A Survey on Human Motion Analysis from Depth Data

Mao Ye[1], Qing Zhang[1], Liang Wang[2], Jiejie Zhu[3], Ruigang Yang[1], and Juergen Gall[4]

[1] University of Kentucky, 329 Rose St., Lexington, KY, 40508, U.S.A.
{mao.ye,qing.zhang}@uky.edu, ryang@cs.uky.edu
[2] Microsoft, One Microsoft Way, Redmond, WA, 98052, U.S.A.
liangwan@microsoft.com
[3] SRI International Sarnoff, 201 Washington Rd, Princeton, NJ, 08540, U.S.A.
jiejie.zhu@sri.com
[4] University of Bonn, Roemerstrasse 164, 53117 Bonn, Germany
gall@iai.uni-bonn.de

Abstract. Human pose estimation has been actively studied for decades. While traditional approaches rely on 2d data like images or videos, the development of Time-of-Flight cameras and other depth sensors created new opportunities to advance the field. We give an overview of recent approaches that perform human motion analysis which includes depth-based and skeleton-based activity recognition, head pose estimation, facial feature detection, facial performance capture, hand pose estimation and hand gesture recognition. While the focus is on approaches using depth data, we also discuss traditional image based methods to provide a broad overview of recent developments in these areas.

1 Introduction

Human motion analysis has been a major topic from the early beginning of computer vision [1,2] due to its relevance to a large variety of applications. With the development of new depth sensors and algorithms for pose estimation [3], new opportunities have emerged in this field. Human motion analysis is, however, more than extracting skeleton pose parameters. In order to understand the behaviors of humans, a higher level of understanding is required, which we generally refer to as activity recognition. A review of recent work of the lower level task of human pose estimation is provided in the chapter *Full-Body Human Motion Capture from Monocular Depth Images*. Here we consider the higher level activity recognition task in Section 2. In addition, the motion of body parts like the head or the hands are other important cues, which are discussed in Section 3 and Section 4. In each section, we give an overview of recent developments in human motion analysis from depth data, but we also put the approaches in context of traditional image based methods.

M. Grzegorzek et al. (Eds.): Time-of-Flight and Depth Imaging, LNCS 8200, pp. 149–187, 2013.

2 Activity Recognition

A large amount of research has been conducted to achieve the high level understanding of human activities. The task can be generally described as: given a sequence of motion data, identify the actions performed by the subjects present in the data. Depending on the complexity, they can be conceptually categorized as gestures, actions and activities with interactions. Gestures are normally regarded as the atomic element of human movements, such as "turning head to the left", "raising left leg" and "crouching". Actions usually refer to a single human motion that consists of one or more gestures, for example "walking", "throwing", etc. In the most complex scenario, the subject could interact with objects or other subjects, for instance, "playing with a dog", "two persons fighting" and "people playing football".

Though it is easy for human being to identify each class of these activities, currently no intelligent computer systems can robustly and efficiently perform such task. The difficulties of action recognition come from several aspects. Firstly, human motions span a very high dimensional space and interactions further complicate searching in this space. Secondly, instantiations of conceptually similar or even identical activities by different subjects exhibit substantial variations. Thirdly, visual data from traditional video cameras can only capture projective information of the real world, and are sensitive to lighting conditions.

However, due to the wide applications of activities recognition, researchers have been actively studying this topic and have achieved promising results. Most of these techniques are developed to operate on regular visual data, i.e. color images or videos. There have been excellent surveys on this line of research [4,5,6,7]. By contrast, in this section, we review the state-of-the-art techniques that investigate the applicability and benefit of depth sensors for action recognition, due to both its emerging trend and lack of such a survey. The major advantage of depth data is alleviation of the third difficulty mentioned above. Consequently, most of the methods that operate on depth data achieve view invariance or scale invariance or both.

Though researchers have conducted extensive studies on the three categories of human motions mentioned above based on visual data, current depth based methods mainly focus on the first two categories, i.e. gestures and actions. Only few of them can deal with interactions with small objects like cups. Group activities that involve multiple subjects have not been studied in this regard. One of the reason is the limited capability of current low cost depth sensors in capturing large scale scenes. We therefore will focus on the first two groups as well as those that involve interactions with objects. In particular, only full-body motions will be considered in this section, while body part gestures will be discussed in Section 3 and Section 4.

The pipeline of activity recognition approaches generally involve three steps: *features extraction*, *quantization/dimension reduction* and *classification*. Our review partly follows the taxonomy used in [4]. Basically we categorize existing methods based on the features used. However, due to the special characteristics of depth sensor data, we feel it necessary to differentiate methods that rely directly on depth

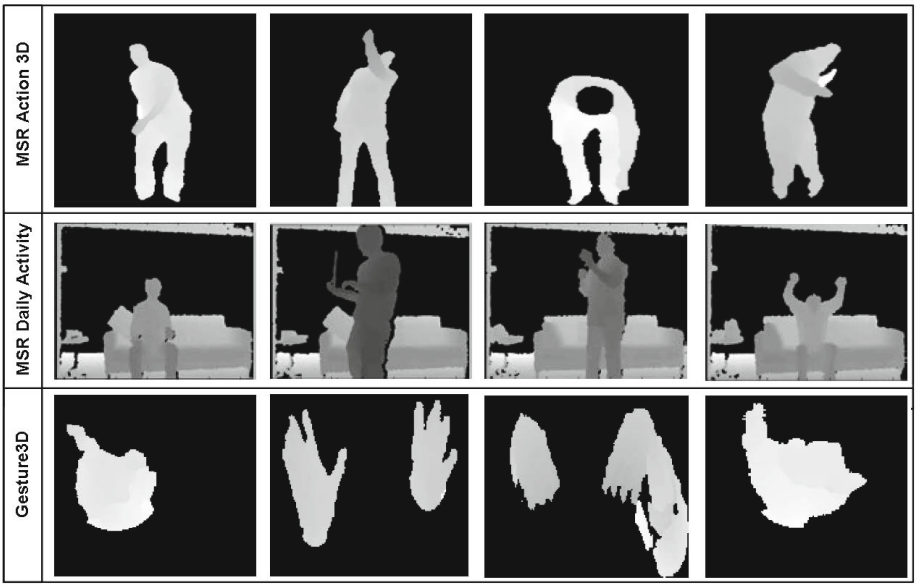


Fig. 1. Examples from the three datasets: MSR Action 3D Dataset [8], MSR Daily Activity Dataset [9] and Gesture3D Dataset [10] ©2013 IEEE

maps or features therein, and methods that take skeleton (or equivalently joints) as inputs. Therefore, the reviewed methods are separated into *depth map-based* and *skeleton-based*. Following [4], each category is further divided into *space time approaches* and *sequential approaches*. The space time approaches usually extract local or global (holistic) features from the space-time volume, without explicit modeling of temporal dynamics. Discriminative classifiers, such as SVM, are then usually used for recognition. By contrast, sequential approaches normally extract local features from data of each time instance and use generative statistical model, such as HMM, to model the dynamics explicitly.

We discuss the *depth map-based* methods in Section 2.2 and the *skeleton-based* methods in Section 2.3. Some methods that utilize both information are also considered in Section 2.3. Before the detailed discussions of the existing methods, we would like to first briefly introduce several publicly available datasets, as well as the mostly adopted evaluation metric in Section 2.1.

2.1 Evaluation Metric and Datasets

The performance of the methods for activity recognition are evaluated mainly based on *accuracy*, that is the percentage of correctly recognized actions. There are several publicly available dataset collected by various authors for evaluation purpose. Here we explicitly list three of them that are most popular, namely the MSR Action 3D Dataset [8], MSR Daily Activity Dataset [9] and Gesture3D

Table 1. Summary of the most popular publicly available datasets for evaluating activity recognition performance

Datasets	#Subjects	#Types of activities	#Data sequences
MSR Action 3D [8]	10	20	567
Gesture3D [10]	10	12	336
MSR Daily Activity 3D [9]	10	16	960

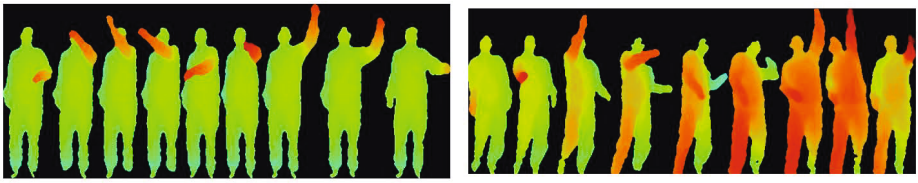
Dataset [10]. Each of the datasets include various types of actions performed by different subjects multiple times. Table 1 provides a summary of these three datasets, while Figure 1 shows some examples. Notice that the MSR Action 3D Dataset [8] is pre-processed to remove the background, while the MSR Daily Activity 3D Dataset [9] keeps the entire captured scene. Therefore, the MSR Daily Activity 3D Dataset can be considered as more challenging. Most of the methods reviewed in the following sections were evaluated on some or all of these datasets, while some of them conducted experiments on their self-collected dataset, for example due to mismatch of focus.

2.2 Depth Maps-Based Approaches

The depth map-based methods rely mainly on features, either local or global, extracted from the space time volume. Compared to visual data, depth maps provide metric, instead of projective, measurements of the geometry that are invariant to lighting. However, designing both effective and efficient depth sequence representations for action recognition is a challenging task. First of all, depth sequences may contain serious occlusions, which makes the global features unstable. In addition, the depth maps do not have as much texture as color images do, and they are usually too noisy (both spatially and temporally) to apply local differential operators such as gradients on. It has been noticed that directly applying popular feature descriptors designed for color images does not provide satisfactory results in this case [11]. These challenges motivate researchers to develop features that are semi-local, highly discriminative and robust to occlusion. The majority of depth maps based methods rely on space time volume features; therefore we discuss this sub-category first, followed by the sequential methods.

2.2.1 Depth Map-Based Space Time Volume Approaches

Li et al. [8] present a study on recognizing human actions from sequences of depth maps. The authors employed the concept of bag-of-points in the expandable graphical model framework to construct the action graph [12] to encode the actions. Each node of the action graph which represents a salient posture is described by a small set of representative 3d points sampled from the depth maps (example depth maps are shown in Figure 2. The key idea is to use a small number of 3d points to characterize the 3d shape of each salient posture and to use a Gaussian Mixture Model to effectively capture the statistical distribution of the points. In terms of 3d points sampling, the paper proposed a simple yet



(a) *Draw tick* (b) *Tennis serve*

Fig. 2. Examples of the sequences of depth maps for actions in [8]: (a) Draw tick and (b) Tennis serve ©2010 IEEE

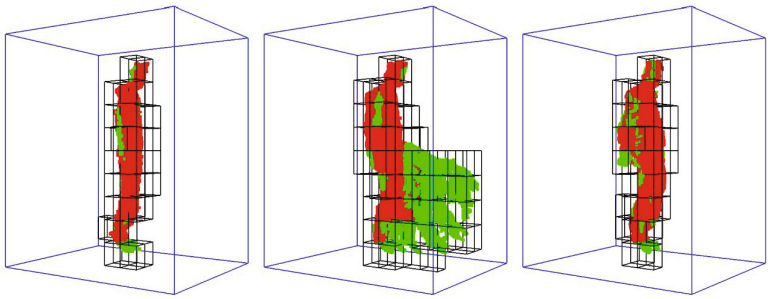


Fig. 3. Examples of the space-time cells of a depth sequence of the action Forward Kick used in [13] ©2010 Springer

effective projection based sampling scheme for sparse sampling from depth maps. Experiments were conducted on the dataset collected by the authors, which is later known as the MSR Action3D Dataset [8]. The results have shown that over 90% recognition accuracy is achieved by only sampling 1% of the 3d points from the depth maps.

One limitation of the approach in [8] is the loss of spatial context information between interest points. Also, due to noise and occlusions in the depth maps, the silhouettes viewed from the side and from the top may not be reliable. This makes it very difficult to robustly sample the interest points given the geometry and motion variations across different persons. To address these issues, Vieira et al. [13] presented a novel feature descriptor, named Space-Time Occupancy Pattens (STOP). The depth sequence is represented in a 4d space-time grid. A saturation scheme is then used to enhance the roles of the sparse cells which typically consist of points on the silhouettes or moving parts of the body. Figure 3 illustrates the space-time cells from a depth sequence of the action Forward Kick. The sequence is divided into three time segments, and each segment contains of about 20 frames. Only the non-empty cells are drawn. The red points are those in the cells that have more than a certain number of points. The accuracy of the STOP features for action classification was shown to be higher in a comparison with [8] on the MSR Action3D Dataset [8].

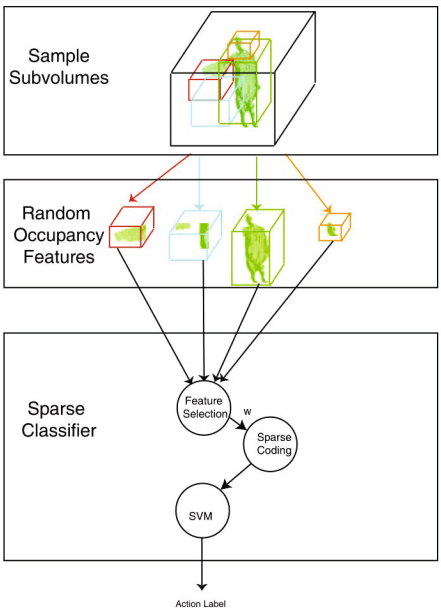


Fig. 4. The framework of the method proposed by [14]. Note that only 3d sub-volumes are shown for illustration. In the real implementation, 4d sub-volumes are used. ©2012 Springer.

Wang et al. [14] also studied the problem of action recognition from depth sequences captured by a single commodity depth camera. In order to address the noise and occlusion issues, the authors treated a three-dimensional action sequence as a 4d shape and proposed Random Occupancy Pattern (ROP) features, which were extracted from randomly sampled 4d sub-volumes with different sizes and at different locations. Since the ROP features are extracted at a larger scale, they are robust to noise. In the meantime, they are less sensitive to occlusion because they encode information from the regions that are most discriminative for the given action. The paper also proposed a weighted random sampling scheme to efficiently explore the large dense sampling space. Sparse coding is employed to further improve the robustness of the proposed method. The general framework of the method proposed in [14] is shown in Figure 4. The authors compared their results against those obtained from [8] and [13] using the MSR Action3D Dataset [8]. Experimental results conclude that [14] outperforms [8] by a large margin ($> 10\%$) and is slightly superior to [13].

Yang et al. [15] developed the so-called Depth Motion Maps (DMM) to capture the aggregated temporal motion energies. More specifically, the depth map is projected onto three pre-defined orthogonal Cartesian planes and then normalized. For each projected map, a binary map is generated by computing and thresholding the difference of two consecutive frames. The binary maps are then summed up to obtain the DMM for each projective view. Histogram of Oriented

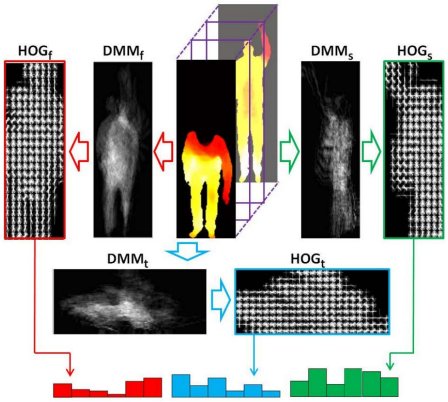


Fig. 5. The framework of the method proposed by [15] ©2012 ACM

Gradients (HOG) is then applied to each view to extract features, and features from three views are concatenated together to form the DMM-HOG descriptors. An SVM classifier is trained on such descriptors for recognition. Compared to many other methods in this category, the computational cost of this approach is relatively low, since HOG is only applied to the final DMM. Evaluations based on the MSR Action3D Dataset [8] showed high recognition rates. However, the hand-crafted projection planes might raise problems related to view-dependency. Their high recognition rate is partly due to the fact that subjects in the MSR-Action3D Dataset mostly face towards the camera. An interesting exploration they performed is to characterize the number of frames required to generate satisfactory recognition results. The conclusion they reached is that only short sub-sequence of roughly 35 frames is sufficient. Nonetheless, the number is in fact largely dependent on the complexity of the actions.

More recently, Oreifej and Liu [11] presented a new descriptor for depth maps. The authors describe the depth video sequence using a histogram capturing the distribution of the surface normal orientation in the 4d volume of time, depth and spatial coordinates. As the depth sequence represents a depth function of space and time, they proposed to capture the observed changing structure using a histogram of oriented 4d surface normals (HON4D). To construct HON4D, the 4d space is initially quantized using the vertices of a regular polychoron. Afterwards, the quantization is refined using a novel discriminative density measure such that additional projectors are induced in the directions where the 4d normals are denser and more discriminative. Figure 6 summarizes the various steps involved in computing the HON4D descriptor. Experimental results from the standard benchmark MSR Action3D Dataset [8] showed that using the proposed HON4D descriptors achieved the state of the art in recognition accuracy.

Rather than using depth maps only, Zhang et al. [16] proposed 4d local spatio-temporal features as the representation of human activities. This 4d feature is a weighted linear combination of a visual component and a geometric component.

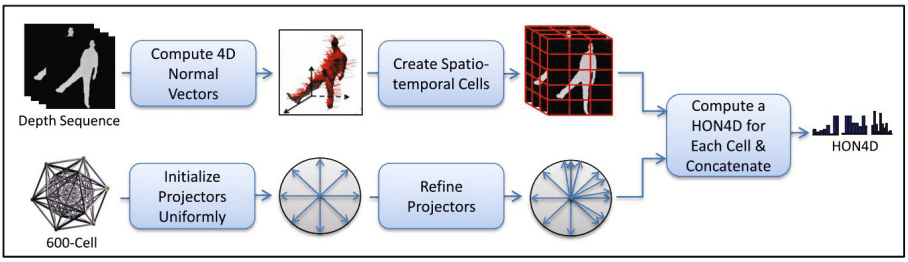


Fig. 6. The steps for computing HON4D descriptor in [11] ©2013 IEEE

This approach then concatenates per-pixel responses and their gradients within a spatial-temporal window into a feature vector which has over 10^5 elements. In order to reduce such a high dimensionality, the approach applies K-means clustering on all feature vectors collected from a training dataset and forms a codebook with 600 vocabularies which is used to code six activity categories: lift, remove, wave, push, walk and signal. In order to predict activities from input videos, the approach formulates this problem as a Latent Dirichlet Allocation (LDA) model where six activity categories are regarded as topics, and codes calculated from 4d features are regarded as words. Gibbs sampling [17] is then adopted for approximate estimation and inference for this high-dimensional model, due to its efficiency. They demonstrated their approach on a self-collected dataset with 198 short video clips, each lasting from 2 to 5 seconds, including 6 activities. Each activity has 33 video clips. The combined features (85.5%) using LDA outperforms features based on intensity (77.67%), demonstrating that depth is an important cue to improve activity recognition accuracy.

Lei et al. [18] also combine depth and color cues, while targeting at recognizing fine-grained kitchen activities. Different from the methods above that are mainly limited to single subject motions, this work demonstrated a successful prototype that tracks the interaction between a human hand and objects in the kitchen, such as mixing flour with water and chopping vegetables. It is shown that the recognition of objects and their state changes through actions is helpful in recognizing very fine-grained kitchen activity from few training samples. The reported system uses object tracking results to study both object and action recognition. For object recognition, the system uses SIFT-like feature from both color and depth data. These features are fed into an SVM to train a classifier. For action recognition, the authors combine a global feature and a local feature. The global feature is defined by PCA on the gradients of 3d hand trajectories since a hand can be tracked using human skin characteristics. The local feature is defined as bag-of-words of snippets of trajectory gradients. The training dataset includes 35 object instances and 28 action instances. Each action instance has only 3 samples compared with 33 in [16]. The reported overall action recognition accuracy is around 82% by combining trajectory-based action recognition with object recognition. This shows that by combining hand-object tracking and object-action recognition, systems like this are capable of identifying and recognizing objects and actions in a real-world kitchen environment with only a

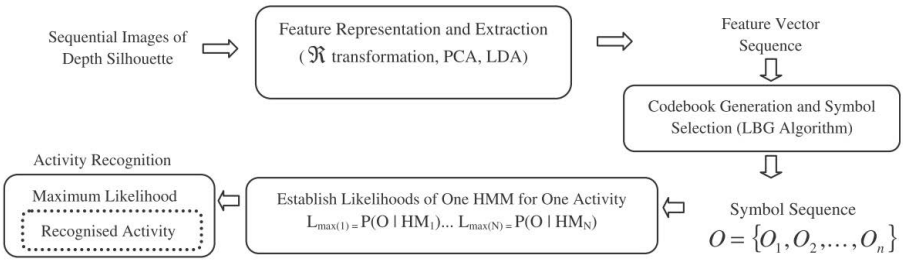


Fig. 7. The flow of the method proposed by Jalal et al. [19] ©2013 SAGE

small dataset. This work is only a proof of concept. Deploying such a system in a real environment requires a larger set of objects and actions, along with variations across people and physical environments that present many challenges not revealed in their work. Nevertheless, there are many possibilities to enhance their system, such as combining multiple sensors including wearable cameras and infrastructure sensors to robustify RGBD cameras in a real world environment.

2.2.2 Depth Maps-Based Sequential Approaches

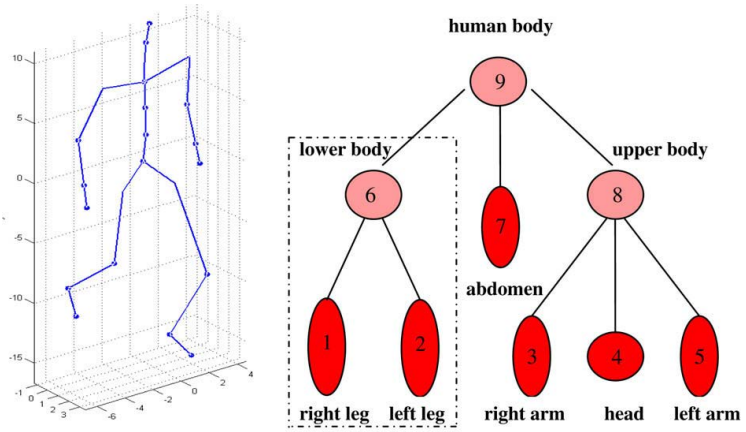
As mentioned before, local differential operators are not suitable for extracting features from depth maps, resulting in difficulties in extracting reliable temporal correspondences. Therefore only few approaches have explored the possibility of explicitly modeling temporal dynamics from depth maps. This line of research lies in between pure depth map-based methods and skeleton-based methods. They try to design features from which reliable temporal motion can be extracted, while skeletons are one of the most natural features that embed such information.

Inspired by the great success of silhouette based methods developed for visual data, Jalal et al. [19] extract depth silhouettes to construct feature vectors. Figure 7 shows the overall flow of their proposed pipeline. The key idea is to apply $\mathcal{R}$ transform [20] on the depth silhouette to obtain compact shape representation reflecting time-sequential profiles of the activities. PCA is then used for dimension reduction and Linear Discriminant Analysis is adopted to extract most discriminant vectors as in [21]. Similar to most sequential methods for visual data, HMM is utilized for recognition. Experiments were performed on 10 daily home activities collected by the authors, each with 15 video clips. Upon this dataset, a recognition rate of 96.55% was achieved.

Together with the skeleton-based methods that will be studied in Section 2.3, the depth map-based approaches are summarized in Table 2 and Table 3.

2.3 Skeleton-Based Approaches

The study of skeleton-based activity recognition dates back to the early work by Johansson [23], which demonstrated that a large set of actions can be recog-



(a) Human skeleton (b) Human hierarchical decomposition

Fig. 8. (a) Example of a typical human skeleton used for recognition. (b) Example of a typical hierarchy of human body parts in a tree structure as in [22] ©2010 Elsevier.

nized solely from the joint positions. This concept has been extensively explored ever since. In contrast to the depth maps-based methods, the majority of the skeleton-based methods model temporal dynamics explicitly. One main reason is the natural correspondence of skeletons across time, while this is difficult to establish for general visual and depth data. There are mainly three ways to obtain skeletons: active motion capture (MoCap) systems, monocular or multi-view color images and single view depth maps [24,25]. One difference worth mentioning is the degree of embedded noise. Overall the MoCap data is the cleanest compared to the other two. A multi-view setup is usually adopted for color images, and therefore produces more stable skeleton estimations than those from monocular depth maps. Early methods were mostly tested on MoCap data and skeletons from multi-view image data; while more recent work operates more on noisy skeleton data from monocular depth maps, mainly due to its simple setup. In the following, we first discuss sequential approaches, followed by space time volume approaches.

2.3.1 Skeleton-Based Sequential Approaches

Though we discuss mainly recent research in this study, the seminal work by Campbell and Bobick [26] is still worth mentioning. They represent human actions as curves in low-dimensional phase spaces obtained via projection of 3d joint trajectories. The phase space is defined with each axis being an independent parameter of the body, for example ankle-ankle, or its first derivative. A static pose is interpreted as a point in the phase space, while an action forms a curve. Multiple 2d subspaces are chosen via supervised learning paradigm and the action curves are projected onto these spaces as the action feature. A given actions is projected as a set of points and recognized by verifying whether they

are on certain action curves. However, due to their cubic polynomial fitting of the projected curves, only simple movements can be recognized. In particular, they succeeded in recognizing various ballet dance moves. Notice that dynamics are not explicitly considered for their recognition, though such information is embedded in the curve representation. Due to the phase space representation, their method is both view invariant and scale invariant.

Similar to the idea of 2d subspace selection above, Lv et al. [27] designed a set of (spatially) local features based on single joints or combinations of small sets of joints. Their observations suggest that using solely the full pose vector might cause loss of some relevant information and reduce the discriminative power. They consider three types of motions that involve motions of different primary body parts: {*leg+torso, arm, head*}. In the end, they construct a 141 dimensional feature vector from seven types of features including the full pose vector. The skeleton is pre-normalized to avoid dependence on initial body orientation and body size variations. An HMM is built for each feature and action class to model the temporal dynamics. A key novelty of their method is to treat each of the HMM models as a weak classifier and combine them with the multi-class AdaBoost classifier [28] to significantly increase the discriminative power. Besides, they propose a method using dynamic programming to extract from a continuous video the segment that involves an activity considered. They tested their method on two datasets: a set of 1979 MoCap sequences with 243,407 frames in total, collected from the internet, and a set of annotated motion sequences [29]. For the first dataset, they achieved recognition rates of $\{92.3\%, 94.7\%, 97.2\%\}$ for the three classes of actions separately when half of the data was used as training data, and $\{88.1\%, 91.9\%, 94.9\%\}$ when the training data was reduced to 1/3. Noticeably, a 30% gain was reached via the use of AdaBoost in this test. A recognition rate of 89.7% was achieved for the second dataset, which is segmented by their proposed method and thus more difficult. Overall their method has achieved promising results with the small classes of actions considered. However, in reality many human actions involve motions of the entire body, such as dancing, and it is not clear how well this method can be generalized to deal with such complex actions.

The recent work by Xia et al. [21] proposed a feature called Histogram of 3d Joint Locations (HOJ3D) that essentially encodes spatial occupancy information relative to the skeleton root, i.e. hip center. Towards this end, they define a modified spherical coordinate system on the hip center and partition the 3d space into n bins, as shown in Figure 9 (a) and (b) respectively. Radial distance is not considered in this spherical coordinate system to make it scale-invariant. Different from other methods that also utilize spatial occupancy information that make binary decision, such as [14] and [9], they perform a probabilistic voting to determine the fractional occupancy, as demonstrated in Figure 9(c). In order to extract dominant features, Linear Discriminant Analysis is applied to reduce the dimensionality from n to (#Class−1). Vector Quantization is performed via K-means to discretize the continuous vectors obtained from the previous step, and discrete HMM is adopted to model the dynamics and recognize actions.

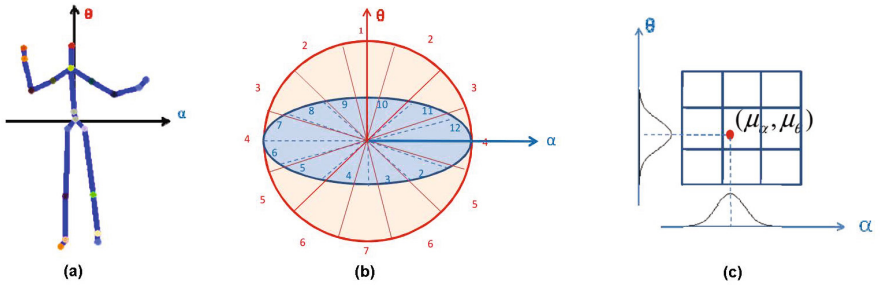


Fig. 9. Reference coordinates of HOJ3D (a) and spherical coordinate system for joint location binning used in [21]. (c) The probabilistic voting for spatial occupancy via a Gaussian weighting function in [21]. ©2012 IEEE.

They tested their approach on both their own dataset and the MSR Action3D dataset [8]. Experiments on the MSR Action3D Dataset [8] showed that their method outperformed [8]. However, the heavy reliance on the hip joint might potentially jeopardize their recognition accuracy, due to the noise embedded in the estimated hip joint location. Currently the estimation of this joint with [25] is not very reliable, especially when the subject is not facing towards the camera.

The above-mentioned methods are mostly limited to single human actions, due to lack of a model of the motion hierarchy. By contrast, Koppula et al. [30,31] explicitly consider human-object interactions. They aimed at joint activity and object affordance labeling from RGBD videos as illustrated in Figure 10. They defined an MRF over the spatio-temporal sequence with two kinds of nodes, namely objects nodes and sub-activity nodes, and edges representing the relationships between object affordances, their relations with sub-activities, and their evolution over time. The explicit modeling of the motion hierarchy enables this method to handle complex activities that involve human-object interactions. Features are defined for both classes of nodes. The object node feature is a vector representing the object's location in the scene and how it changes within the temporal segment including the transformation matrix and displacement of the corresponding points from the SIFT tracker. The sub-activity node feature map gives a vector of features computed using the human skeleton information obtained from a skeleton tracker on RGBD video. By defining the feature vectors, they then train a multi-class SVM classifier on the training data. Given the model parameters, the inference problem is to find the best labeling for the input video. Its equivalent formulation has a linear relaxation which can be solved efficiently using a graph-cut method. Evaluations are conducted based on the Cornell 60 dataset [32] and a new dataset acquired by the authors, named Cornell 120.

Similar to the work of Koppula et al. [30,31], Sung et al. [34,35] also explicitly model the activity hierarchy, however, with a two-layer Maximum Entropy Markov Model (MEMM) [36]. The lower layer nodes of the MEMM represent sub-activities such as "lifting left hand", while higher level nodes represent more

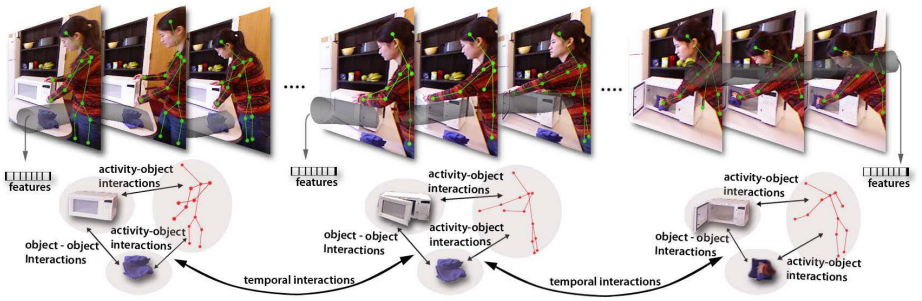


Fig. 10. The MRF graph of [33]. Different types of nodes and relationships modeled in part of the *cleaning objects* activity comprising three sub-activities: reaching, opening and scrubbing. ©2012 IEEE.

general and complex activities such as "pouring water". The features used in their work consist of four components. The first one are body pose features based on joint orientations that are transformed to the local coordinate system of torso to remove view dependency. The angles are represented as quaternions to avoid the well-known gimbal lock phenomenon when using Euler angles. Besides, the angle between each foot and the torso is explicitly emphasized to tell apart sitting poses from standing poses. The second component consists of the positions of the hands relative to the torso and the head, due to the discriminative power of hand positions. The third considers the motion of joints with a temporally sliding window. Besides these skeleton features, they incorporate image and point cloud features as the last component. Specifically, Histogram of Oriented Gradients (HOG) [37] descriptors are used on both RGB and depth data. A key component of their model is dynamic association of the sub-activities with the higher-level activities. In general, they do not assume that the input videos are segmented. Instead, they use GMM to group the training data into clusters that represent sub-activities and utilize the proposed probabilistic model to infer an optimal association of these two layers on-the-fly. Experiments are conducted based on the dataset acquired by the authors.

The work by Wang et al. [9] also utilizes both skeleton and point cloud information. The key idea is that some actions differ mainly due to the objects in interactions, while skeleton information is not sufficient in such cases. Towards this end, they introduced a novel actionlet ensemble model to represent each action and capture the intra-class variance via occupancy information, as illustrated in Figure 11. In terms of skeleton information, one important observation made by them is that the pairwise relative positions of the joints are more discriminative than the joint positions themselves. Interactions between humans and environmental objects are characterized by Local Occupancy Patterns (LOP) at each joint. The LOP features are computed based on the 3d point cloud around a particular joint. The local space of each joint is discretized using a spatial grid as shown in Figure 11. Moreover, they concatenate both feature vectors and apply Short Fourier Transform to obtain the coefficients as

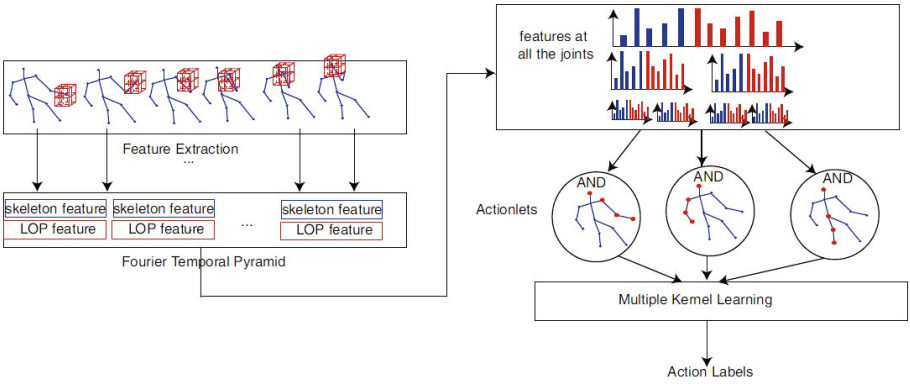


Fig. 11. The actionlet framework proposed by Wang et al. [9] ©2012 IEEE

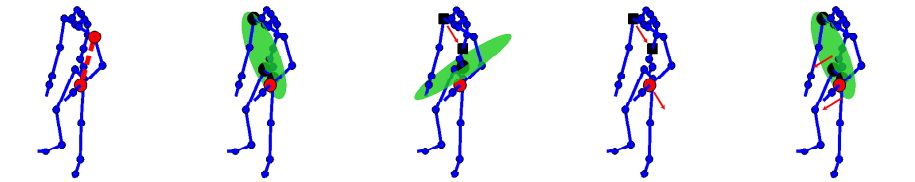


Fig. 12. Relational pose features [38]. (a) Euclidean distance between two joints (red). (b) Distance between a joint (red) and a plane (green) defined by three joints. (c) Distance between a joint (red) and a plane (green) defined by one joint and the normal direction of two joints (black). (d) Velocity of a joint (red) in the direction of two joints (black). (e) Velocity of a joint in normal direction of the plane.

the Fourier Temporal Pyramid features at each joint. The Fourier Temporal Pyramid is insensitive to temporal misalignment and robust to noise, and also can characterize the temporal structure of the actions. An *actionlet* is defined as a conjunctive structure on the base features (Fourier Pyramid features). They learn the discriminative actionlet by iteratively optimizing parameters through a generic SVM solver and obtain an SVM model defining a joint feature map on the data and labels as a linear output function. Once they have training pairs, they employed a mining algorithm to output a discriminative actionlet pool which contains the actionlets meeting the criteria: having a large confidence and a small ambiguity. They evaluated their method using CMU MoCap dataset, MSR Action3D Dataset [8] and a new dataset named MSR Daily Activity 3D. Experiments demonstrated the superior performance of their method compared to other state-of-the-art methods.

A more general approach has been proposed by Yao et al. [38] where skeleton motion is encoded by relational pose features [39], as shown in Figure 12. These features describe geometric relations between specific joints in a single pose or

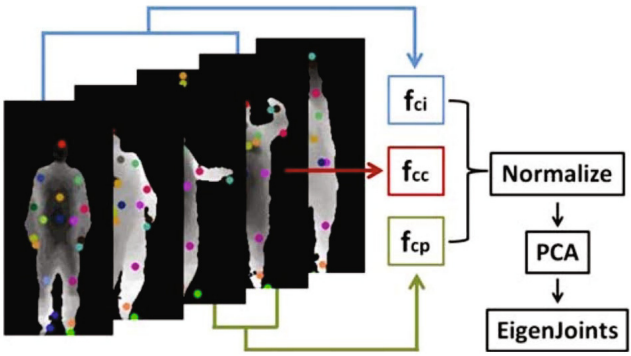


Fig. 13. The EigenJoints features developed by Yang et al. [33] ©2012 IEEE

a short sequence of poses. For action recognition, a Hough forest [40] has been used. Furthermore, a system for coupling the closely intertwined tasks of action recognition and pose estimation is presented. Experiments on a multi-view kitchen dataset [41] indicate that the quality of estimated poses with an average error between 42mm-70mm is sufficient for reliable action recognition.

Similar to the depth maps based category, the sequential methods usually require a larger set of training data. However, the explicit modeling of motion dynamics provide the potential to capture complex and general activities. A major difference is that the dynamics are well defined due to exact semantic definition of joints.

2.3.2 Skeleton-Based Space Time Volume Approaches

The space time volume approaches using skeleton information usually extract global features from the joints, sometimes combined with point cloud data. This line of research is relatively new and only a few methods lie in this category.

Yang et al. [33] developed the EigenJoints features from RGBD sequences as shown in Figure 13. The features include posture features f_{cc}, motion features f_{cp} and offset features f_{ci}. The posture and motion features encode spatial and temporal configuration with pair-wise joint differences in single frames and between consecutive frames, respectively. The offset features represent the difference of a pose with respect to the initial pose with the assumption that the initial pose is generally neutral. They normalize the three channels and apply PCA to reduce redundancy and noise to obtain the EigenJoints descriptor. For classification, they adopt a Naive-Bayes-Nearest-Neighbor (NBNN) classifier due to its simplicity. The video-to-class NN search is accelerated using a KD-tree. Their evaluation on the MSR Action3D dataset [8] demonstrated the effectiveness of their approach. One limitation of their method is the assumption about the initial pose. It is not clear how this assumption affects the recognition accuracy in a more general context.

Table 2. Accuracy of the reviewed activity recognition methods on the popular datasets. Notice that the numbers are based on those reported in the corresponding papers and the specific evaluation methodology can be slightly different even for the same dataset.

	MSR Action 3D [8]			Gesture3D [10]	MSR Daily Activity 3D [9]
	$\frac{1}{3}$ training	$\frac{2}{3}$ training	cross subject		
[8]	91.36%	94.2%	74.7%	-	-
[13]	96.8%	98.25%	84.8%	-	-
[14]	-	-	86.2%	88.5%	-
[15]	95.83%	97.37%	91.63%	89.20%	-
[11]	-	-	88.89%	92.45%	80%
[16]	-	-	-	-	-
[18]	-	-	-	-	-
[19]	95.8%	97.78%	91.63%	-	-
[26]	-	-	-	-	-
[27]	-	-	-	-	-
[21]	96.2%	97.15%	78.97%	-	-
[30,31]	-	-	-	-	-
[34,35]	-	-	-	-	-
[9]	-	-	88.2%	-	85.75%
[38]	-	-	-	-	-
[33]	95.8%	97.77%	82.33%	-	-

2.4 Summary

A summary of the methods reviewed above, both depth maps-based and skeleton-based, are presented in Table 3. Since all the reviewed methods are capable of dealing with both gestures and actions, only the capability of handling interactions is enumerated. The accuracy of the reviewed methods on the popular datasets are summarized in Table 2. Originally, Wang et al [8] performed three set of tests on the MSR Action 3D Dataset. The first two use one third and two thirds of the data for training, respectively, while the last one was designed for a cross-subject test. This evaluation method is adopted by most of the works that follow, as can be seen in the table. As the MSR Daily Activity Dataset is relatively new, not many methods were evaluated on it. The methods that were not evaluated on these datasets generally performed evaluations on other datasets that are not listed here.

In conclusion, with the excellent opportunities provided by the low-cost depth sensors for activity recognition, promising results have been achieved as evidenced in recent research work. The unique characteristics of the depth sensor data inspire the researchers to investigate effective and efficient approaches for this task, partly based on the traditional work on regular visual data. One line of ongoing research attempts to design more discriminative and meanwhile compact feature vectors from depth and skeleton data to describe human activities. Another possible direction is to extend the current methods to deal with more complicated activities, such as interactions or group activities. In this case, existing works that operate on regular visual data might provide some good insights [4].

Table 3. Summary of methods for action recognition based on data from depth sensors. Here "Seq" refers to "Sequential", "STV" refers to "Space Time Volume" and "Skel" means "Skeleton".

	Taxonomy	Features	Representation	Classifier	View-invariant	Scale-invariant	Interactions
[8]	Depth+STV	Bag of 3d points	2d Projection	Action graph	yes	yes	no
[13]	Depth+STV	STOP	PCA	Action graph	yes	yes	no
[14]	Depth+STV	ROP	Sparse Coding	SVM	yes	yes	no
[15]	Depth+STV	DMM + HOG		SVM	no	yes	no
[11]	Depth+STV	HON4D	Histogram	SVM	yes	yes	no
[16]	Depth+STV	4d Local Spatio-Temporal Features	PCA	Latent Dirichlet Allocation	no	no	yes
[18]	Depth+STV	SIFT-like	PCA + Bag of Words	SVM	no	no	yes
[19]	Depth+Seq	Depth silhouettes + $\mathcal{R}$ Transform		HMM	no	yes	no
[26]	Skel+Seq	3d joint trajectories	Projection in phase spaces	Similar to NN	yes	yes	no
[27]	Skel+Seq	Poses of single and multiple joints		HMM + AdaBoost	yes	yes	no
[21]	Skel+Seq	HOJ3D	Linear Discriminant Analysis	HMM	yes	yes	no
[30,31]	Skel+Seq	Object and Pose features		Multi-class SVM	no	no	yes
[34,35]	Skel+Seq	Pose features + HOG		MEMM	yes	yes	no
[9]	Skel+Seq	LOP	Actionlet	SVM	yes	yes	yes
[38]	Skel+Seq	Relational Pose Features		Hough Forest	yes	yes	no
[33]	Skel+STV	EigenJoints	PCA	NBNN	yes	yes	no

3 Face Motion

Human motion analysis is not restricted to full body motion, but can also be applied to body parts like the face or the hands. In this section, we give an overview of different approaches that capture head or facial motion at different

levels of details; see Figures 14, 18 and 19. The lowest level estimates the head pose only, i.e., location and orientation of the head. Approaches for head pose estimation are discussed in Section 3.1. Facial feature points or low-resolution shape models provide more information and are often extracted for applications like face recognition, speech recognition or analysis of facial expressions. While Section 3.2 discusses works for extracting facial feature points, Section 3.3 discusses methods that aim at capturing all details of facial motion. The latter is mainly used in the context of facial animations. Parts of this section appeared in [42].

3.1 Head Pose Estimation

With application ranging from face recognition to driver drowsiness detection, automatic head pose estimation is an important problem. Since the survey [43] gives already an excellent overview of approaches until the year 2007, we focus on more recent approaches for head pose estimation that appeared in 2007 or later. Although the focus is head pose estimation from depth data, we give a broader view that also includes methods that estimate the head pose from RGB data like images or videos. Methods based on 2d images can be subdivided into appearance-based and feature-based approaches, depending on whether they analyze the face as a whole or instead rely on the localization of some specific facial features for head pose estimation.

3.1.1 RGB Appearance-Based Methods

Appearance-based methods usually discretize the head pose space and learn separate detectors for subsets of poses [44,45]. Chen et al. [46] and Balasubramanian et al. [47] present head pose estimation systems with a specific focus on the mapping from the high-dimensional space of facial appearance to the lower-dimensional manifold of head poses. The latter work considers face images with varying poses as lying on a smooth low-dimensional manifold in a high-dimensional feature space. The proposed Biased Manifold Embedding uses the pose angle information of the face images to compute a biased neighborhood of each point in the feature space, prior to determining the low-dimensional embedding. In the same vein, Osadchy et al. [48] instead use a convolutional network to learn the mapping, achieving real-time performance for the face detection problem, while also providing an estimate of the head pose. A very popular family of methods use statistical models of the face shape and appearance, like Active Appearance Models (AAMs) [49], multi-view AAMs [50] and 3d Morphable Models [51,52]. Such methods, however, focus more on tracking facial features rather than estimating the head pose. In this context, the authors of [53] coupled an Active Appearance Model with the POSIT algorithm for head pose tracking.

3.1.2 RGB Feature-Based Methods

Feature-based methods rely on some specific facial features to be visible, and therefore are sensitive to occlusions and to large head rotations. Vatahska et

al. [54] use a face detector to roughly classify the pose as frontal, left, or right profile. After this, they detect the eyes and nose tip using AdaBoost classifiers. Finally, the detections are fed into a neural network which estimates the head orientation. Similarly, Whitehill et al. [55] present a discriminative approach to frame-by-frame head pose estimation. Their algorithm relies on the detection of the nose tip and both eyes, thereby limiting the recognizable poses to the ones where both eyes are visible. Morency et al. [56] propose a probabilistic framework called Generalized Adaptive View-based Appearance Model integrating frame-by-frame head pose estimation, differential registration and keyframe tracking.

3.1.3 Head Pose Estimation from Depth or 3D

In general, approaches relying solely on 2d images are sensitive to illumination changes and lack of distinctive features. Moreover, the annotation of head poses from 2d images is intrinsically problematic. Since 3d sensing devices have become available, computer vision researchers have started to leverage the additional depth information for solving some of the inherent limitations of image-based methods. Some of the recent works thus use depth as primary cue [57] or in addition to 2d images [58,59,60].

Seemann et al. [60] presented a neural network-based system fusing skin color histograms and depth information. It tracks at 10 fps but requires the face to be detected in a frontal pose in the first frame of the sequence. The approach in [61] uses head pose estimation only as a pre-processing step to face recognition, and the low reported average errors are only calculated on faces of subjects that belong to the training set. Still in a tracking framework, Morency et al. [59] use instead intensity and depth input images to build a prior model of the face using 3d view-based eigenspaces. Then, they use this model to compute the absolute difference in pose for each new frame. The pose range is limited and manual cropping is necessary. In [58], a 3d face model is aligned to an RGB-depth input stream for tracking features across frames, taking into account the very noisy nature of depth measurements coming from commercial sensors.

Considering instead pure detectors on a frame-by-frame basis, Lu and Jain [62] create hypotheses for the nose position in range images based on directional maxima. For verification, they compute the nose profile using PCA and a curvature-based shape index. Breitenstein et al. [57] presented a real-time system working on range scans provided by the scanner of [63]. Their system can handle large pose variations, facial expressions and partial occlusions, as long as the nose remains visible. Their method relies on several candidate nose positions, suggested by a geometric descriptor. Such hypotheses are all evaluated in parallel on a GPU, which compares them to renderings of a generic template with different orientations. Finally the orientation which minimizes a predefined cost function is selected. Breitenstein et al. also collected a dataset of over 10k annotated range scans of heads. The subjects, both male and female, with and without glasses, were recorded using the scanner of [63] while turning their heads around, trying to span all possible yaw and pitch rotation angles they could. The scans were semi-automatically annotated by a template-based tracking approach for head pose estimation [64] as illustrated in Figure 14. The tracker requires a user-specific head model that has

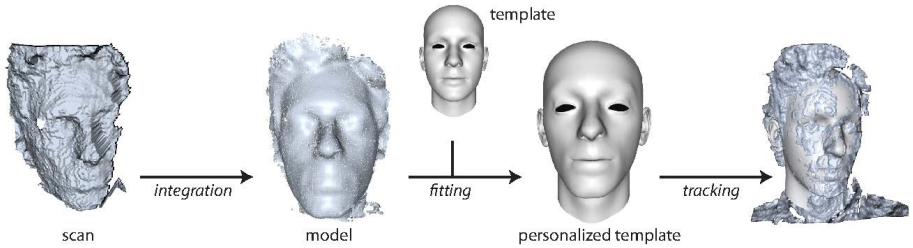
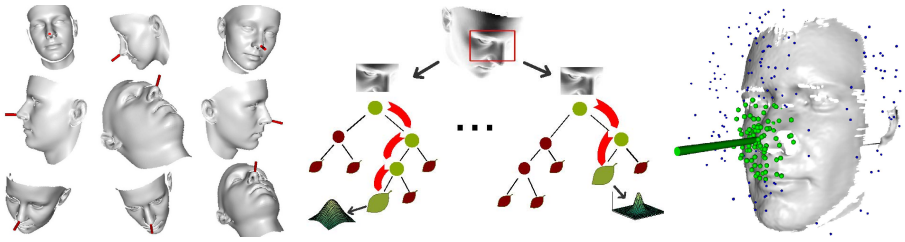


Fig. 14. Head pose tracking with a head template [64]. A user turns the head in front of the depth sensor, the scans are integrated into a point cloud model [69] and a generic template is fit to it using graph-based non-rigid ICP [70]. The personalized template is used for rigid tracking.



(a) Training data (b) Regression forest (c) Pose prediction

Fig. 15. Head pose estimation with regression forests [66]. (a) A head model is used to generate a large set of training data. (b) Based on the training data, a forest of regression trees is trained. Each tree takes a depth patch as input and regresses the pose parameters. (c) The regressed values of each patch can be considered as votes for the pose parameters. The final estimate is obtained by mean-shift.

been acquired before recording the dataset. The same authors also extended their system to use lower quality depth images from a stereo system [65].

While GPUs allow the evaluation of many hypotheses in real-time, they are not available for embedded systems where power consumption matters. In order to achieve real-time performance without the need of a GPU and to be robust to occlusions, a random forests framework for head pose estimation from depth data has been employed in [66]. The approach is illustrated in Figure 15. In [67], the approach has been further extended to handle noisy sensor data and a dataset with annotated head pose has been collected. The dataset comprises 24 sequences of 20 different subjects (14 men and 6 women, 4 subjects with glasses) that move their heads while sitting about 1 meter away from a Kinect sensor. Some examples of the dataset are shown in Figure 16. The biggest advantage of depth data for head pose estimation in comparison to 2d data is the simplicity of generating an abundance of training data with perfect ground truth. In [66], depth images of head poses are synthetically generated by rendering the 3d morphable model of [68].

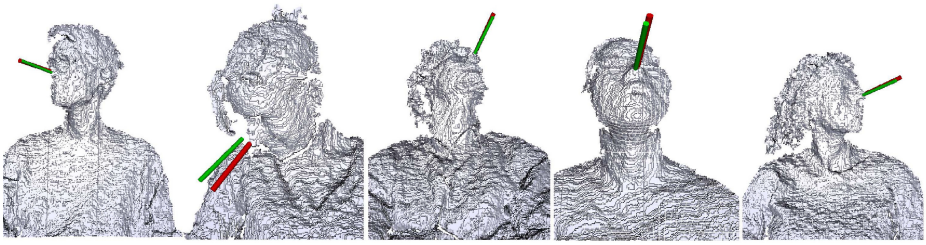


Fig. 16. Database for benchmarking head pose estimation from depth data [67]. The green cylinder represents the estimated head pose, while the red one encodes the ground truth.

3.2 Facial Feature Detection

3.2.1 Facial Feature Detection from 2D Data

Facial feature detection from standard images is a well studied problem, often performed as preprocessing for face recognition. Previous contributions can be classified into two categories, depending on whether they use global or local features. Holistic methods, e.g., Active Appearance Models [49,71,72], use the entire facial texture to fit a generative model to a test image. As discussed in Section 3.1, they can also be used for head pose estimation. They are usually affected by lighting changes and a bias towards the average face. The complexity of the modeling is an additional issue. Moreover, these methods perform poorly on unseen identities [73] and cannot handle low-resolution images well.

In recent years, there has been a shift towards methods based on independent local feature detectors [74,75,76,77]. These detectors are discriminative models of image patches centered around facial landmarks. To improve accuracy and reduce the influence of inaccurate detections and false positives, global models of the facial features configuration like pictorial structures [78,79] or Active Shape Models [80] can be used.

3.2.2 Facial Feature Detection from 3D Data

Similar to the 2d case, methods focusing on facial feature localization from range data can be subdivided into categories using global or local information. Among the former class, the authors of [81] deform a bi-linear face model to match a scan of an unseen face in different expressions. Yet, the paper's focus is not on the localization of facial feature points and real-time performance is not achieved. Also Kakadiaris et al. [82] non-rigidly align an annotated model to face meshes. However, constraints need to be imposed on the initial face orientation. Using high quality range scans, Weise et al. [83] presented a real-time system that is capable of tracking facial motion in detail, but requires personalized templates. The same approach has been extended to robustly track head pose and facial deformations using RGB-depth streams provided by commercial sensors like the Kinect [64].

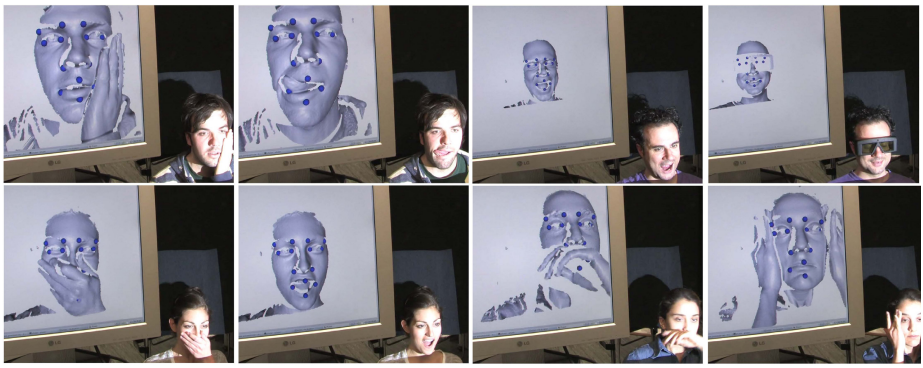


Fig. 17. Real-time facial feature localization using depth from a structured light system as input [42]

Most works that try to directly localize specific feature points from 3d data take advantage of surface curvatures. For example, the authors of [84,85,86] all use curvature to roughly localize the inner corners of the eyes. Such an approach is very sensitive to missing depth data, particularly for the regions around the inner eye corners that are frequently occluded by shadows. Also, Mehryar et al. [87] use surface curvatures by first extracting ridge and valley points, which are then clustered. The clusters are refined using a geometric model imposing a set of distance and angle constraints on the arrangement of candidate landmarks. Colbry et al. [88] use curvature in conjunction with the Shape Index proposed by [89] to locate facial feature points from range scans of faces. The reported execution time of this anchor point detector is 15 seconds per frame. Wang et al. [90] use point signatures [91] and Gabor filters to detect some facial feature points from 3d and 2d data. The method needs all desired landmarks to be visible, thus restricting the range of head poses while being sensitive to occlusions. Yu et al. [92] use genetic algorithms to combine several weak classifiers into a 3d facial landmark detector. Fanelli et al. [42] proposed a real-time system that relies on random forests for localizing fiducials. The system is shown in Figure 17. Ju et al. [93] detect the nose tip and the eyes using binary neural networks, and propose a 3d shape descriptor invariant to pose and expression.

The authors of [94] propose a 3d Statistical Facial Feature Model (SFAM), which models both the global variations in the morphology of the face and the local structures around the landmarks. The low reported errors for the localization of 15 points in scans of neutral faces come at the expense of processing time: over 10 minutes are needed to process one facial scan. In [95], fitting the proposed PCA shape model containing only the upper facial features, i.e., without the mouth, takes on average 2 minutes per face.

For evaluating facial feature detectors on depth data, there are two datasets available. $BU3DFE$ [97] contains 100 subjects (56 females and 44 males) posing six basic expressions plus neutral in front of a 3d face scanner. Each of the six prototypic expressions (happiness, disgust, fear, angry, surprise and sadness)

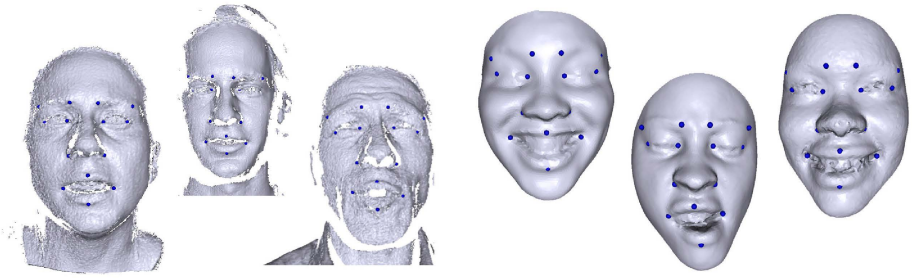


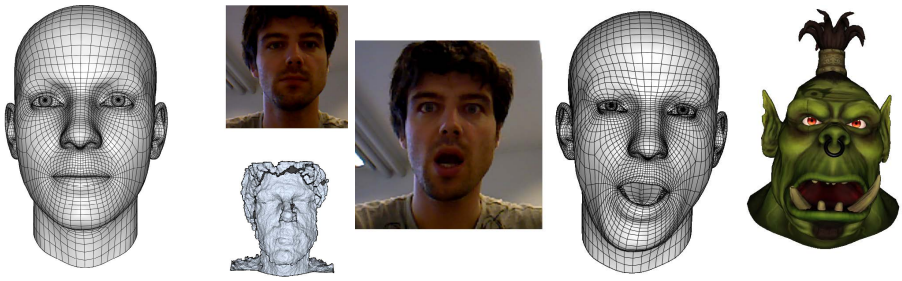
Fig. 18. Successfully localized facial features using the approach [42] on some test scans from the $B3D(AC)^2$ database [96] (left) and the $BU3DFE$ dataset [97] (right)

includes four levels of intensity, i.e., there are 25 static 3d expression models for each subject, resulting in a total of 2500 faces. Each face is annotated with 83 facial points. $B3D(AC)^2$ [96] comprises over 120k depth images and includes 14 subjects, 8 females and 6 males, repeating sentences from 40 movie sequences, both in a neutral and in an induced emotional state.

3.3 Facial Performance Capture

Facial performance capture goes beyond simple shape models or feature point detection and aims at capturing the full geometry of the face, mainly for facial animations. A typical application is performance-based facial animation where the non-rigid surface of the face is tracked and the motion is transferred to a virtual face [98,99]. Most of the work has focused so far on the acquisition of high-quality data using structured light systems [100,101,102,103,104] or passive multi-camera systems [105,106,101,107] in controlled setups. These methods propose different acquisition setups that are optimized for acquisition time, acquisition accuracy, or budget.

There are a few works that go beyond capturing facial motion in studio environments. The approach [108] uses a time-of-flight camera to estimate a few basic facial action units based on the Facial Action Coding System (FACS). The method fits a high-resolution statistical 3d expression morph model to the noisy depth data by an iterative closest point algorithm and regresses the action units from the fitted model. The method [83] achieves real-time performance-based facial animation by generating a user-specific facial expression model offline. During tracking the PCA components of the expression model are estimated and transferred to a PCA model of a target face in real-time. In [64], a robust method based on user-specific blendshapes has been proposed for real-time facial performance capture and animation. In contrast to most other works for facial animation, the approach also works with noisy depth data. The approach is illustrated in Figure 19.



(a) Model generation (b) Real-time capture and facial animation

Fig. 19. (a) Data from a depth sensor is used to build a user-specific blendshape model. (b) Having build the model, the motion can be transferred to a virtual head in real-time. ©2013 Faceshift AG `http://www.faceshift.com`.

3.4 Summary

Capturing facial motion from depth data has progressed fast in the last years and several real-time systems for different applications have been developed. While for some applications head pose estimation might be sufficient, more details like facial feature points, facial action units, or full geometry can be captured. Interestingly, the richer output does not necessarily require much higher computational cost, still allowing real-time performance, but it requires more effort for acquiring training data or an additional offline acquisition process, e.g., to acquire a user-specific model. While the discussed methods already perform well in terms of runtime and accuracy, there is further room for improving the accuracy without compromising runtime. For evaluation, several datasets have been released as shown in Table 4. Although each dataset has been recorded for a specific task like head pose estimation [57,42], facial expression recognition [97,109,110], face recognition [111,112,113], or audiovisual speech recognition [96], they can be also used for benchmarking methods for other tasks. Current datasets and methods assume that the head is clearly visible, the handling of crowded scenes for instance with many faces has not been addressed so far.

4 Hand Motion

Capturing the motion of hands shares many similarities with full body pose estimation. However, hands impose some additional challenges like uniform skin color, very large pose variations and severe occlusions that are even difficult to resolve from depth data. Since hands interact with other hands or objects nearly all the time, capturing hand motion is still a very challenging task. Parts of this section appeared in [114].

4.1 Hand Pose Estimation

In the survey [115], various methods for hand pose estimation have been reviewed in the context of human-computer interaction. We also follow the taxonomy used

Table 4. Datasets for evaluating depth-based approaches for head pose estimation and facial feature detection

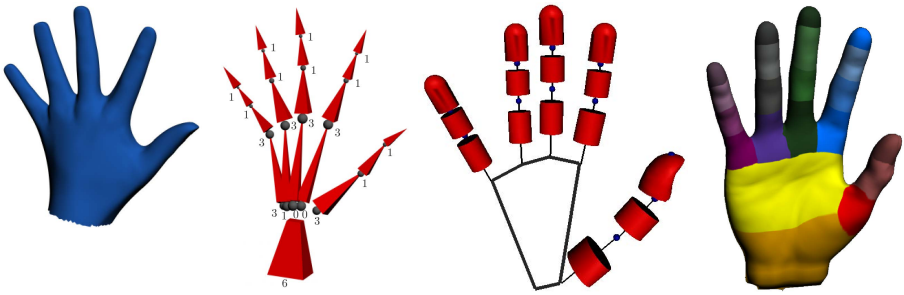
Dataset	Annotation	Data	Subjects
ETH Face Pose Range Image [57]	head pose	10k depth	20
Biwi Kinect Head Pose [42]	head pose	15k RGBD	20
Binghamton 3D Facial Expression [97]	6 facial expressions, 3 facial points	3k 3d models	100
Bosphorus Database [109]	24 facial points, FACS	5k 3d models	105
3D Dynamic Facial Expression [110]	6 facial expressions, 83 facial points	60k 3d models	101
Texas 3D Face Recognition [111]	25 facial points	1k RGBD	105
Biwi 3D Audiovisual Corpus [96]	face model, emotions, segmented speech	120k RGBD	14
UMB 3D Face [112]	7 facial points	1k RGBD	143
EURECOM Kinect Face [113]	6 facial points	1k RGBD	52

in [115] that splits the approaches in discriminative methods that use classification or regression to estimate the hand pose from image data and generative methods that use a hand model to recover the hand pose.

The model-based approaches mainly differ in the used cues and techniques for solving the problem. The most commonly used image features are silhouettes and edges, but also other cues like shading, color, or optical flow have been used [115]. For instance, edges, optical flow and shading information have been combined in [116] for articulated hand tracking. In [117], a method based on texture and shading has been proposed. A very important cue is depth [118,119] that has been recently revisited in the context of depth sensors [120].

In order to recover the hand pose based on some cues, several techniques have been proposed. One of the first methods for 3d hand pose estimation used local optimization [121], which is still a very popular method due to its efficiency, but it also requires a careful design of the objective function to avoid that the method gets stuck in local minima [114,117]. Other methods rely on filtering techniques like Kalman filter [122] or particle filter [123]. While particle filtering and local stochastic optimization have been combined in [119] to improve the performance of filtering techniques in the high-dimensional space of hand poses, [124,125] proposed to reduce the pose space by using linear subspaces. The methods, however, considered only very few hand poses. Other methods rely on belief propagation [126,127] or particle swarm optimization [128].

Depending on the method, the used hand models also differ as shown in Figure 20. The highest accuracy is achieved with user-specific hand models, e.g., [114,130]. These models need to be acquired offline, similar to user-specific head models, but anatomical properties like fix limb length are retained during tracking. More flexible are graphical models that connect limbs modeled by shape primitives and use often Gaussian distributions to model the allowed distance of two limbs, e.g., [126,127]. For each limb a likelihood is computed and the best hand configuration is inferred from the graphical model connecting the



(a) Articulated Hand Model (b) Connected Parts (c) Labeled Parts

Fig. 20. Different models for hand pose estimation: (a) Detailed 3d mesh with underlying skeleton [114] (b) Connected body parts [127] (c) Labeled surface for training a body part detector [129]

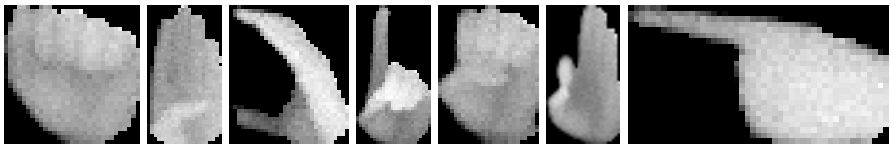
limbs. A hand model based on a self-organizing map [131] is discussed in the chapter *Gesture Interfaces with Depth Sensors*.

Discriminative approaches like [132,133,134] do not require an explicit hand model, but learn a mapping from image features to hand poses from a large set of training data. Although these methods process the frames independently, temporal consistency can be enforced [135,136,137]. While discriminative methods can recover from errors, the accuracy and type of hand poses that can be handled depends on the training data. Discriminative approaches that process the full hand are therefore not suitable for applications that require accurate hand poses of a-priori unknown actions. However, instead of learning a mapping for the full hand, also a mapping only for body parts can be learned [129] as shown in Figure 20(c). Breaking the hand into parts has the advantage that a larger variation of poses can be handled. Similar approaches have been successfully applied to human pose [3] or head pose estimation [66].

Recently, the focus has been on hand motion capture in the context of interactions. [127] has considered hand tracking from depth data in the context of object manipulation. While the objects were originally treated as occluders, [139] proposed to learn an object dependent hand pose prior to assist tracking. The method assumes that object manipulations of similar objects have been previously observed for training and exploits contact points and geometry of the objects. Such dependencies can also be used to create hand animations [140,141] as shown in Figure 21. In the context of object grasping, a database of 10 000 hand poses with and without grasped objects has been created to recover the hand pose from images using nearest neighbor search [136]. Recently, it has been proposed to track the manipulated object and the hand at the same time to constrain the search space using collision constraints [142]. The object is assumed to be a shape primitive like a cylinder, ellipsoid or box whose parameters are estimated. In [130], particle swarm optimization (PSO) has been applied to hand tracking of two interacting hands from depth data.

(a) Skeleton (b) Hand Model (c) Skeleton (d) Hand Model

Fig. 21. The relations between hands and object classes can be modeled to synthesize hand poses or hand-object animations [138].



(a) (b) (c) (d) (e) (f) (g)

Fig. 22. Alphabet *(A-G)* of the American sign language captured with a ToF camera

Instead of using off-the-shelf depth sensors, some approaches have focused on the sensor design in the context of human-computer interaction (HCI) applications. For instance, Leap Motion[1] developed a controller that allows to capture the motion of finger tips with high accuracy. While the volume that can be captured by the controller is very small, a wrist-worn sensor for hand pose estimation has been proposed in [143]. In [144], RGBD data is used to improve a marker-based system. A more detailed overview of approaches for HCI, including commercial systems, is given in the chapter *Gesture Interfaces with Depth Sensors*.

4.2 Hand Gesture Recognition

Even if the resolution of the hands is too small to estimate the full articulated hand, the gesture of the hand can still be estimated given a suitable training set. In this section, we give an overview of different methods for recognizing hand gestures, in particular letters of a sign language, as shown in Figure 22. Parts of this section appeared in [140].

Recognizing signs of visual-gestural languages like the American sign language (ASL) is a very active field [145,146,147,148,149,150,151]. For instance, the SignSpeak project [152] aims at developing vision-based technology for translating continuous sign language to text. It is also related to gesture recognition from depth data and optional color data [153,154,155,156,157,158,159,160,161,162], which is discussed in the chapter *Gesture Interfaces with Depth Sensors*. While for gesture recognition usually only a small set of distinctive gestures are used, the letters of a sign language are pre-defined and not very distinctive in low

[1] `http://www.leapmotion.com`

resolution depth images. In the following, we structure the methods into user-specific systems, i.e., the systems are trained and designed for a specific user, and general systems, i.e., the user does not provide any training data:

4.2.1 User-Specific Systems

Polish finger alphabet symbols have been classified in [163] with an off-line setup. The input for each of the considered 23 gestures consisted of a gray-scale image at a relatively high resolution and depth data acquired by a stereo setup. In [164], a real-time recognition system has been developed for Spanish sign language letters where a colored glove was used. The real-time system [165] recognizes 46 gestures including symbols of the ASL. It assumes constant lighting conditions for training and testing and uses a wristband and special background for accurate hand segmentation. More recently, British sign language finger spelling has been investigated in [166] where the specialty is that both hands are involved in the 26 static gestures. Working on skin color, it is assumed that the signer wears suitable clothing and the background is of a single uniform color. The system recognizes also spelled words contained in a pre- defined lexicon.

4.2.2 General Systems

Using a stereo camera to acquire 3d and color data, Takimoto et al. [160] proposed a method for recognizing 41 Japanese sign language characters. Data was acquired from 20 test subjects and the achieved classifier runtime is about 3 frames per second. Although the approach does not require special background or lighting conditions, segmenting the hand, which is a challenging task by itself, is greatly simplified by the use of a black wristband. Colored gloves have been used in [167] for recognizing 23 symbols of the Irish sign language in real-time. A method for recognizing the ASL finger alphabet off-line has been proposed in [168]. Input data was acquired in front of a white background and the hand bounding box was defined for each image manually. A similar setup has been used in [169]. In [140], a method based on average neighborhood margin maximation has been proposed that recognizes the ASL finger alphabet from low-resolution depth data in real-time.

The methods are summarized in Table 5.

4.3 Summary

In contrast to activity recognition and facial motion capture, there is a lack of publicly available datasets for benchmarking and comparing methods for hand pose estimation. Even for RGB data, there are very few datasets that provide ground-truth data [170]. While many depth datasets have been used for hand gesture recognition or recognizing finger alphabet symbols, there is no dataset available that has been consistently used. As shown in Table 5, many methods use a different number of gestures or recording setups. In order to evaluate the progress in this area, publicly available datasets with ground-truth data are needed.

Table 5. Overview of methods for recognizing hand gestures

Method	# of Gest.	Setup	Depth	Resolution	Markers	Real-time
[165]	46	user-specific	no	320x240	wristband	yes
[159]	11	user-specific	yes	160x120		yes
[163]	23	user-specific	yes	320x240 768x576(gray)	black long sleeve	no
[164]	19	user-specific	no		colored glove	yes
[162]	6	user-specific	yes	176x144 640x480(rgb)		yes
[140]	26	user-specific	yes	176x144		yes
[157]	12	general	yes	160x120		yes
[158]	6	general	yes	176x144		yes
[156]	5	general	yes	176x144		yes
[160]	41	general	yes	320x240 1280x960(rgb)	wristband	no
[167]	23	general	no		color glove	yes
[168]	26	general	no		bounding box	no
[169]	26	general	no		bounding box	no
[140]	26	general	yes	176x144		yes

5 Conclusion

Over the last years, a significant progress has been made in the field of human motion analysis from depth data. The success is attested by commercial systems that estimate full body poses for computer games, hand poses for gesture interfaces, or capture detailed head motions for facial animations. It is expected that more approaches in the field will make the transition from the lab to a business. The main advantages of developing applications for depth sensors compared to purely 2d color sensors are (i) the better robustness to lighting conditions, at least in indoor environments, (ii) the resolved scale-distance ambiguity of 2d sensors, making it easier to develop real-time algorithms, (iii) the possibility to synthesize an abundance of training data. Nevertheless, there are still many research challenges that need to be addressed and that cannot be resolved by improving only the data quality provided by the sensors. So far, the most successful approaches for capturing full body motion or specific body parts like hands or the head assume that the subjects are within a specific distance range to the sensor. Dealing with a larger range of distances, however, requires to smoothly blend between analyzing full body motion and the motion of body parts. If the person is far away from the sensor, full body motion can be better analyzed than facial motion. As soon as the person gets closer to the sensor, only parts of the person remain visible and motion analysis is limited to the upper body, hands, or the face. For some applications, even all aspects of human body language need to be taken into account to understand the intend of the user. In sign languages, for instance, it is not only hand gestures that matter, but also the motion of the arms, facial expressions and the movements of the lips. Bring-

ing all these components of motion analysis together, which have been mainly addressed independently, is a big challenge for the future. Another challenge is motion analysis in the context of crowded scenes and interactions. While first approaches address the problem of human-human or human-object interactions, more work needs to be done in this area to achieve performances that are good enough for real-world applications.

References

1. Klette, R., Tee, G.: Understanding human motion: A historic review. In: Rosenhahn, B., Klette, R., Metaxas, D. (eds.) Human Motion. Computational Imaging and Vision, vol. 36, pp. 1–22. Springer, Netherlands (2008)
2. Aggarwal, J.: Motion analysis: Past, present and future. In: Bhanu, B., Ravishankar, C.V., Roy-Chowdhury, A.K., Aghajan, H., Terzopoulos, D. (eds.) Distributed Video Sensor Networks, pp. 27–39. Springer, London (2011)
3. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. In: IEEE Conference on Computer Vision and Pattern Recognition (2011)
4. Aggarwal, J., Ryoo, M.: Human activity analysis: A review. ACM Computing Surveys 43(2), 16:1–16:43 (2011)
5. Mitra, S., Acharya, T.: Gesture recognition: A survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 37(3), 311–324 (2007)
6. Moeslund, T.B., Hilton, A., Krüger, V.: A survey of advances in vision-based human motion capture and analysis. Computer Vision and Image Understanding 104(2), 90–126 (2006)
7. Poppe, R.: A survey on vision-based human action recognition. Image and Vision Computing 28(6), 976–990 (2010)
8. Li, W., Zhang, Z., Liu, Z.: Action recognition based on a bag of 3d points. In: Workshop on Human Activity Understanding from 3D Data, pp. 9–14 (2010)
9. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Mining actionlet ensemble for action recognition with depth cameras. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1290–1297 (2012)
10. Kurakin, A., Zhang, Z., Liu, Z.: A real time system for dynamic hand gesture recognition with a depth sensor. In: 2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO), pp. 1975–1979 (2012)
11. Oreifej, O., Liu, Z.: Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In: IEEE Conference on Computer Vision and Pattern Recognition (2013)
12. Li, W., Zhang, Z., Liu, Z.: Expandable data-driven graphical modeling of human actions based on salient postures. IEEE Transactions on Circuits and Systems for Video Technology 18(11), 1499–1510 (2008)
13. Vieira, A.W., Nascimento, E.R., Oliveira, G.L., Liu, Z., Campos, M.F.M.: STOP: Space-time occupancy patterns for 3D action recognition from depth map sequences. In: Alvarez, L., Mejail, M., Gomez, L., Jacobo, J. (eds.) CIARP 2012. LNCS, vol. 7441, pp. 252–259. Springer, Heidelberg (2012)
14. Wang, J., Liu, Z., Chorowski, J., Chen, Z., Wu, Y.: Robust 3D action recognition with random occupancy patterns. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part II. LNCS, vol. 7573, pp. 872–885. Springer, Heidelberg (2012)

15. Yang, X., Zhang, C., Tian, Y.: Recognizing actions using depth motion maps-based histograms of oriented gradients. In: ACM International Conference on Multimedia, pp. 1057–1060 (2012)
16. Zhang, H., Parker, L.: 4-dimensional local spatio-temporal features for human activity recognition. In: International Conference on Intelligent Robots and Systems, pp. 2044–2049 (2011)
17. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America 101(suppl. 1), 5228–5235 (2004)
18. Lei, J., Ren, X., Fox, D.: Fine-grained kitchen activity recognition using rgb-d. In: ACM Conference on Ubiquitous Computing (2012)
19. Jalal, A., Uddin, M.Z., Kim, J.T., Kim, T.S.: Recognition of human home activities via depth silhouettes and transformation for smart homes. Indoor and Built Environment 21(1), 184–190 (2011)
20. Wang, Y., Huang, K., Tan, T.: Human activity recognition based on r transform. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2007)
21. Xia, L., Chen, C.C., Aggarwal, J.: View invariant human action recognition using histograms of 3d joints. In: Workshop on Human Activity Understanding from 3D Data, pp. 20–27 (2012)
22. Han, L., Wu, X., Liang, W., Hou, G., Jia, Y.: Discriminative human action recognition in the learned hierarchical manifold space. Image and Vision Computing 28(5), 836–849 (2010)
23. Johansson, G.: Visual motion perception. Scientific American (1975)
24. Ye, M., Wang, X., Yang, R., Ren, L., Pollefeys, M.: Accurate 3d pose estimation from a single depth image. In: IEEE International Conference on Computer Vision, pp. 731–738 (2011)
25. Criminisi, A., Shotton, J., Robertson, D., Konukoglu, E.: Regression forests for efficient anatomy detection and localization in CT studies. In: Menze, B., Langs, G., Tu, Z., Criminisi, A. (eds.) MICCAI 2010. LNCS, vol. 6533, pp. 106–117. Springer, Heidelberg (2011)
26. Campbell, L., Bobick, A.: Recognition of human body motion using phase space constraints. In: IEEE International Conference on Computer Vision, pp. 624–630 (1995)
27. Lv, F., Nevatia, R.: Recognition and segmentation of 3-D human action using HMM and multi-class adaBoost. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3954, pp. 359–372. Springer, Heidelberg (2006)
28. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 55(1), 119–139 (1997)
29. Lee, M.W., Nevatia, R.: Dynamic human pose estimation using markov chain monte carlo approach. In: IEEE Workshops on Application of Computer Vision, pp. 168–175 (2005)
30. Koppula, H.S., Gupta, R., Saxena, A.: Human activity learning using object affordances from rgb-d videos. CoRR abs/1208.0967 (2012)
31. Koppula, H.S., Gupta, R., Saxena, A.: Learning human activities and object affordances from rgb-d videos. CoRR abs/1210.1207 (2012)
32. Lai, K., Bo, L., Ren, X., Fox, D.: Sparse distance learning for object recognition combining rgb and depth information. In: International Conferences on Robotics and Automation, pp. 4007–4013 (2011)

33. Yang, X., Tian, Y.: Eigenjoints-based action recognition using naive-bayes-nearest-neighbor. In: Workshop on Human Activity Understanding from 3D Data, pp. 14–19 (2012)
34. Sung, J., Ponce, C., Selman, B., Saxena, A.: Human activity detection from rgbd images. In: Plan, Activity, and Intent Recognition (2011)
35. Sung, J., Ponce, C., Selman, B., Saxena, A.: Unstructured human activity detection from rgbd images. In: IEEE International Conference on Robotics and Automation, pp. 842–849 (2012)
36. McCallum, A., Freitag, D., Pereira, F.C.N.: Maximum entropy markov models for information extraction and segmentation. In: International Conference on Machine Learning, pp. 591–598 (2000)
37. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 886–893 (2005)
38. Yao, A., Gall, J., Van Gool, L.: Coupled action recognition and pose estimation from multiple views. International Journal of Computer Vision 100(1), 16–37 (2012)
39. Müller, M., Röder, T., Clausen, M.: Efficient content-based retrieval of motion capture data. ACM Transactions on Graphics 24, 677–685 (2005)
40. Gall, J., Yao, A., Razavi, N., Van Gool, L., Lempitsky, V.: Hough forests for object detection, tracking, and action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2011)
41. Tenorth, M., Bandouch, J., Beetz, M.: The TUM kitchen data set of everyday manipulation activities for motion tracking and action recognition. In: IEEE Workshop on Tracking Humans for the Evaluation of their Motion in Image Sequences (2009)
42. Fanelli, G., Dantone, M., Gall, J., Fossati, A., Van Gool, L.: Random forests for real time 3d face analysis. International Journal of Computer Vision 101(3), 437–458 (2013)
43. Murphy-Chutorian, E., Trivedi, M.: Head pose estimation in computer vision: A survey. Transactions on Pattern Analysis and Machine Intelligence 31(4), 607–626 (2009)
44. Jones, M., Viola, P.: Fast multi-view face detection. Technical Report TR2003-096, Mitsubishi Electric Research Laboratories (2003)
45. Huang, C., Ding, X., Fang, C.: Head pose estimation based on random forests for multiclass classification. In: International Conference on Pattern Recognition (2010)
46. Chen, L., Zhang, L., Hu, Y., Li, M., Zhang, H.: Head pose estimation using fisher manifold learning. In: Analysis and Modeling of Faces and Gestures (2003)
47. Balasubramanian, V.N., Ye, J., Panchanathan, S.: Biased manifold embedding: A framework for person-independent head pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition (2007)
48. Osadchy, M., Miller, M.L., LeCun, Y.: Synergistic face detection and pose estimation with energy-based models. In: Neural Information Processing Systems (2005)
49. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. IEEE Transactions on Pattern Analysis and Machine Intelligence 23, 681–685 (2001)
50. Ramnath, K., Koterba, S., Xiao, J., Hu, C., Matthews, I., Baker, S., Cohn, J., Kanade, T.: Multi-view aam fitting and construction. International Journal of Computer Vision 76(2), 183–204 (2008)

51. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: ACM International Conference on Computer Graphics and Interactive Techniques, pp. 187–194 (1999)
52. Storer, M., Urschler, M., Bischof, H.: 3d-mam: 3d morphable appearance model for efficient fine head pose estimation from still images. In: Workshop on Subspace Methods (2009)
53. Martins, P., Batista, J.: Accurate single view model-based head pose estimation. In: Automatic Face and Gesture Recognition (2008)
54. Vatahska, T., Bennewitz, M., Behnke, S.: Feature-based head pose estimation from images. In: International Conference on Humanoid Robots (2007)
55. Whitehill, J., Movellan, J.R.: A discriminative approach to frame-by-frame head pose tracking. In: Automatic Face and Gesture Recognition (2008)
56. Morency, L.P., Whitehill, J., Movellan, J.R.: Generalized adaptive view-based appearance model: Integrated framework for monocular head pose estimation. In: Automatic Face and Gesture Recognition (2008)
57. Breitenstein, M.D., Kuettel, D., Weise, T., Van Gool, L., Pfister, H.: Real-time face pose estimation from single range images. In: IEEE Conference on Computer Vision and Pattern Recognition (2008)
58. Cai, Q., Gallup, D., Zhang, C., Zhang, Z.: 3D deformable face tracking with a commodity depth camera. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part III. LNCS, vol. 6313, pp. 229–242. Springer, Heidelberg (2010)
59. Morency, L.P., Sundberg, P., Darrell, T.: Pose estimation using 3d view-based eigenspaces. In: Automatic Face and Gesture Recognition (2003)
60. Seemann, E., Nickel, K., Stiefelhagen, R.: Head pose estimation using stereo vision for human-robot interaction. In: Automatic Face and Gesture Recognition (2004)
61. Mian, A., Bennamoun, M., Owens, R.: Automatic 3d face detection, normalization and recognition. In: 3D Data Processing, Visualization, and Transmission (2006)
62. Lu, X., Jain, A.K.: Automatic feature extraction for multiview 3d face recognition. In: Automatic Face and Gesture Recognition (2006)
63. Weise, T., Leibe, B., Van Gool, L.: Fast 3d scanning with automatic motion compensation. In: IEEE Conference on Computer Vision and Pattern Recognition (2007)
64. Weise, T., Bouaziz, S., Li, H., Pauly, M.: Realtime performance-based facial animation. ACM Transactions on Graphics 30(4) (2011)
65. Breitenstein, M.D., Jensen, J., Høilund, C., Moeslund, T.B., Van Gool, L.: Head pose estimation from passive stereo images. In: Salberg, A.-B., Hardeberg, J.Y., Jenssen, R. (eds.) SCIA 2009. LNCS, vol. 5575, pp. 219–228. Springer, Heidelberg (2009)
66. Fanelli, G., Gall, J., Van Gool, L.: Real time head pose estimation with random regression forests. In: IEEE Conference on Computer Vision and Pattern Recognition (2011)
67. Fanelli, G., Weise, T., Gall, J., Van Gool, L.: Real time head pose estimation from consumer depth cameras. In: Mester, R., Felsberg, M. (eds.) DAGM 2011. LNCS, vol. 6835, pp. 101–110. Springer, Heidelberg (2011)
68. Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T.: A 3d face model for pose and illumination invariant face recognition. In: Advanced Video and Signal based Surveillance (2009)
69. Weise, T., Wismer, T., Leibe, B., Van Gool, L.: In-hand scanning with online loop closure. In: 3-D Digital Imaging and Modeling (2009)
70. Li, H., Adams, B., Guibas, L.J., Pauly, M.: Robust single-view geometry and motion reconstruction. ACM Transactions on Graphics 28(5) (2009)

71. Cootes, T.F., Wheeler, G.V., Walker, K.N., Taylor, C.J.: View-based active appearance models. Image and Vision Computing 20(9-10), 657–664 (2002)
72. Matthews, I., Baker, S.: Active appearance models revisited. International Journal of Computer Vision 60(2), 135–164 (2003)
73. Gross, R., Matthews, I., Baker, S.: Generic vs. person specific active appearance models. Image and Vision Computing 23(12), 1080–2093 (2005)
74. Valstar, M., Martinez, B., Binefa, X., Pantic, M.: Facial point detection using boosted regression and graph models. In: IEEE Conference on Computer Vision and Pattern Recognition (2010)
75. Amberg, B., Vetter, T.: Optimal landmark detection using shape models and branch and bound slides. In: IEEE International Conference on Computer Vision (2011)
76. Belhumeur, P.N., Jacobs, D.W., Kriegman, D.J., Kumar, N.: Localizing parts of faces using a consensus of exemplars. In: IEEE Conference on Computer Vision and Pattern Recognition (2011)
77. Dantone, M., Gall, J., Fanelli, G., Van Gool, L.: Real-time facial feature detection using conditional regression forests. In: IEEE Conference on Computer Vision and Pattern Recognition (2012)
78. Felzenszwalb, P.F., Huttenlocher, D.P.: Pictorial structures for object recognition. International Journal of Computer Vision 61(1), 55–79 (2005)
79. Everingham, M., Sivic, J., Zisserman, A.: Hello! my name is... buffy - automatic naming of characters in tv video. In: British Machine Vision Conference (2006)
80. Cristinacce, D., Cootes, T.: Automatic feature localisation with constrained local models. Journal of Pattern Recognition 41(10), 3054–3067 (2008)
81. Mpiperis, I., Malassiotis, S., Strintzis, M.: Bilinear models for 3-d face and facial expression recognition. IEEE Transactions on Information Forensics and Security 3(3), 498–511 (2008)
82. Kakadiaris, I.A., Passalis, G., Toderici, G., Murtuza, M.N., Lu, Y., Karampatziakis, N., Theoharis, T.: Three-dimensional face recognition in the presence of facial expressions: An annotated deformable model approach. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(4), 640–649 (2007)
83. Weise, T., Li, H., Van Gool, L., Pauly, M.: Face/off: live facial puppetry. In: Symposium on Computer Animation, pp. 7–16 (2009)
84. Sun, Y., Yin, L.: Automatic pose estimation of 3d facial models. In: International Conference on Pattern Recognition (2008)
85. Segundo, M., Silva, L., Bellon, O., Queirolo, C.: Automatic face segmentation and facial landmark detection in range images. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 40(5), 1319–1330 (2010)
86. Chang, K.I., Bowyer, K.W., Flynn, P.J.: Multiple nose region matching for 3d face recognition under varying facial expression. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(10), 1695–1700 (2006)
87. Mehryar, S., Martin, K., Plataniotis, K., Stergiopoulos, S.: Automatic landmark detection for 3d face image processing. In: Evolutionary Computation (2010)
88. Colbry, D., Stockman, G., Jain, A.: Detection of anchor points for 3d face verification. In: IEEE Conference on Computer Vision and Pattern Recognition (2005)
89. Dorai, C., Jain, A.K.: COSMOS - A Representation Scheme for 3D Free-Form Objects. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(10), 1115–1130 (1997)
90. Wang, Y., Chua, C., Ho, Y.: Facial feature detection and face recognition from 2d and 3d images. Pattern Recognition Letters 10(23), 1191–1202 (2002)

91. Chua, C.S., Jarvis, R.: Point signatures: A new representation for 3d object recognition. International Journal of Computer Vision 25, 63–85 (1997)
92. Yu, T.H., Moon, Y.S.: A novel genetic algorithm for 3d facial landmark localization. In: Biometrics: Theory, Applications and Systems (2008)
93. Ju, Q., O'keefe, S., Austin, J.: Binary neural network based 3d facial feature localization. In: International Joint Conference on Neural Networks (2009)
94. Zhao, X., Dellandréa, E., Chen, L., Kakadiaris, I.: Accurate landmarking of three-dimensional facial data in the presence of facial expressions and occlusions using a three-dimensional statistical facial feature model. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 41(5), 1417–1428 (2011)
95. Nair, P., Cavallaro, A.: 3-d face detection, landmark localization, and registration using a point distribution model. IEEE Transactions on Multimedia 11(4), 611–623 (2009)
96. Fanelli, G., Gall, J., Romsdorfer, H., Weise, T., Van Gool, L.: A 3-d audio-visual corpus of affective communication. IEEE Transactions on Multimedia 12(6), 591–598 (2010)
97. Yin, L., Wei, X., Sun, Y., Wang, J., Rosato, M.J.: A 3d facial expression database for facial behavior research. In: International Conference on Automatic Face and Gesture Recognition (2006)
98. Lewis, J.P., Pighin, F.: Background mathematics. In: ACM SIGGRAPH Courses (2006)
99. Alexander, O., Rogers, M., Lambeth, W., Chiang, M., Debevec, P.: The digital emily project: photoreal facial modeling and animation. In: ACM SIGGRAPH Courses (2009)
100. Zhang, S., Huang, P.: High-resolution, real-time 3d shape acquisition. In: Workshop on Real-time 3D Sensors and Their Use (2004)
101. Zhang, L., Snavely, N., Curless, B., Seitz, S.M.: Spacetime faces: high resolution capture for modeling and animation. ACM Transactions on Graphics 23(3), 548–558 (2004)
102. Borshukov, G., Piponi, D., Larsen, O., Lewis, J.P., Tempelaar-Lietz, C.: Universal capture - image-based facial animation for "the matrix reloaded". In: ACM SIGGRAPH Courses (2005)
103. Ma, W.C., Hawkins, T., Peers, P., Chabert, C.F., Weiss, M., Debevec, P.: Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination. In: Eurographics Conference on Rendering Techniques, pp. 183–194 (2007)
104. Wilson, C.A., Ghosh, A., Peers, P., Chiang, J.Y., Busch, J., Debevec, P.: Temporal upsampling of performance geometry using photometric alignment. ACM Transactions on Graphics 29(2) (2010)
105. Beeler, T., Bickel, B., Beardsley, P., Sumner, B., Gross, M.: High-quality single-shot capture of facial geometry. ACM Transactions on Graphics 29 (2010)
106. Bradley, D., Heidrich, W., Popa, T., Sheffer, A.: High resolution passive facial performance capture. ACM Transactions on Graphics 29(4) (2010)
107. Furukawa, Y., Ponce, J.: Dense 3d motion capture from synchronized video streams. In: IEEE Conference on Computer Vision and Pattern Recognition (2008)
108. Breidt, M., Buelthoff, H., Curio, C.: Robust semantic analysis by synthesis of 3d facial motion. In: Automatic Face and Gesture Recognition (2011)

109. Savran, A., Celiktutan, O., Akyol, A., Trojanová, J., Dibeklioglu, H., Esenlik, S., Bozkurt, N., Demirkir, C., Akagunduz, E., Caliskan, K., Alyuz, N., Sankur, B., Ulusoy, I., Akarun, L., Sezgin, T.M.: 3d face recognition performance under adversarial conditions. In: Workshop on Multimodal Interfaces, pp. 87–102 (2007)
110. Yin, L., Chen, X., Sun, Y., Worm, T., Reale, M.: A high-resolution 3d dynamic facial expression database. In: Automatic Face and Gesture Recognition (2008)
111. Gupta, S., Markey, M., Bovik, A.: Anthropometric 3d face recognition. International Journal of Computer Vision 90(3), 331–349 (2010)
112. Colombo, A., Cusano, C., Schettini, R.: Umb-db: A database of partially occluded 3d faces. In: Workshop on Benchmarking Facial Image Analysis Technologies, pp. 2113–2119 (2011)
113. Huynh, T., Min, R., Dugelay, J.-L.: An efficient LBP-based descriptor for facial depth images applied to gender recognition using RGB-D face data. In: Park, J.-I., Kim, J. (eds.) ACCV Workshops 2012, Part I. LNCS, vol. 7728, pp. 133–145. Springer, Heidelberg (2013)
114. Ballan, L., Taneja, A., Gall, J., Van Gool, L., Pollefeys, M.: Motion capture of hands in action using discriminative salient points. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VI. LNCS, vol. 7577, pp. 640–653. Springer, Heidelberg (2012)
115. Erol, A., Bebis, G., Nicolescu, M., Boyle, R.D., Twombly, X.: Vision-based hand pose estimation: A review. Computer Vision and Image Understanding 108(1-2), 52–73 (2007)
116. Lu, S., Metaxas, D., Samaras, D., Oliensis, J.: Using multiple cues for hand tracking and model refinement. In: IEEE Conference on Computer Vision and Pattern Recognition (2003)
117. de La Gorce, M., Fleet, D.J., Paragios, N.: Model-based 3d hand pose estimation from monocular video. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(9), 1793–1805 (2011)
118. Delamarre, Q., Faugeras, O.D.: 3d articulated models and multiview tracking with physical forces. Computer Vision and Image Understanding 81(3), 328–357 (2001)
119. Bray, M., Koller-Meier, E., Van Gool, L.: Smart particle filtering for high-dimensional tracking. Computer Vision and Image Understanding 106(1), 116–129 (2007)
120. Oikonomidis, I., Kyriazis, N., Argyros, A.: Efficient model-based 3d tracking of hand articulations using kinect. In: British Machine Vision Conference (2011)
121. Rehg, J.M., Kanade, T.: Visual tracking of high dof articulated structures: an application to human hand tracking. In: Eklundh, J.-O. (ed.) ECCV 1994. LNCS, vol. 801, pp. 35–46. Springer, Heidelberg (1994)
122. Stenger, B., Mendonca, P., Cipolla, R.: Model-based 3D tracking of an articulated hand. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 310–315 (2001)
123. MacCormick, J., Isard, M.: Partitioned sampling, articulated objects, and interface-quality hand tracking. In: Vernon, D. (ed.) ECCV 2000. LNCS, vol. 1843, pp. 3–19. Springer, Heidelberg (2000)
124. Heap, T., Hogg, D.: Towards 3d hand tracking using a deformable model. In: International Conference on Automatic Face and Gesture Recognition (1996)
125. Wu, Y., Lin, J., Huang, T.: Capturing natural hand articulation. In: IEEE International Conference on Computer Vision, pp. 426–432 (2001)

126. Sudderth, E., Mandel, M., Freeman, W., Willsky, A.: Visual Hand Tracking Using Nonparametric Belief Propagation. In: Workshop on Generative Model Based Vision, pp. 189–189 (2004)
127. Hamer, H., Schindler, K., Koller-Meier, E., Van Gool, L.: Tracking a hand manipulating an object. In: IEEE International Conference on Computer Vision, pp. 1475–1482 (2009)
128. Oikonomidis, I., Kyriazis, N., Argyros, A.A.: Markerless and efficient 26-DOF hand pose recovery. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) ACCV 2010, Part III. LNCS, vol. 6494, pp. 744–757. Springer, Heidelberg (2011)
129. Keskin, C., Kra, F., Kara, Y., Akarun, L.: Real time hand pose estimation using depth sensors. In: Fossati, A., Gall, J., Grabner, H., Ren, X., Konolige, K. (eds.) Consumer Depth Cameras for Computer Vision. Advances in Computer Vision and Pattern Recognition, pp. 119–137. Springer, London (2013)
130. Oikonomidis, I., Kyriazis, N., Argyros, A.A.: Tracking the articulated motion of two strongly interacting hands. In: IEEE Conference on Computer Vision and Pattern Recognition (2012)
131. State, A., Coleca, F., Barth, E., Martinetz, T.: Hand tracking with an extended self-organizing map. In: Estevez, P.A., Principe, J.C., Zegers, P. (eds.) Advances in Self-Organizing Maps. AISC, vol. 198, pp. 115–124. Springer, Heidelberg (2013)
132. Rosales, R., Athitsos, V., Sigal, L., Sclaroff, S.: 3d hand pose reconstruction using specialized mappings. In: IEEE International Conference on Computer Vision, pp. 378–387 (2001)
133. Athitsos, V., Sclaroff, S.: Estimating 3d hand pose from a cluttered image. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 432–439 (2003)
134. de Campos, T., Murray, D.: Regression-based hand pose estimation from multiple cameras. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 782–789 (2006)
135. Stenger, B., Thayananthan, A., Torr, P.: Model-based hand tracking using a hierarchical bayesian filter. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(9), 1372–1384 (2006)
136. Romero, J., Kjellström, H., Kragic, D.: Hands in action: Real-time 3d reconstruction of hands in interaction with objects. In: International Conferences on Robotics and Automation, pp. 458–463 (2010)
137. Lee, C.S., Chun, S.Y., Park, S.W.: Articulated hand configuration and rotation estimation using extended torus manifold embedding. In: International Conference on Pattern Recognition, pp. 441–444 (2012)
138. Hamer, H., Gall, J., Urtasun, R., Van Gool, L.: Data-driven animation of hand-object interactions. In: International Conference on Automatic Face and Gesture Recognition, pp. 360–367 (2011)
139. Hamer, H., Gall, J., Weise, T., Van Gool, L.: An object-dependent hand pose prior from sparse training data. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 671–678 (2010)
140. Uebersax, D., Gall, J., den Bergh, M.V., Van Gool, L.: Real-time sign language letter and word recognition from depth data. In: IEEE Workshop on Human Computer Interaction: Real-Time Vision Aspects of Natural User Interfaces (2011)
141. Ye, Y., Liu, C.K.: Synthesis of detailed hand manipulations using contact sampling. ACM Transactions on Graphics 31(4), 41 (2012)
142. Oikonomidis, I., Kyriazis, N., Argyros, A.: Full dof tracking of a hand interacting with an object by modeling occlusions and physical constraints. In: IEEE International Conference on Computer Vision (2011)

143. Kim, D., Hilliges, O., Izadi, S., Butler, A.D., Chen, J., Oikonomidis, I., Olivier, P.: Digits: Freehand 3d interactions anywhere using a wrist-worn gloveless sensor. In: ACM Symposium on User Interface Software and Technology, pp. 167–176 (2012)
144. Zhao, W., Chai, J., Xu, Y.Q.: Combining marker-based mocap and rgb-d camera for acquiring high-fidelity hand motion data. In: Symposium on Computer Animation, pp. 33–42 (2012)
145. Starner, T., Weaver, J., Pentland, A.: Real-time american sign language recognition using desk and wearable computer based video. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(12), 1371–1375 (1998)
146. Derpanis, K.G., Wildes, R.P., Tsotsos, J.K.: Hand gesture recognition within a linguistics-based framework. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3021, pp. 282–296. Springer, Heidelberg (2004)
147. Ong, S., Ranganath, S.: Automatic sign language analysis: A survey and the future beyond lexical meaning. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(6), 873–891 (2005)
148. Pei, T., Starner, T., Hamilton, H., Essa, I., Rehg, J.: Learnung the basic units in american sign language using discriminative segmental feature selection. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4757–4760 (2009)
149. Yang, H.D., Sclaroff, S., Lee, S.W.: Sign language spotting with a threshold model based on conditional random fields. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(7), 1264–1277 (2009)
150. Theodorakis, S., Pitsikalis, V., Maragos, P.: Model-level data-driven sub-units for signs in videos of continuous sign language. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2262–2265 (2010)
151. Zafrulla, Z., Brashear, H., Hamilton, H., Starner, T.: A novel approach to american sign language (asl) phrase verification using reversed signing. In: IEEE Workshop on CVPR for Human Communicative Behavior Analysis, pp. 48–55 (2010)
152. Dreuw, P., Ney, H., Martinez, G., Crasborn, O., Piater, J., Moya, J.M., Wheatley, M.: The signspeak project - bridging the gap between signers and speakers. In: International Conference on Language Resources and Evaluation (2010)
153. Liu, X., Fujimura, K.: Hand gesture recognition using depth data. In: International Conference on Automatic Face and Gesture Recognition (2004)
154. Mo, Z., Neumann, U.: Real-time hand pose recognition using low-resolution depth images. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1499–1505 (2006)
155. Breuer, P., Eckes, C., Müller, S.: Hand gesture recognition with a novel IR time-of-flight range camera–A pilot study. In: Gagalowicz, A., Philips, W. (eds.) MIRAGE 2007. LNCS, vol. 4418, pp. 247–260. Springer, Heidelberg (2007)
156. Soutschek, S., Penne, J., Hornegger, J., Kornhuber, J.: 3-d gesture-based scene navigation in medical imaging applications using time-of-flight cameras. In: Workshop on Time of Flight Camera based Computer Vision (2008)
157. Kollorz, E., Penne, J., Hornegger, J., Barke, A.: Gesture recognition with a time-of-flight camera. International Journal of Intelligent Systems Technologies and Applications 5, 334–343 (2008)
158. Penne, J., Soutschek, S., Fedorowicz, L., Hornegger, J.: Robust real-time 3d time-of-flight based gesture navigation. In: International Conference on Automatic Face and Gesture Recognition (2008)
159. Li, Z., Jarvis, R.: Real time hand gesture recognition using a range camera. In: Australasian Conference on Robotics and Automation (2009)

160. Takimoto, H., Yoshimori, S., Mitsukura, Y., Fukumi, M.: Classification of hand postures based on 3d vision model for human-robot interaction. In: International Symposium on Robot and Human Interactive Communication, pp. 292–297 (2010)
161. Lahamy, H., Litchi, D.: Real-time hand gesture recognition using range cameras. In: Canadian Geomatics Conference (2010)
162. Van den Bergh, M., Van Gool, L.: Combining rgb and tof cameras for real-time 3d hand gesture interaction. In: IEEE Workshop on Applications of Computer Vision (2011)
163. Marnik, J.: The polish finger alphabet hand postures recognition using elastic graph matching. In: Kurzynski, M., Puchala, E., Wozniak, M., Zolnierek, A. (eds.) Computer Recognition Systems 2. ASC, vol. 45, pp. 454–461. Springer, Heidelberg (2007)
164. Incertis, I., Garcia-Bermejo, J., Casanova, E.: Hand gesture recognition for deaf people interfacing. In: International Conference on Pattern Recognition, pp. 100–103 (2006)
165. Lockton, R., Fitzgibbon, A.W.: Real-time gesture recognition using deterministic boosting. In: British Machine Vision Conference (2002)
166. Liwicki, S., Everingham, M.: Automatic recognition of fingerspelled words in british sign language. In: IEEE Workshop on CVPR for Human Communicative Behavior Analysis (2009)
167. Kelly, D., Mc Donald, J., Markham, C.: A person independent system for recognition of hand postures used in sign language. Pattern Recognition Letters 31, 1359–1368 (2010)
168. Amin, M., Yan, H.: Sign language finger alphabet recognition from gabor-pca representation of hand gestures. In: Machine Learning and Cybernetics (2007)
169. Munib, Q., Habeeb, M., Takruri, B., Al-Malik, H.: American sign language (asl) recognition based on hough transform and neural networks. Expert Systems with Applications 32(1), 24–37 (2007)
170. Tzionas, D., Gall, J.: A comparison of directional distances for hand pose estimation. In: Weickert, J., Hein, M., Schiele, B. (eds.) GCPR 2013. LNCS, vol. 8142, pp. 131–141. Springer, Heidelberg (2013)