3D Morphable Face Model (GM-3DMM) to represent a global population of 3D faces as a mixture of Gaussian subpopulations. It

extends the traditional 3DMM [91] naturally by adopting a shared covariance structure to mitigate small sample estimation problems associated with data in high dimensional spaces. Experiments in fitting the GM-3DMM to 2D face images to facilitate their geometric and photometric normalisation for pose and illumination invariant face recognition demonstrates the merit of the proposed multiple cohort 3D face model.

4. Conclusion

The aim of this guest editorial was to introduce this special issue on Articulated Motion and Deformable Object Recognition. The 20 papers accepted for the special issue cover four of the main subtopics of AMDO: human motion analysis (action/gesture), human pose estimation, deformable shape segmentation, and face analysis. The papers were introduced in the context of the recent developments in the field reviewed in this editorial.

Limitations and Challenges of AMDO. Although the accepted papers push the boundaries of the state of the art, there are still some limitations and challenges. First of all, there is a scope for exploring hybrid deep learning networks, as pioneered in SI_1, to capture spatial-temporal structure information more comprehensively. Second, the problem of fusing multiple modalities remains an open issue. Thanks to the recent trends in the development of cheap sensors, which provide complementary sources of information, multimodal data analysis will continue to grow in importance. One can therefore expect that future efforts in this direction will increase dramatically.

Finally, although deep learning-based methods have been demonstrated to show impressive promise in the field of AMDO, the need to collect large scale labeled data is an unwelcome obstacle. Training from only a few samples is still a challenging problem in machine learning. Although some previous works [20,69,92] have attempted zero/one-shot learning in the field of AMDO, the results achieved are not yet accurate enough. Therefore, a few-shot (i.e. one-shot or zero-shot) learning is a research direction where new advances can be expected in a foreseeable future.

We hope the contributed papers in this special issue, together with the survey of the recent developments presented in this editorial, paint a broad picture of the state of the art in the subject area of AMDO that will jointly promote future developments in this exciting field.

Acknowledgments

This work has been partially supported by the Chinese National Natural Science Foundation Projects ♯61502491, by the Spanish project TIN2016-74946-P (MINECO/ FEDER, UE) and CERCA Programme/ Generalitat de Catalunya, Spanish project TIN2015-67149-C3-2-R (Ministerio de Economia y Competitividad, Gobierno de España), EPSRC/dstl supported projects EP/K014307/1 and EP/R018456/1, and EPSRC supported projects EP/N007743/1 and EP/P022529/1.

References

[1] W. Li, Z. Zhang, Z. Liu, Action recognition based on a bag of 3D points, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2010, pp. 9–14.
[2] S. Fothergill, H. Mentis, P. Kohli, S. Nowozin, Instructing people for training gestural interactive systems, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2012, pp. 1737–1746.
[3] J. Wang, Z. Liu, Y. Wu, J. Yuan, Mining actionlet ensemble for action recognition with depth cameras, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 1290–1297.
[4] L. Xia, C. Chen, J.K. Aggarwal, View invariant human action recognition using histograms of 3d joints, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2012, pp. 20–27.
[5] K. Yun, J. Honorio, D. Chattopadhyay, T.L. Berg, D. Samaras, Two-person interaction detection using body-pose features and multiple instance learning, in: Proceedings of 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, 2012, pp. 28–35.
[6] A. Shahroudy, J. Liu, T. Ng, G. Wang, Ntu rgb+d: A large scale dataset for 3d human activity analysis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1010–1019.
[7] Q. De Smedt, H. Wannous, J.-P. Vandeborre, Skeleton-based dynamic hand gesture recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016, pp. 1–9.
[8] S. Escalera, X. Baró, J. Gonzalez, M.Á. Bautista, M. Madadi, M. Reyes, V. Ponce-López, H.J. Escalante, J. Shotton, I. Guyon, Chalearn looking at people challenge 2014: dataset and results., in: Proceedings of ECCV Workshops, 2014, pp. 459–473.
[9] V. Carletti, P. Foggia, G. Percannella, A. Saggese, M. Vento, Recognition of human actions from RGB-D videos using a reject option, in: Proceedings of International Conference on Image Analysis and Processing, Springer, 2013, pp. 436–445.
[10] Y. Song, D. Demirdjian, R. Davis, Tracking body and hands for gesture recognition: Natops aircraft handling signals database, in: Proceedings of IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011, IEEE, 2011, pp. 500–506.
[11] Y. Yang, C. Deng, S. Gao, W. Liu, D. Tao, X. Gao, Discriminative multi-instance multitask learning for 3D action recognition, IEEE Transactions on Multimedia 19 (3) (2017) 519–529.
[12] R. Qiao, L. Liu, C. Shen, A. van den Hengel, Learning discriminative trajectorylet detector sets for accurate skeleton-based action recognition, Pattern Recogn. 66 (2017) 202–212.
[13] C. Chen, K. Liu, N. Kehtarnavaz, Real-time human action recognition based on depth motion maps, J. Real-time Image Process. 12 (1) (2016) 155–163.
[14] X. Zhang, Y. Wang, M. Gou, M. Sznaier, O. Camps, Efficient temporal sequence comparison and classification using gram matrix embeddings on a riemannian manifold, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4498–4507.
[15] M. Liu, H. Liu, C. Chen, Enhanced skeleton visualization for view invariant human action recognition, Pattern Recogn. 68 (2017) 346–362.
[16] Y. Hou, Z. Li, P. Wang, W. Li, Skeleton optical spectra based action recognition using convolutional neural networks, IEEE Trans. Circuits Syst. Video Technol. PP (99) (2016) 1.
[17] P. Wang, Z. Li, Y. Hou, W. Li, Action recognition based on joint trajectory maps using convolutional neural networks, in: Proceedings of the 2016 ACM on Multimedia Conference, ACM, 2016, pp. 102–106.
[18] A. Shahroudy, T.-T. Ng, Y. Gong, G. Wang, Deep multimodal feature analysis for action recognition in RGB+ d videos, IEEE Trans. Pattern Anal. Mach. Intell. PP (99) (2017) 1.
[19] Z. Luo, B. Peng, D.A. Huang, A. Alahi, L. Fei-Fei, Unsupervised learning of long-term motion dynamics for videos, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7101–7110.
[20] J. Wan, G. Guo, S.Z. Li, Explore efficient local features from RGB-D data for one-shot learning gesture recognition, IEEE Trans. Pattern Anal. Mach. Intell. 38 (8) (2016) 1626–1639.
[21] S. Zhang, X. Liu, J. Xiao, On geometric features for skeleton-based action recognition using multilayer LSTM networks, in: Proceedings of 2017 Winter Conference on Applications of Computer Vision (WACV),, IEEE, 2017, pp. 148–157.
[22] D.C. Luvizon, H. Tabia, D. Picard, Learning features combination for human action recognition from skeleton sequences, Pattern Recogn. Lett. (2017).
[23] M. Li, H. Leung, Graph-based approach for 3D human skeletal action recognition, Pattern Recogn. Lett. 87 (2017) 195–202.
[24] F. Ahmed, P.P. Paul, M.L. Gavrilova, Joint-triplet motion image and local binary pattern for 3D action recognition using kinect, in: Proceedings of the 29th International Conference on Computer Animation and Social Agents, ACM, 2016, pp. 111–119.
[25] Q. Ke, S. An, M. Bennamoun, F. Sohel, F. Boussaid, Skeletonnet: mining deep part features for 3-D action recognition, IEEE Signal Process. Lett. 24 (6) (2017) 731–735.
[26] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, X. Xie, et al., Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks., in: Proceedings of AAAI, 2, 2016, p. 8.
[27] C. Li, Y. Hou, P. Wang, W. Li, Joint distance maps based action recognition with convolutional neural networks, IEEE Signal Process. Lett. 24 (5) (2017) 624–628.
[28] N. Neverova, C. Wolf, G. Taylor, F. Nebout, Moddrop: adaptive multi-modal gesture recognition, IEEE Trans. Pattern Anal. Mach. Intell. 38 (8) (2016) 1692–1706.
[29] E.P. Ijjina, K.M. Chalavadi, Human action recognition in RGB-D videos using motion sequence information and deep learning, Pattern Recogn. 72 (2017) 504–516.
[30] L. Brun, G. Percannella, A. Saggese, M. Vento, Action recognition by using kernels on aclets sequences, Comput. Vis. Image Underst. 144 (2016) 3–13.
[31] A. Joshi, C. Monnier, M. Betke, S. Sclaroff, Comparing random forest approaches to segmenting and classifying gestures, Image Vis. Comput. 58 (2017) 86–95.
[32] J. Zhang, W. Li, P.O. Ogunbona, P. Wang, C. Tang, RGB-D-based action recognition datasets: a survey, Pattern Recogn. 60 (2016) 86–105.
[33] F.J. Perales, J. Santos-Victor, 8th articulated motion and deformable objects, in: Proceedings of the 8th International Conference AMDO 2014, 2014.

[34] F.J. Perales, J. Kittler, 9th articulated motion and deformable objects, in: Proceedings of the 9th International Conference AMDO 2016, 2016.
[35] S. Escalera, V. Athitsos, I. Guyon, Challenges in multi-modal gesture recognition, in: Gesture Recognition, Springer, 2017, pp. 1–60.
[36] M. Asadi-Aghbolaghi, A. Clapés, M. Bellantonio, H.J. Escalante, V. Ponce-López, X. Baró, I. Guyon, S. Kasaei, S. Escalera, Deep learning for action and gesture recognition in image sequences: a survey, in: Gesture Recognition, Springer, 2017, pp. 539–578.
[37] L. Sigal, A.O. Balan, M.J. Black, Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion, Int. J. Comput. Vis. 87 (1) (2010) 4–27.
[38] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, M.J. Black, Keep it SMPL: automatic estimation of 3D human pose and shape from a single image, in: Proceedings of European Conference on Computer Vision, Springer, 2016, pp. 561–578.
[39] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, M.J. Black, SMPL: A skinned multi-person linear model, ACM Trans. Graph. (TOG) 34 (6) (2015) 248.
[40] X. Zhou, M. Zhu, S. Leonardos, K.G. Derpanis, K. Daniilidis, Sparseness meets deepness: 3D human pose estimation from monocular video, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4966–4975.
[41] B. Tekin, P. Marquez Neila, M. Salzmann, P. Fua, Learning to fuse 2D and 3D image cues for monocular body pose estimation, in: Proceedings of International Conference on Computer Vision (ICCV), in: EPFL-CONF-230311, 2017.
[42] A. Elhayek, E. de Aguiar, A. Jain, J. Thompson, L. Pishchulin, M. Andriluka, C. Bregler, B. Schiele, C. Theobalt, Marconiconvnet-based marker-less motion capture in outdoor and indoor scenes, IEEE Trans. Pattern Anal. Mach. Intell.e 39 (3) (2017) 501–514.
[43] P. Shilane, P. Min, M. Kazhdan, T. Funkhouser, The princeton shape benchmark, in: Proceedings of Shape Modeling Applications, 2004, IEEE, 2004, pp. 167–178.
[44] T. Le, G. Bui, Y. Duan, A multi-view recurrent neural network for 3D mesh segmentation, Comput. Graph. 66 (2017) 103–112.
[45] Z. Shu, C. Qi, S. Xin, C. Hu, L. Wang, Y. Zhang, L. Liu, Unsupervised 3D shape segmentation and co-segmentation via deep learning, Comput. Aid. Geom. Des. 43 (2016) 39–52.
[46] X. Chen, A. Golovinskiy, T. Funkhouser, A benchmark for 3D mesh segmentation, ACM Trans. Graph. (TOG) 28 (2009) 73. ACM.
[47] D. Riccio, G. Tortora, M. De Marsico, H. Wechsler, Egaethnicity, gender and age, a pre-annotated face database, in: Proceedings of 2012 IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications (BIOMS), IEEE, 2012, pp. 1–8.
[48] http://www.face-rec.org/databases/.
[49] M. Castrillón-Santana, M. De Marsico, M. Nappi, D. Riccio, MEG: Texture operators for multi-expert gender classification, Comput. Vis. Image Underst. 156 (2017) 4–18.
[50] A. Dantcheva, P. Elia, A. Ross, What else does your biometric data reveal? A survey on soft biometrics, IEEE Trans. Inf. Forensics Secur. 11 (3) (2016) 441–467.
[51] E. Eidinger, R. Enbar, T. Hassner, Age and gender estimation of unfiltered faces, IEEE Trans. Inf. Forensics Secur. 9 (12) (2014) 2170–2179.
[52] A.C. Gallagher, T. Chen, Understanding images of groups of people, in: Proceedings of 7th International Conference on IEEE Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2009, pp. 256–263.
[53] K. Ricanek, T. Tesafaye, Morph: a longitudinal image database of normal adult age-progression, in: Proceedings of Automatic Face and Gesture Recognition, 2006. FGR, IEEE, 2006, pp. 341–345.
[54] Z. Tan, J. Wan, Z. Lei, R. Zhi, G. Guo, S.Z. Li, Efficient group-n encoding and decoding for facial age estimation, IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI), 2017 PP (99) (2017) 1.
[55] Y. Chen, Z. Tan, A.P. Leung, J. Wan, J. Zhang, Multi-region ensemble convolutional neural networks for high accuracy age estimation, in: British Machine Vision Conference (BMVC), 2017.
[56] R. Rothe, R. Timofte, L. Van Gool, Deep expectation of real and apparent age from a single image without facial landmarks, Int. J. Comput. Vis. (2016) 1–14.
[57] Z. Tan, S. Zhou, J. Wan, Z. Lei, S.Z. Li, Age estimation based on a single network with soft softmax of aging modeling, in: Proceedings of Asian Conference on Computer Vision, Springer, 2016, pp. 203–216.
[58] J.-C. Chen, A. Kumar, R. Ranjan, V.M. Patel, A. Alavi, R. Chellappa, A cascaded convolutional neural network for age estimation of unconstrained faces, in: Proceedings of IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS), IEEE, 2016, pp. 1–8.
[59] K. Zhang, et al., Age group and gender estimation in the wild with deep ROR architecture, in: IEEE Access, 5, 2017, pp. 22492–22503.
[60] Z. Niu, M. Zhou, L. Wang, X. Gao, G. Hua, Ordinal regression with multiple output CNN for age estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4920–4928.
[61] Y. Sun, Y. Chen, X. Wang, X. Tang, Deep learning face representation by joint identification-verification, in: Proceedings of Advances in Neural Information Processing Systems, 2014, pp. 1988–1996.
[62] Y. Sun, D. Liang, X. Wang, X. Tang, Deepid3: face recognition with very deep neural networks, CoRR, abs/1502.00873, 2015.
[63] C.A. Corneanu, M.O. Simon, J.F. Cohn, S.E. Guerrero, Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications, IEEE Trans. Pattern Anal. Mach. Intell. 38 (8) (2016) 1548–1568.
[64] E. Agustsson, R. Timofte, S. Escalera, X. Baro, I. Guyon, R. Rothe, Apparent and real age estimation in still images with deep residual regressors on appa-real database, in: Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition, 2006. FGR 2006., 2017.
[65] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, T. Darrell, Long-term recurrent convolutional networks for visual recognition and description, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2625–2634.
[66] K. Schindler, L. Van Gool, Action snippets: how many frames does human action recognition require? in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, IEEE, 2008, pp. 1–8.
[67] A. Srivastava, E. Klassen, S.H. Joshi, I.H. Jermyn, Shape analysis of elastic curves in euclidean spaces, IEEE Trans. Pattern Anal. Mach. Intell. 33 (7) (2011) 1415–1428.
[68] M. Braham, M. Van Droogenbroeck, Deep background subtraction with scene-specific convolutional neural networks, in: Proceedings of International Conference on Systems, Signals and Image Processing (IWSSIP), IEEE, 2016, pp. 1–4.
[69] J. Wan, Q. Ruan, W. Li, S. Deng, One-shot learning gesture recognition from RGB-D data using bag of features, J. Mach. Learn. Res. 14 (1) (2013) 2549–2582.
[70] M.S. Ryoo, Human activity prediction: early recognition of ongoing activities from streaming videos, in: Proceedings of IEEE International Conference on Computer Vision (ICCV), IEEE, 2011, pp. 1036–1043.
[71] X. Zhao, S. Wang, X. Li, H.L. Zhang, Online action recognition by template matching, in: Proceedings of International Conference on Health Information Science, Springer, 2013, pp. 269–272.
[72] A. Galata, N. Johnson, D. Hogg, Learning variable-length Markov models of behavior, Comput. Vis. Image Underst. 81 (3) (2001) 398–413.
[73] Y. Du, Y. Fu, L. Wang, Skeleton based action recognition with convolutional neural network, in: Proceedings of 3rd IAPR Asian Conference on Pattern Recognition (ACPR), IEEE, 2015, pp. 579–583.
[74] D.C. Mocanu, H.B. Ammar, D. Lowet, K. Driessens, A. Liotta, G. Weiss, K. Tuyls, Factored four way conditional restricted Boltzmann machines for activity recognition, Pattern Recogn. Lett. 66 (2015) 100–108.
[75] Y. Dai, H. Li, M. He, A simple prior-free method for non-rigid structure-from–motion factorization, Int. J. Comput. Vis. 107 (2) (2014) 101–122.
[76] M. Lee, J. Cho, C.-H. Choi, S. Oh, Procrustean normal distribution for non-rigid structure from motion, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 1280–1287.
[77] S. He, Q. Yang, R.W. Lau, J. Wang, M.-H. Yang, Visual tracking via locality sensitive histograms, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2427–2434.
[78] M. Danelljan, F. Shahbaz Khan, M. Felsberg, J. Van de Weijer, Adaptive color attributes for real-time visual tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1090–1097.
[79] L. Zhou, W. Li, P. Ogunbona, Learning a pose lexicon for semantic action recognition, in: Proceedings of IEEE International Conference on Multimedia and Expo (ICME),, IEEE, 2016, pp. 1–6.
[80] M. Meshry, M.E. Hussein, M. Torki, Linear-time online action detection from 3D skeletal data using bags of gesturelets, in: Proceedings of IEEE Winter Conference on Applications of Computer Vision (WACV), 2016, IEEE, 2016, pp. 1–9.
[81] Y. Li, D. Xu, Y. Yue, Y. Wang, S.-F. Chang, E. Grinspun, P.K. Allen, Regrasping and unfolding of garments using predictive thin shell modeling, in: Proceedings of IEEE International Conference on Robotics and Automation (ICRA), 2015, IEEE, 2015, pp. 1382–1388.
[82] D.F. Fouhey, V. Delaitre, A. Gupta, A.A. Efros, I. Laptev, J. Sivic, People watching: human actions as a cue for single view geometry, Int. J. Comput. Vis. 110 (3) (2014) 259–274.
[83] Y. Yang, D. Ramanan, Articulated pose estimation with flexible mixtures-of–parts, in: Proceedings of Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE, 2011, pp. 1385–1392.
[84] J. Wang, Z. Liu, Y. Wu, Learning actionlet ensemble for 3D human action recognition, in: Proceedings of Human Action Recognition with Depth Cameras, Springer, 2014, pp. 11–40.
[85] K. Buys, C. Cagniart, A. Baksheev, T. De Laet, J. De Schutter, C. Pantofaru, An adaptable system for RGB-D based human body detection and pose estimation, J. Vis. Commun. Image Represent. 25 (1) (2014) 39–52.
[86] K. Nishi, J. Miura, Generation of human depth images with body part labels for complex human pose recognition, Pattern Recogn. 71 (2017) 402–413.
[87] A. Dosovitskiy, J.T. Springenberg, M. Riedmiller, T. Brox, Discriminative unsupervised feature learning with convolutional neural networks, in: Proceedings of Advances in Neural Information Processing Systems, 2014, pp. 766–774.
[88] F. Chazal, L.J. Guibas, S.Y. Oudot, P. Skraba, Persistence-based clustering in Riemannian manifolds, J. ACM (JACM) 60 (6) (2013) 41.
[89] B. Rieck, H. Leitte, Exploring and comparing clusterings of multivariate data sets using persistent homology, in: Computer Graphics Forum, 35, Wiley Online Library, 2016, pp. 81–90.
[90] V. Mnih, N. Heess, A. Graves, et al., Recurrent models of visual attention, in: Proceedings of Advances in neural information processing systems, 2014, pp. 2204–2212.
[91] V. Blanz, T. Vetter, A morphable model for the synthesis of 3D faces, in: Proceedings of the 26th annual conference on Computer graphics and interactive techniques, ACM Press/Addison-Wesley Publishing Co., 1999, pp. 187–194.
[92] P. Mettes, C.G. Snoek, Spatial-aware object embeddings for zero-shot localization and classification of actions, in: Proceedings of ICCV, 2017, pp. 4443–4452.