Pattern Recognition Letters 48 (2014) 70–80

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Review Article

Human activity recognition from 3D data: A review ☆

J.K. Aggarwal, Lu Xia *

The University of Texas at Austin, Austin, TX 78705, USA

ARTICLE INFO

Article history:
Available online 4 May 2014

Keywords:
Computer vision
Human activity recognition
3D data
Depth image

ABSTRACT

Human activity recognition has been an important area of computer vision research since the 1980s. Various approaches have been proposed with a great portion of them addressing this issue via conventional cameras. The past decade has witnessed a rapid development of 3D data acquisition techniques. This paper summarizes the major techniques in human activity recognition from 3D data with a focus on techniques that use depth data. Broad categories of algorithms are identified based upon the use of different features. The pros and cons of the algorithms in each category are analyzed and the possible direction of future research is indicated.

1. Introduction

This paper is a tribute to the work of Dr. **Maria Petrou** on 3D data. She pioneered many techniques and methodologies dealing with 3D data which have benefited researchers as well as industry. She developed techniques for generating a 3D map from photometric stereo [3]; she proposed two methodologies of characterizing 3D textures: one based on gradient vectors and one on generalized co-occurrence matrices [45]; she built algorithms to reconstruct 3D horizons from 3D seismic datasets [8] and to inference the shape of a block of granite from cameras placed at 90 degrees to each other [30]. Her passing on October 15, 2012, just as the new advanced 3D sensors were becoming available, was a great loss. As far as we know, she did not present work on activity recognition from 3D data. Even though, she made important contributions to the computer vision community and inspired the later works on 3D which includes the application on activity recognition [17]. In this paper, we are going to give a review of the recent works on activity recognition using 3D data.

Recognizing human activity is one of the important areas of computer vision research today. The goal of human activity recognition is to automatically detect and analyze human activities from the information acquired from sensors, e.g. a sequence of images, either captured by RGB cameras, range sensors, or other sensing modalities. Its applications include surveillance, video analysis, robotics and a variety of systems that involve interactions between persons and electronic devices. The development of human activity recognition from depth sensors began in the early 1980s. Past research has mainly focused on learning and recognizing activities from video sequences taken by visible-light cameras. Those works were summarized at different depths from different perspectives in several survey papers [1,83]. The major issue with visible-light videos is that capturing articulated human motion from monocular video sensors results in a considerable loss of information. This limits the performance of video-based human activity recognition. Despite the efforts of the past decades, recognizing human activities from videos is still a challenging task.

After the recent release of cost-effective depth sensors, we see another growth of research on 3D data. We divide the methods to obtain 3D data from the past 20 years into three categories. One way is by using marker-based motion capture systems such as MoCap.[1] The second way is from stereo: capture 2D image sequences from multiple views to reconstruct 3D information [3]. The third way is to use range sensors. The development of range cameras has progressed rapidly over the past few years. Recently, the advent of depth cameras at relatively inexpensive costs and smaller sizes gives us easy access to the 3D data at a higher frame rate resolution, leading to the emergence of many new works on action recognition from 3D data. We will discuss the state-of-the-art algorithms on human activity recognition using 3D data in each category in this article with a focus on recent developments on depth data.

Depending on the environments, human activity may have different forms ranging from simple actions to complex activities. They can be conceptually categorized into four categories [1]: atomic actions, activities that contain a sequence of different actions, interactions that include person–object interactions and

☆ This paper has been recommended for acceptance by Edwin Hancock.

* Corresponding author. Fax: +1 512 471 5532.
E-mail address: xialu@utexas.com (L. Xia).

[1] http://mocap.cs.cmu.edu/.

http://dx.doi.org/10.1016/j.patrec.2014.04.011
0167-8655/© 2014 Elsevier B.V. All rights reserved.

person-to-person interactions, and group activities. Research on atomic action recognition from 3D data has developed for many years, while complex activities and interactions have been studied recently after the easy access of 3D data became available. The research on group activities using 3D data is limited, either due to the difficulty of obtaining the data or limitation of the sensors.

Here we enumerate four major challenges to vision based human action recognition. The first is low level challenges [93,13]. Occlusions, cluttered background, shadows, and varying illumination conditions can produce difficulties for motion segmentation and alter the way actions are perceived. This is a major difficulty of activity recognition from RGB videos. The introduction of 3D data largely alleviates the low-level difficulties by providing the structure information of the scene. The second challenge is view changes [63,93,37,92,68]. The same actions can generate a different "appearance" from different perspectives. Solving this issue with a traditional RGB camera is achieved by introducing multiple synchronized cameras, which is not an easy task for some applications. But this is not a serious problem for recognition algorithms using a 3D motion capture system. For recognition from range images, this problem is partially alleviated because the appearance from a slightly rotated view can be inferred from the depth data. Even though, the problem is not totally solved because the range image only provides information on one side of the object in view, nothing is known about the other side. If skeletal joint information can be inferred accurately using a single depth camera, a view-invariant recognition algorithm may be constructed from the skeletal joint information. The third challenge is scale variance, which can result from a subject appearing at different distances to the camera or different subjects of different body sizes. In RGB videos, this can be solved by extracting features at multiple scales [14]. In depth videos, this can be easily adjusted because the true 3D dimension of the subject is known [78,97]. The fourth challenge is intra-class variability and inter-class similarity of actions [64]. Individuals can perform an action in different directions with different characteristics of body part movements, and two actions may be only distinguished by very subtle spatio-temporal details. This still remains a hard problem for algorithms using various types of data.

The objective of this article is to provide an overview of the state-of-the-art methodologies on human-activity recognition using 3D data. We discuss various techniques to acquire 3D spatio-temporal data, and summarize recognition methodologies using each type of data source. In particular, we will talk about 3D from stereo, 3D from motion capture system and 3D from depth sensors. Most of the recent works on human activity recognition reside in the third category, and some of the techniques used are adapted from previous works in the first two categories. We, therefore, put our focus on the third category in this survey. Although there are also works on 3D shape from shading [66], 3D from focus/defocus [41,27], 3D from texture [53], 3D from motion [16] and so on, they are not covered in this survey. These problems are usually ill-posed and the solutions are not unique [65]. Even with parameters such as the light source, the surface reflectance and the camera, it is still hard to get rid of some ambiguities. Hence, the algorithms are mostly proposed to solve the 3D shape of static objects. To the best of our knowledge, there is no work on activity recognition that uses shading, defocus texture, or motion to acquire 3D spatio-temporal information.

Recently, [12] wrote a survey summarizing the human motion analysis algorithms using depth imagery. Their paper summarizes depth sensing techniques, preprocessing of depth images, pose recognition, and action recognition from depth video, datasets, and libraries. However, the activities recognition methodologies were only coarsely categorized and introduced. This article will follow a different framework, and we will show a deeper analysis of the activity recognition algorithms. Especially, we will give an inside look at the features that were proposed in different scenarios and cover more recent publications. The activities discussed in this review range from atomic actions to complex activities, which include whole body activities, upper-body gestures, hand gestures, person–object interactions and person-to-person interactions. Static gestures, gait, and pose recognition are not in the scope of the paper.

2. 3D from stereo

Stereo, the reconstruction of three-dimensional shapes from two or more intensity images is a classic research problem in computer vision [54,67]. Early range sensors are expensive and cumbersome; the low-cost digital stereo camera has therefore generated interest in vision-based systems. A stereo camera is equipped with two or more lenses with a separate image sensor or film frame for each lens. This allows the camera to simulate human binocular vision, and therefore gives it the ability to generate three-dimensional images. By comparing the two images, the relative depth information can be obtained, in the form of disparities, which are inversely proportional to the differences in distance to the objects. The detail of stereo matching and computing the disparity/depth from the matching is not discussed in this paper, please refer to [25] for details.

Stereo vision is highly important in fields such as robotics, and it also has application in entertainment, information transfer, and automated systems. For instance, Dr. Maria Petrou developed a photometric stereo technique, which uses a fixed digital camera and three lights to illuminate the object from different angles. All the data are combined into one 3D image by analyzing the shadows and highlights. This technique has been used to find flaws in industrial surfaces and to capture a 3D map of a human face [3]. The techniques of recovering and analyzing depth images made a great contribution to the computer vision community. Kanade et al. developed a video-rate stereo machine that generated a dense depth map at the video rate [43] and demonstrate its application in merging real and virtual worlds in real time.

As mentioned above, we draw the ideas of acquisition of 3D information from Petrou, Kanade, and other researchers. Further, researchers combine the 3D with other information to analyze sequences of images. In particular, researchers further explored its application on tracking [20,59,69,105], pose recognition [18,58,75,32] and human activity recognition from stereo images or sequences [33,85]. The commonly used activity recognition schemes fall into three categories.

The first category is model based approaches which fit a known parametric 3D model, usually a kinematic model, to the stereo videos and estimate motion parameters of the 3D model. For instance, [84] co-register a 3D body model to the stereo images and extract joints locations from the body model and use joint-angle features for action recognition. The fitted 3D body model is less ambiguous than silhouettes, however, recovering the parameters and pose of the model is usually a difficult problem without the help of landmarks.

The second category is holistic approaches, which directly model actions using image formation, silhouettes, or optical flow [92]. Action is recognized by comparing the observations with learned templates of the same type. This usually requires that test and learned templates are obtained from similar configurations. [61] represent human actions as short sequences of atomic body poses, and the body poses are represented by a set of 3D silhouettes seen from multiple viewpoints. In this approach, no explicit 3D poses or body models are used, and individual body parts are not identified. [2] encode action using the Cartesian component of optical flow vectors, human body silhouettes feature vector, and the combined feature. A set of HMMs were built for each action to represent action from different views. [92] generate action exemplars as 3D visual hulls that have been computed using five calibrated cameras. The 3D exemplar is used to generate an arbitrary 2D view and compare with the observation silhouettes obtained from a single camera and background subtraction. An exemplar-based hidden Markov model (HMM) was proposed for recognition. [37] estimate 2D optical flow from each view and extend to 3D using 3D model reconstructions and pixel-to-vertex correspondences. They also use the 3D motion vector fields for action recognition.

The third category is to calculate a depth map or disparity map from the stereo images and build features using the depth map [33,68]. First, matching is found between a pair of images that identify corresponding points, then disparity d is calculated from the matches which in turn gives the depth information of the imaged scene [73,23]: $Z = \frac{fB}{d}$ where Z is the distance along the camera Z axis, which is the depth. f is the focal length of the camera in pixels, B is the baseline in meters, and d is the disparity in pixels. After Z is determined, the X, Y of the real world coordinates can be computed from $X = \frac{uZ}{f}$, $Y = \frac{vZ}{f}$ where u, v are the pixel location of the point in the 2D image. The X, Y, Z are thus the real world position. This gives the same type of input as range sensors, so we will discuss them together in Section 4.

Because of the complexity of the geometry, reconstruction of 3D information from stereo images still remains a challenging task. Reflections, transparencies, depth discontinuities, and lack of texture in images confound the matching process and result in ambiguities in depth assignments caused by the presence of multiple good matches. Meanwhile, the reconstruction relies on the intensity images which means that it is sensitive to lighting changes and background clutters. The need for multiple calibrated and synchronized cameras followed by an exhaustive training phase is obviously not desirable. These issues limited the application of stereo vision.

3. 3D from motion capture

Another technique for getting 3D information of humans is directly using a motion capture system (Mocap). It is an important technique for capturing and analyzing human articulations. Mocap has been widely used to animate computer graphic figures in motion pictures and video games. It is also used for analyzing and perfecting the sequencing mechanics of premiere athletes, as well as monitoring the recovery progress of physical therapies.

There are many ways to generate motion capture data, e.g. using optical sensing of markers placed in specific positions usually at, or near, boney regions, and using triangulation from multiple cameras to estimate the 3D position of each marker. Motion capture data can also be obtained using a marker-less motion tracker, RFID tag readings, and magnetic sensor readings. For an overview of the techniques to generate motion capture data, please refer to the online document[2].

Several motion capture datasets are available that provide large collections of actions, such as the CMU Motion Capture Database[3], the MPI HDM05 Motion Capture Database2[4], The CMU Kitchen Data Set[5], the LACE Indoor Activity Benchmark Data Set [6], and the TUM Kitchen dataset[7].

In a motion capture system, only the 3D locations of the selected points are recorded, so the action recognition algorithms using this type of data commonly build features based on joint positions or joint angles. Parameters are calculated from a single joint or a combination of multiple joints [29,55,46]. [22] partition the human model into 5 parts and compute disjoint submotion matrices from the joint angles to describe the submotions of each parts. [77] group the joints into sets of point triplets that form planes, and use fundamental ratios to identify plane motions.

The skeletal joint information that has been explored here is in fact similar to what is acquired later on from the modern range sensors. Thus we will discuss the types of skeletal joints features in more detail in Section 4.2.

Generally speaking, the skeletal joints information collected by the Mocap system is more reliable and relatively less noisy than the ones estimated from the depth sensor. Algorithms working with the range sensor need to pay closer attention to handle the noise. On the other hand, collecting data with a single depth sensor is much easier than using the Mocap system. Thus more varied datasets are becoming available with the depth sensors, which expands the realm of the problems that the computer vision algorithms are going to solve.

4. 3D from range sensors

The acquisition of 3D data from one camera has certainly been a desirable goal for a long time. Devices that capture 3D surface characteristics have been around for over thirty years. These devices, known as range scanners, output a two-dimensional array of distances corresponding to each point in the imaged scene. Earlier range sensors were either too expensive, difficult to use in human environments, slow at acquiring data, or provided a poor estimation of distance. Computer vision research on range data was focused on representation and recognition of static objects [99,79,4], body parts [35], or static gestures [56]. There is also research on motion estimation from a pair of range images [71,72]. Before the advent of easy to use range sensors, there was no publication on human activity recognition from range images to the best of our knowledge because the acquisition of real-time range images is a difficult task.

Recent advances in sensing technology have enabled us to capture the depth information in real-time, which inspired the research on activity recognition from 3D data. The widely used sensors at acquiring depth images include time-of-flight (ToF) cameras and structured light cameras (e.g. Microsoft Kinect). ToF cameras compute depth by measuring the time-of-flight of a light signal between the camera and the subject for each point of the image. Structured light cameras calculate the depth by projecting a structured light onto the scene and comparing the reflected pattern with the stored pattern. For details on the two types of sensors and comparisons, please refer to [12].

In a depth image, the value of each pixel corresponds to the distance between the real world point and the sensor, which provides the 3D structural information of the scene. It is sometimes called a 2.5D image because only the 3D structure of the point visible to the sensor is contained in the image, nothing is known about the other side of the object or scene. In spite of this, depth still provides important complementary information to visual color images, it is more robust to illumination changes and can even work in total darkness.

Various methods have been proposed in the last few years addressing the problem of human activity recognition from depth images. Based on the features they use, we divide them into five categories: features from 3D silhouettes, features from skeletal joint or body part locations, local spatio-temporal features, local

[2] http://en.wikipedia.org/wiki/Motion_capture/.
[3] http://mocap.cs.cmu.edu/.
[4] http://www.mpi-inf.mpg.de/resources/HDM05/.
[5] http://kitchen.cs.cmu.edu/.
[6] http://www.cs.rochester.edu/%18spark/muri/.
[7] http://ias.in.tum.de/software/kitchen-activity-data.

occupancy patterns, and 3D scene flow features. We will give an overview of the algorithms proposed using the features in each category, and analyze the advantages and limitations of each type of features. Fig. 1 shows the taxonomy of the features and the list of publications that we are going to cover.

4.1. Recognition from 3D silhouettes

Among the early attempts on action recognition from intensity images, researchers have successfully extracted 2D silhouettes as a simple representation of human body shape from the intensity or RGB images and model the evolution of silhouettes in the temporal domain to recognize actions. It was shown that the silhouettes, or, extremities of the silhouettes, carry a great deal of shape information of the body. By tracking the person's silhouette over time, [21] generated a Motion History Image (MHI) which is a scalar-valued image where intensity is a function of recency of motion. [28] extracted a "star" skeleton from silhouettes for motion analysis. [104] extracted extremities from 2D silhouettes as semantic posture representation in their application for the detection of fence climbing. However, the silhouettes extracted from intensity images are view-dependent, and only suitable for describing actions parallel to the camera. Also, extracting the correct silhouettes of the actor can be difficult when there is background clutter or bad lighting conditions.

In a depth image, the silhouette of a person can usually be extracted more easily and accurately. In addition, the depth image provides the body shape information not only along the silhouettes, but also the whole side facing the camera. Thus, more information can be acquired from depth images. Many algorithms have been proposed to recognize actions using representations built from 3D silhouettes. [52] sample a bag of 3D points on the contours of the planar projections of the 3D depth map to characterize a set of salient postures that correspond to the nodes in the action graph (see Fig. 2). The number of points can be controlled by the number of projection planes used. [101] also project depth maps onto three orthogonal planes. They propose Depth Motion Maps (DMM) which stack the motion energy through the entire video sequences on each plane. HOG is employed to describe the DMM. [60] propose a Three-Dimensional Motion History Image (3D-MHI) which equip the original MHI with two additional channels, i.e. two depth change induced motion history images (DMHIs): forward-DMHI and backward-DHMI which encode forward and backward motion history. [40] use Radon transform ($\mathcal{R}$ transform) to compute a 2D projection of depth silhouettes along specified view directions, and employ $\mathcal{R}$ transform to transform the 2D Radon projection into a 1D profile for every frame. [26] propose a Global Histogram of Oriented Gradient (GHOG) by extending the classic HOG [19] which was designed for pedestrian detection from RGB images. The GHOG describes the appearance of the whole silhouettes without splitting the image into cells. The gradient of the depth stream shows the highest response on the contours of the person thus indicating the posture of the person. [96] propose extended-MHIs by fusing MHI with gait energy information (GEI) and inversed recording (INV) at an early stage. GEI compensates for non-moving regions and multiple-motion-instance regions. INV provides complementary information by assigning a larger value at initial motion frames instead of the last motion frames. The extended-MHI was proved to outperform the original MHI on an action recognition scenario. [47] divide the depth image into sectors and compute the average distance from the hand silhouettes in each sector to the center of the normalized hand mesh as a feature vector to recognize hand gestures.

Current algorithms using 3D silhouettes are suitable for single person action recognition and perform best on simple atomic actions. There is difficulty in recognizing complex activities due to the information lost either when computing 2D projection of 3D data or when stacking the information along temporal domains.

Features for activity recognition using depth images

	3D silhouettes	Skeletal joints/body parts	Local spatio-temporal interest features	Local occupancy pattern	3D optical flow
Gaming Action	Li et al.(2010) Yang et al. (2012)	Wang et al. (CVPR2012) Bloom et al. (2012) Yang and Tian (2012) Xia et al.(2012)	Xia and Aggarwal (2013)	Wang et al. (ECCV 2012) Vieira et al. (2012)	
Daily Action		Sempena et al. (2011) Masood et al. (2011) Xia et al. (2012)			Ballin et al. (2013)
Daily activity	Jalal et al.(2011) Ni et al. (2010)	Wang et al. (CVPR2012) Zhang and Tian (2012) Sung et al. (2011)	Zhang and Parker (2011) Zhao et al. (2012) Cheng et al. (2012) Ni et al. (2011) Xia and Aggarwal (2013)	Wang et al. (CVPR 2012)	
Upper arm gestures	Wu et al.(2012) Fanello et al. (2013)		Hernandez-Vela et al. (2012)		Holte et al. (2010) Fanello et al. (2013)
Hand gestures	Kurakin et al. (2012)			Kurakin et al. (2012) Wang et al. (ECCV2012)	
Human interaction		Yun et al. (2012)			

Fig. 1. Taxonomy for feature on human activity recognition from depth data and list of corresponding publications in each category.

Occlusion and noise can mar the silhouettes dramatically, and the extraction of accurate silhouettes may be difficult when the person interacts with background objects (e.g. sitting on sofa). Special algorithms have to be devised to model person-to-person interactions. Furthermore, the depth map in fact only gives the 3D silhouettes of the person facing the camera. So the 3D silhouettes based algorithm are usually view-dependent, even though they are not limited to only modeling parallel motions as in intensity image.

4.2. Recognition from skeletal joints or body parts tracking

The human body is an articulated system of rigid segments connected by joints and human action is considered as a continuous evolution of the spatial configuration of these segments. Back in 1975, Johansson's experiments show that humans can recognize activity with only seeing the light spots attached to the person's major joints [42]. In computer vision, there is plenty of research on extracting the joints or detecting body parts and tracking them in the temporal domain for activity recognition. In intensity images, researchers tried to extract "skeletons" from silhouettes [28] or label main body parts such as arms, legs, torso, and head for activity recognition [7]. Researchers also tried to extract joints or body parts from stereo images or to directly get them from motion capture systems.

In 2011, [78] proposed to extract 3D body joint locations from a single depth image using an object recognition scheme. The human body is labeled as body parts based on the per-pixel classification results. The parts include LU/ RU/ LW/ RW head, neck, L/R shoulder, LU/ RU/ LW/ RW arm, L/ R elbow, L/ R wrist, L/ R hand, LU/ RU/ LW/ RW torso, LU/ RU/ LW/ RW leg, L/ R knee, L/ R ankle, and L/ R foot (Left, Right, Upper, Lower). Skeletal joints tracking algorithms were built into the Kinect device, which offers easy access to the skeletal joint locations (see Fig. 3). This excited considerable interest in the computer vision community. Many algorithms have been proposed recognizing activities using skeletal joint information. The most straight forward feature is the pairwise joint location difference feature, which is a compact representation of the structure of the skeleton posture of a single frame. By computing the difference of the joint positions from the current frame and previous frame, one can get the joint motion between the two frames. Especially, supposing the first frame is a neutral pose, taking the joint position difference between the current frame and first frame can generate an offset feature. [57,107,100] concatenates these features and test its effectiveness at recognizing activities. Besides direct concatenation, [98] cast the joint positions into 3D cone bins and build a histogram of 3D joints locations (HOJ3D) as representation of postures.

Furthermore, the position of certain joints may be related to certain types of activities. [107] use head-floor distance and joint angles between every pair of joints to detect abnormal actions related to a falling event. [50] use depth to extract hands and objects from a tabletop and track hands using color. Then they combine a global feature using PCA on 3D hand trajectories, and a local feature using bag-of-words of trajectory gradients to recognize kitchen activities from an overhead Kinect-style camera.

From the skeletal joint locations, joint orientation can be computed. It is also a good feature as it is invariant to human body size. [76] built a feature vector from joint orientation along time series and apply dynamic time warping onto the feature vector for action recognition. [9] concatenates five types of features: pairwise joint position difference, joint velocity, velocity magnitude, joint angle velocity w.r.t. the x–y plane and x–z plane, and a 3D joint angle between three distinct joints. In total, 170 features were computed to recognize gaming actions.

There are also researchers that group the joints and construct planes from joints and measure joint-to-plane distance and motion as features. [106] constructed a feature that captures the geometric relationship between a joint and a plane spanned by three joints. This feature is intended to describe information such as how far the right foot lies in front of the plane spanned by the left knee, hip, and torso. [80] computed each joint's rotation matrix with respect to the person's torso, hand position w.r.t. the torso and joint rotation motion as features and use maximum-entropy Markov model (MEMM) to learn the actions. We collect a number of related features in Table 1 [90,102,106].

Compared to the features from 3D silhouettes, the skeletal joint features are invariant to the camera location and subject appearance. In a good setting, the skeletal tracking algorithms can accurately extract the joint positions from either front view, side view, or back view making the skeletal joint feature based algorithm view-invariant. Furthermore, joint orientation feature is invariant to human body size. If the skeletal joint locations of multiple persons are known, the features can be extended to model person-to-person interaction [106]. The recognition scheme based on the skeletal joints feature is better at modeling finer activities compared to the 3D silhouettes based algorithms. The limitation of the skeletal feature is that is does not give information about the surrounding objects. When modeling person-to-object interaction, object detection and tracking has to be combined.

In addition, the current algorithm to estimate skeleton joint position from depth image is not perfect. The skeleton tracking from the Kinect works well when the human subject is in an upright position facing the camera and there are no occlusions. However, the result is not very reliable when the human body is partly in view, the person touches the background, or the person is not in an upright position (e.g. patient lying on a bed). In surveillance or senior home monitoring scenarios, the camera is mounted in a higher location and the subjects are not facing the camera, this may also create difficulty for the skeletal estimation algorithm.

4.3. Recognition using local spatio-temporal features

The local spatio-temporal feature has been a popular description for action recognition from intensity video. The video is regarded as a 3D volume along space (x, y) and temporal t axis. Generally, local spatio-temporal interest points (STIPs) are first detected, then descriptors are built around the STIPs from the volume. After that, classification can be made from the descriptors (e.g. bag-of-words approach). Many different STIP detectors [24,48,62,94] and descriptors [24,44,49,74,94] have been proposed in the literature during the past decade. The local spatio-temporal features have been demonstrated successful at recognizing a number of action classes with varying difficulties [88].

Table 1
Skeletal joint features in literature. i and j are any joints of the persons x and y, $t \in T$. For single person activity recognition, $x = y$. $< p^y_{j,t}, p^y_{k,t}, p^y_{l,t} >$ indicates the plane spanned by p^y_j, p^y_k, p^y_l. $< p^y_{j,t}, p^y_{k,t}, p^y_{l,t} >_n$ indicates the normal vector of the plane spanned by p^y_j, p^y_k, p^y_l.

Features	Equation
Joint location	$F^l = p_{i,t}$
Joint distance	$F^{jd}(i,j;t) = \lVert p^x_{i,t} - p^y_{j,t} \rVert$
Joint orientation	rotation quaternion of each joint w.r.t. torso
Joint motion	$F^{jm}(i,j;t_1,t_2) = \lVert p^x_{i,t1} - p^y_{j,t2} \rVert$
Plane feature	$F^{pl}(i,j,k,l;t) = dist(p^x_{i,t}, < p^y_{j,t}, p^y_{k,t}, p^y_{l,t} >)$
Normal plane	$F^{pl}(i,j,k,l;t) = dist(p^x_{i,t}, < p^y_{j,t}, p^y_{k,t}, p^y_{l,t} >_n)$
Joint velocity	$F^{jv}(i,j,k;t) = \frac{v^x_{i,t} \cdot (p^y_{j,t} - p^y_{k,t})}{\lVert p^x_{1,t} - p^y_{1,t} \rVert}$
Normal velocity	$F^{nv}(i,j,k,l;t) = v^x_{i,t} \cdot \hat{n} < p^y_{j,t}, p^y_{k,t}, p^y_{l,t} >$

Table 2
Activity recognition algorithms using skeletal joint features.

Algorithms	Joint location	Joint orientation	Joint motion	Planes
[76]		✓		
[106]	✓		✓	✓
[9]	✓	✓	✓	
[107]	✓	✓	✓	
[57]	✓		✓	
[80]	✓	✓	✓	✓
[98]	✓	✓		
[100]	✓		✓	

Table 3
Activity recognition using spatio-temporal volume feature.

Algorithms	channel	STIP detector	STIP descriptor
	Depth	only for layer partition	
[60]	RGB	Harris3D	HOGHOF
	Depth	response function	x, y, t gradient
[108]	RGB		x, y, t gradient
	Depth	Harris3D	VFHCRH
[34]	RGB	–	–
	Depth	Harris3D	CCD
[15]	RGB	–	–
	Depth	Harris3D	HOGHOF + LOP
[109]	RGB	Harris3D	HOGHOF
[97]	Depth	filtering scheme with noise suppression	DCSF
	RGB	–	–

Encouraged by the success of intensity video, spatio-temporal interest points and features are also explored on depth data. Among the earliest attempts, [60] used depth information to partition the space into layers, extracted STIPs from RGB channels of each layer using a Harris3D detector [48], and used HOG/HOF [49] to describe the neighborhood of STIPs in the RGB channel. In this approach, depth was only served as an auxiliary for the extraction of STIPs from RGB videos, the detector and descriptor were applied on the RGB channels. [15,34,109] all used a Harris3D detector [48] to extract STIPs from depth videos. Differently, [34] proposed a 2D descriptor VFHCRH by extending [70] to describe the 2D depth image patches. Note that VFHCRH is not a spatio-temporal feature since there is no temporal information encoded in the feature. [15] propose a Comparative Coding Descriptor (CCD) to describe the $3 \times 3 \times 3$ depth cuboid by comparing the depth value of the center point with nearby 26 points. [109] combined HOG/HOF [49] and the proposed local depth pattern (LDP) feature for representation. The LDP feature is defined by the difference of average depth values between nearby cells of the 3D cuboid. They also showed that IPs extracted from RGB channels perform better than IPs extracted from a depth channel using Harris3D for about 2.5% on RGBD-HuDaAct dataset [60]. This is understandable since the Harris3D was designed for intensity videos, but the depth videos are usually noisy and have many missing values. To deal with this, [97] proposed a filtering scheme to find the STIPs from depth videos with noise suppression functions, and it has been proven to give better STIPs in depth videos on the MSRDailyActivity3D dataset [90] compared to Harris3D [48] and Cuboid detector [24]. They also proposed a self-similarity depth cuboid feature (DCSF) as the descriptor for a spatio-temporal depth cuboid which handles the noisy measurements and missing values (see Fig. 4).

The above work all treat the depth channel and RGB channels separately when extracting spatio-temporal interest points. In contrast, [108] extracted interest points from a combined response given by intensity and depth channels. The response is defined by a linear combination of the filtered result from intensity and depth channels. Furthermore, they compute the gradients along x, y, t directions as the local feature for both depth and intensity channels. This feature is tested on a self-collected dataset containing six human activities and achieves good results.

Local spatio-temporal features capture the shape and motion characteristics in video and provide independent representation of events. It is invariant to spatio-temporal shifts and scales, and it naturally deals with occlusions, multiple motions, person-to-person interaction, and person–object interaction. Because such features are usually directly extracted without the need for motion segmentation and tracking, it makes the algorithm more robust and has a wider range of applications. The limitation falls in the following aspects. First, the feature is view-dependent, since the cuboid is extracted directly from the x,y,t volume. Secondly, the current method requires the whole video as input, and the feature computation algorithm is not very fast, thus limiting its real-time application. In addition, the current spatio-temporal interest points extraction and feature description techniques regard the depth channel and RGB channels independently. This may not be the theoretically optimal way since the RGB and D reside together in the 3D world; it might be interesting to develop algorithms with a better fusion of the two types of data.

4.4. Recognition using local 3D occupancy features

Instead of representing the depth video as 3D spatio-temporal volume, the points may be projected to the 4D (x,y,z,t) space. In this 4D space, some location will be occupied by the data points from the video, i.e. the points that the sensor captured from the real world, those locations will have a value of 1, others 0. In general, the local occupancy pattern is quite sparse, that is, the majority of its elements are zero. The local occupancy pattern has been proposed individually by several researchers for activity recognition. In fact, the local occupancy feature can be defined in the (x, y, z) space or (x, y, z, t), the former one describes the local depth appearance at a certain time instant while the latter describes the local atomic events within a certain time range. [90] design a 3D Local Occupancy Patterns (LOP) feature to describe the local "depth appearance" at joint locations to capture the information for person–object interactions. The intuition is, when the person fetches a cup, the space around the hand is "occupied" by the cup. The x,y,z space around the joint is partitioned into a $N_x \times N_y \times N_z$ spatial grid, the number of points that fall into each bin are counted and normalized to obtain the occupancy feature of that bin. This work is an example of combining skeleton joints features and local occupancy features to recognize activities and also to model person–object interactions. [89] defined the random occupancy patterns in the (x, y, z, t) domains. Similarly, the occupancy feature is the sum of the pixels in a sub-volume of the 4D space normalized by a sigmoid function. A weighted sampling approach was proposed to sample sub-volumes from the 4D space, and occupancy patterns were extracted from those locations to give an overall description of the depth video. Instead of random sampling, [87] divided the whole space–time volume into 4D grids, and extracted occupancy patterns from every partition (see Fig. 5). Interestingly, a saturation scheme was proposed to enhance the role of the sparse cells, which typically lie on the silhouettes or moving parts of the body. To deal with the sparsity of the feature, a modified-PCA called Orthogonal Class Learning (OCL) is employed to cut the length of the feature to 1/10 of its original.

The local occupancy feature defined in the (x, y, z, t) space is similar to local spatio-temporal features in that they both describe local "appearance" in the space–time domains. Local spatio-temporal features treat the z dimension as "pixel values" in the (x, y, t) volume while local occupancy patterns project the data onto a (x, y, z, t) 4D space containing 0–1 values. They may both be extracted from selected locations or random sampling. However, the local occupancy features can be very sparse while the spatio-temporal feature is not. Furthermore, spatio-temporal features contain information on the background since the cuboid is extracted from the (x, y, t) space, while local occupancy features only contain information around a specific point at a (x, y, z, t) space. This characteristic is not positive or negative as the background is helpful in certain scenarios while disturbing in some other cases.

4.5. Recognition from 3D optical flow

Optical flow is the distribution of apparent velocities of movement of brightness patterns in an image, which arises both from the relative objects' and the viewer's motion [31]. It is widely used in intensity images for motion detection, object segmentation and stereo disparity measurement [6]. Also, it is a popular feature in activity recognition from videos [11,103]. When multiple cameras are available, the integration over different viewpoints allows a 3D motion field, the scene flow [86]. However, intensity variations alone are not sufficient to estimate motion and additional constraints such as smoothness must be introduced in most scenarios. Some of the promising works on estimating 3D scene flow from stereoscopic include [91,10,39]. These algorithms usually have a high computational cost due to the fact that they estimate both the 3D motion field and disparity changes at the same time. Depth cameras advantageously provide useful geometric information from which additional consistent 3D smoothness constraints can be derived. With a stream of depth and color images coming from calibrated and synchronized cameras, we have a simpler way of getting optical flow in (x, y, z) space. Among the most straight forward and fastest methods, [82,26] compute 3D scene flow (see Fig. 6) by directly transforming the 2D optical flow vectors to 3D using the 3D correspondence information of each point, i.e., each 2D pixel x,y is projected into 3D using the depth value z and the focal length $f : X = (x - x_0)Z/f$, $Y = (y - y_0)Z/f$. (x_0, y_0) is the principal point of the sensor. Compute the 2D optical flow using traditional methods such as [38] or [54], The 3D scene flow can be obtained by differencing the two corresponding 3D vectors in two successive frames F_{t-1} and F_t using equation $\mathbf{D} = (X_t - X_{t-1}, Y_t - Y_{t-1}, Z_t - Z_{t-1})$ [26]. The 3D scene flow estimated using the above method has been proven effective at recognizing arm gestures [36] and upper/full body gestures (ChaLearn dataset) [26]. However, this method is not a very accurate estimation of the 3D scene flow since only the 2D information is considered when finding the correspondence between frames.

Recently, [51] cast the problem of estimating 3D scene flow from a calibrated depth and RGB image sequence as an optimization problem with photometric consistency constraints and motion field regularization. This method has theoretical improvement over the previous projection method since it considers both the photometric consistency and global smoothness of the 3D motion field.

[5] compute the 3D scene flow from point cloud data using 1-nearest neighbor search driven by both the 3D geometric coordinates and the RGB color information. The 3D scene flow is only computed for relevant portions of the 3D scene. They represent each tracked person by a cluster which is defined as a 4D point cloud. The 3D scene flow vector is then summarized within a 3D grid surrounding each cluster, and 3D average velocity vector is computed for each 3D cube and all these vectors are concatenated into a column vector. This feature is tested on a human action recognition task and shows reasonable performance on a new dataset containing six simple human actions.

Overall, the exploration on 3D optical flow or scene flow using RGB and depth imagery has been quite limited. Compared to the success of traditional 2D optical flow, the research on scene flow is still in its preliminary stage [110]. Currently, 3D scene flow is often computed for all the 3D points for the subject or scene, resulting in a large computational cost. Computing the 3D scene flow with real-time performance is a challenging task. We can imagine after the emergence of more effective ways to compute 3D scene flow, it has the potential to be a more popular type of feature for human action recognition and benefit more applications.

5. Kinect activity dataset

The datasets on human activity recognition came out rapidly after the release of the Kinect. Table 4 summaries the datasets that are currently publicly available. It covers gaming actions, atomic actions, daily activities, upper arm gestures, and hand gestures. Most of the datasets provide RGB and depth information, with a few also providing skeleton information. The reader is referred to the individual publications for details.

6. Conclusion and future directions

As the recent development of range sensing technology progresses, we have easy access to the 3D data as a complement to the traditional RGB imagery. Acquiring 3D data from depth sensors is more convenient than estimating it from stereo images or using motion capture systems. This is an important cornerstone in com-

Table 4
Kinect human activity recognition dataset.

Dataset	Scenario	Class	Data source	Subjects	Samples	Remark
MSRAction3D [52]	gaming	20	depth + skeleton	10	402	fronto-parallel
HuDaAct [60]	human daily activities	12	color + depth	30	1189	multiple scenes
Cornell Activity Datasets CAD-60 [81]	human daily activities	12	color + depth + skeleton	4	60	5 scenes
Cornell Activity Datasets CAD-120 [81]	human daily activities	10	color + depth + skeleton	4	120	with 10 sub-activity labels and 12 object affordance labels
Act4[2] [15]	human daily activity	14	color + Depth		6844	from 4 views
ChaLearn Gesture competition[8]	upper arm gestures	30	color + Depth	20	50,000	
MSRDailyActivity3D [90]	human daily activity	16	RGB + depth + skeleton	10	320	person–object interaction
MSRGesture3D [47]	dynamic American Sign Language (ASL) gestures	12	RGB + depth	10	336	
UTKinectAction [98]	atomic actions	10	RGB + depth + skeleton	10	200	person–object interaction
G3D [9]	gaming	7(20)	RGB + depth + skeleton	10	213	
Falling Event [107]	falling related activity	5	RGB + depth	5	150	
LIRIS [95]	human daily activities	10	grayscale + RGB + depth		828	multiple persons
UMD-Telluride[9]	kitchen actions	11	RGB + depth	2	22	person–object interaction

[8] http://gesture.chalearn.org/data/cgd2011.
[9] http://www.umiacs.umd.edu/research/POETICON/telluride_dataset/.

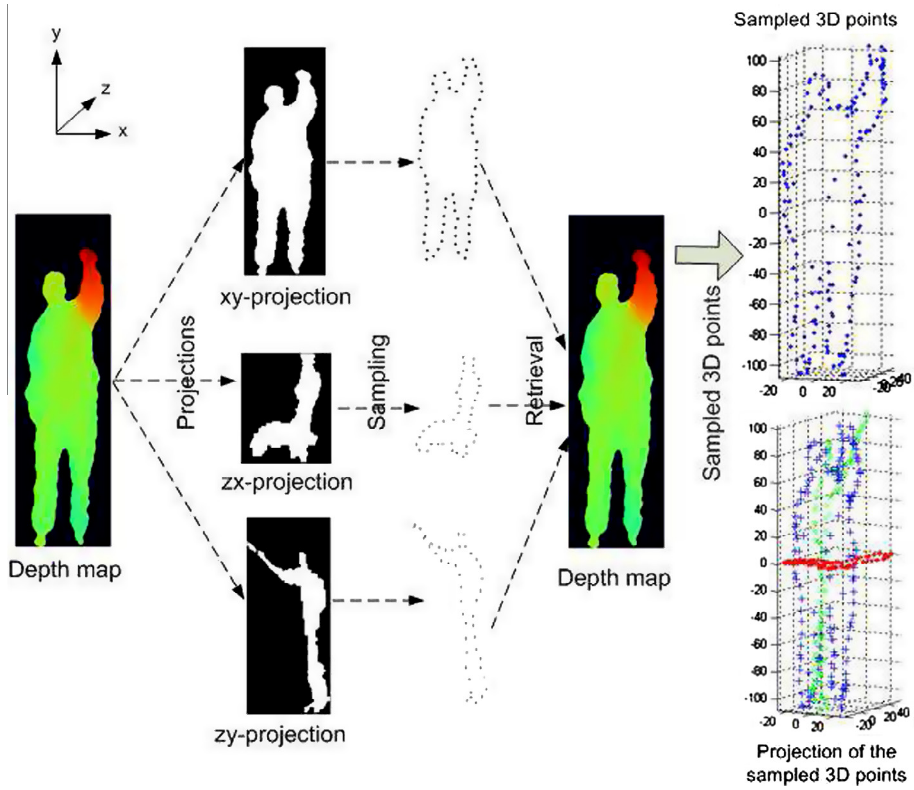

Fig. 2. Example of extracting features from 3D silhouettes [52].

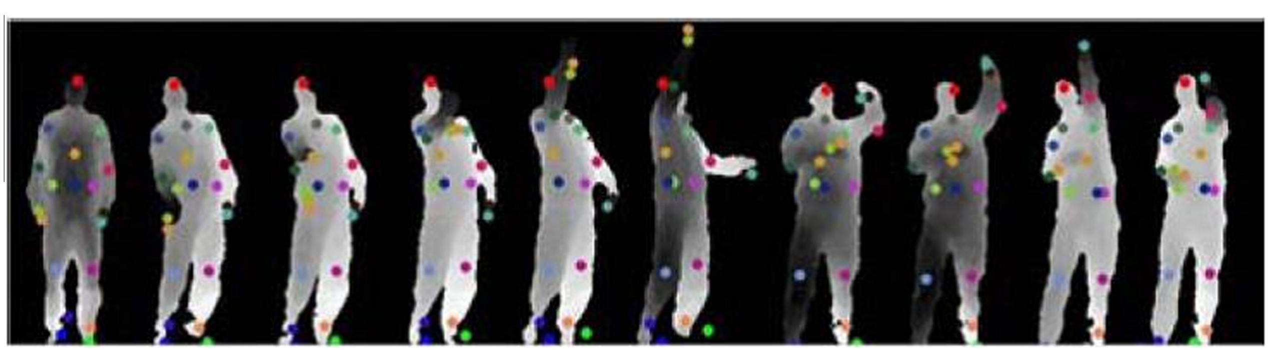

Fig. 3. Example of skeletal joints extracted from depth image sequence for the action *tennis serve* in MSRAction3D dataset [100].

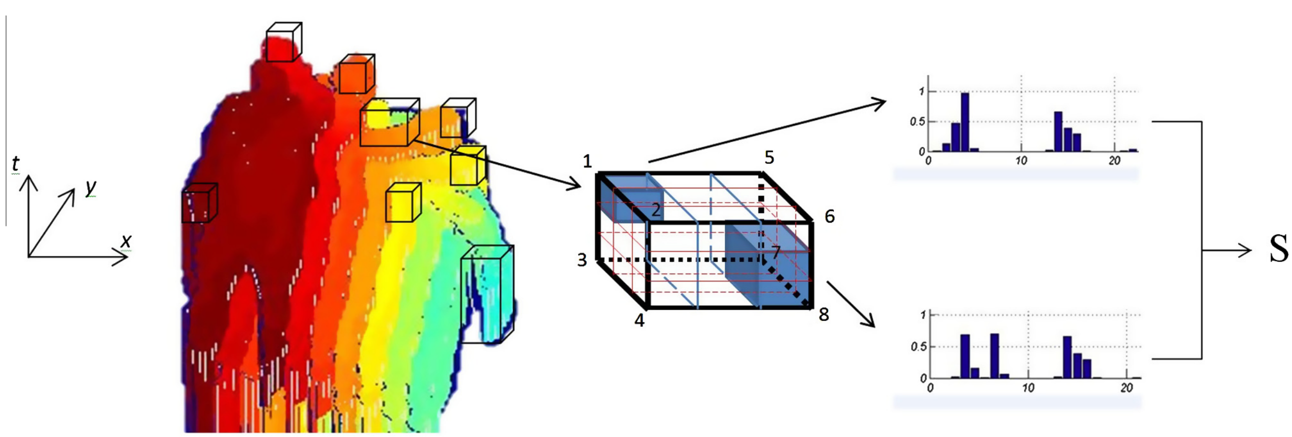

Fig. 4. Example of extracting local spatio-temporal feature from depth video [97].

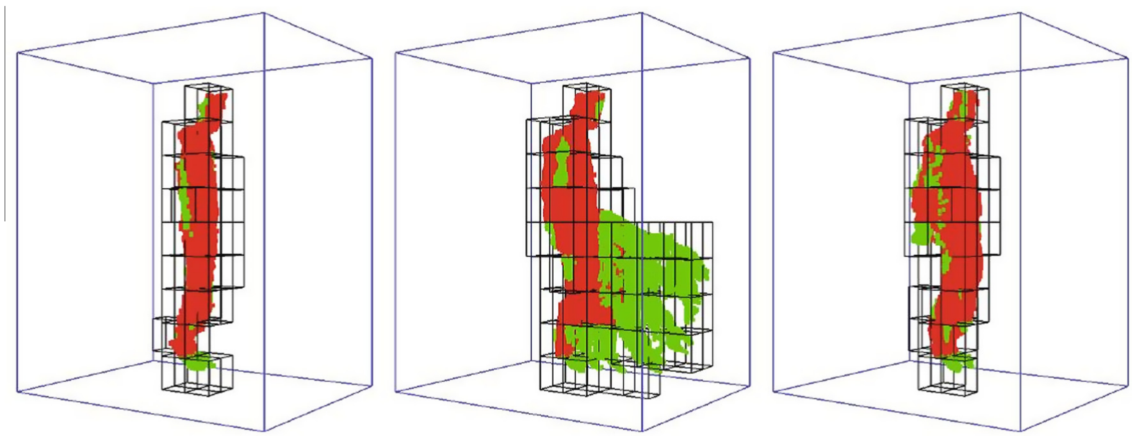

Fig. 5. Example of space–time occupancy pattern of action *Forward Kick* [87].

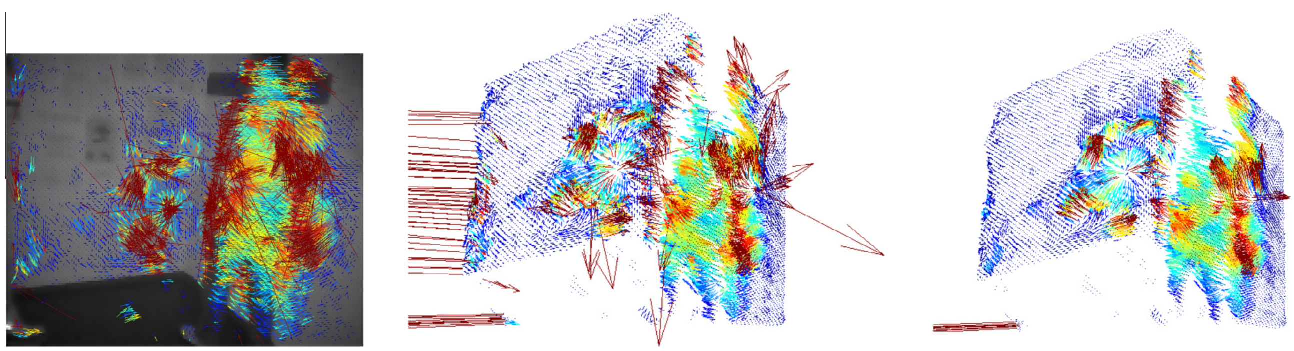

Fig. 6. Example of 3D scene flow: (a) 2D velocity vectors computed using optical flow, (b) 3D velocity vectors utilizing point correspondences, (c) the latter smoothed component wise by a median filter. Each 3th velocity vector is displayed and color coded with respect to its length: red denoting a big motion vector and blue a small one. [82].

puter vision, as the information lost in projection from 3D to 2D in the traditional intensity image may partially be recovered from the sensor. Many algorithms have been proposed to address human activity recognition using 3D data. This article gives an overview of this rapidly growing field, with a focus on recent works on activity recognition from range data. The algorithms are categorized based on the features that are employed to describe the depth data. Details of extracting the features in different ways are presented. Furthermore, the advantages and limitations of each type of feature are analyzed. (see Tables 2 and 3).

We believe this is just the beginning of the low-cost high-end range sensors and the related research. With future developments, range sensors will have higher resolution, less noise, and an extended sensing range. Furthermore, depth sensors accompanied by traditional cameras on laptops and cell phones are coming in the near future, which will provide broader computer vision research topics and applications. In all, the fast growing techniques of 3D sensing and the increasing availability of depth sensors creates a bright future for computer vision research.

Different algorithms on 3D data for activity recognition are emerging rapidly. Currently, most of the algorithms regard the depth channel and RGB channels independently. This may not be the optimal way since the RGB and D reside together in the 3D world; it may be productive to develop algorithms with improved fusion of the two types of data. In addition, since more structural information about the scene and the subject is given by the sensor, the realization of a view-invariant recognition system is becoming easier than before. However, the existing works on view-invariant analysis and solutions with depth data is still lacking when compared to the intensive studies in RGB imagery. Future works may better explore this aspect to make the algorithm view-invariant with the help of depth data.

Furthermore, low level challenges such as shadows and illumination changes are alleviated by the acquisition of the depth channel. However, occlusion is still a problem since current depth sensing techniques give a 2.5D estimation of the scene. Most existing datasets do not involve enough occlusion cases. Future recognition algorithms may devote more attention to obtaining robustness to occlusion. This step is essential for the algorithm to work in real scenarios. In addition, current approaches mostly deal with a single human subject, partially due to the limitation of the existing range sensor. With the development of better sensing technologies, depth data with more subjects or even groups of people may become available. Also, algorithms on group activity recognition will emerge. Current applications include gaming and human–computer interactions. Future applications may cover various aspects of a person's daily life, making our life convenient and totally changing the way we interact with the world.

References

[1] J. Aggarwal, M.S. Ryoo, Human activity analysis: a review, ACM Comput. Surv. (CSUR) 43 (2011) 16.

[2] M. Ahmad, S.W. Lee, Hmm-based human action recognition using multiview image sequences in: ICPR, (2006) 263–266.

[3] V. Argyriou, M. Petrou, S. Barsky, Photometric stereo with an arbitrary number of illuminants, CVIU 114 (2010) 887–900.

[4] F. Arman, J. Aggarwal, Model-based object recognition in dense-range images review, ACM Comput. Surv. (CSUR) 25 (1993) 5–43.

[5] G. Ballin, M. Munaro, E. Menegatti, Human action recognition from rgb-d frames based on real-time 3d optical flow estimation, in: Biologically Inspired Cognitive Architectures. Springer, 2013, pp. 65–74.

[6] J.L. Barron, D.J. Fleet, S. Beauchemin, Performance of optical flow techniques, IJCV 12 (1994) 43–77.
[7] J. Ben-Arie, Z. Wang, P. Pandit, S. Rajaram, Human activity recognition using multidimensional indexing, Pattern Anal. Mach. Intell. IEEE Trans. 24 (2002) 1091–1104.
[8] A. Blinov, M. Petrou, Reconstruction of 3-d horizons from 3-d seismic datasets, Geosci. Remote Sens. 43 (2005) 1421–1431.
[9] V. Bloom, D. Makris, V. Argyriou, G3d: a gaming action dataset and real time action recognition evaluation framework, in: CVPRW, 2012, pp. 7–12.
[10] J. Cech, J. Sanchez-Riera, R. Horaud, Scene flow estimation by growing correspondence seeds, in: CVPR, IEEE. 2011, pp. 3129–3136.
[11] R. Chaudhry, A. Ravichandran, G. Hager, R. Vidal, Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions, in: CVPR, IEEE, 2009, pp. 1932–1939.
[12] L. Chen, H. Wei, J.M. Ferryman, A survey of human motion analysis using depth imagery, Pattern Recognit. Lett. (2013).
[13] L.C. Chen, J.W. Hsieh, C.H. Chuang, C.Y. Huang, D.Y. Chen, Occluded human action analysis using dynamic manifold model, ICPR, IEEE (2012) 1245–1248.
[14] M.Y. Chen, A. Hauptmann, Mosift: Recognizing human actions in surveillance videos, 2009.
[15] Z. Cheng, L. Qin, Y. Ye, Q. Huang, Q. Tian, Human daily action analysis with multi-view and color-depth data, in: ECCV Workshops and Demonstrations, Springer, 2012, pp. 52–61.
[16] A. Chiuso, P. Favaro, H. Jin, S. Soatto, Structure from motion causally integrated over time, Pattern Anal. Mach. Intell. IEEE Trans. 24 (2002) 523–535.
[17] W.J. Christmas, J. Kittler, M. Petrou, Structural matching in computer vision using probabilistic relaxation, PAMI 17 (1995) 749–764.
[18] I. Cohen, H. Li, Inference of human postures by classification of 3d human body shape, AMFG Workshop, IEEE (2003) 74–81.
[19] N. Dalal, B. Triggs, C. Schmid, Human detection using oriented histograms of flow and appearance, in: ECCV. Springer, 2006, pp. 428–441.
[20] T. Darrell, G. Gordon, M. Harville, J. Woodfill, Integrated person tracking using stereo, color, and pattern detection, IJCV 37 (2000) 175–185.
[21] J.W. Davis, A.F. Bobick, The representation and recognition of human movement using temporal templates, in: CVPR, IEEE, 1997, pp. 928–934.
[22] L. Deng, H. Leung, N. Gu, Y. Yang, Generalized model-based human motion recognition with body partition index maps, in: Computer Graphics Forum, Wiley Online Library, 2012, pp. 202–215.
[23] U.R. Dhond, J.K. Aggarwal, Structure from stereo-a review, Syst. Man Cybern. IEEE Trans. 19 (1989) 1489–1510.
[24] P. Dollár, V. Rabaud, G. Cottrell, S. Belongie, Behavior recognition via sparse spatio-temporal features, PETS (2005) 65–72.
[25] M. Domínguez-Morales, A. Jiménez-Fernández, R. Paz-Vicente, A. Linares-Barranco, G. Jiménez-Moreno, 2012. Stereo matching: From the basis to neuromorphic engineering.
[26] S.R. Fanello, I. Gori, G. Metta, F. Odone, Keep it simple and sparse: real-time action recognition, J. Mach. Learn. Res. 14 (2013) 2617–2640.
[27] P. Favaro, S. Soatto, Learning shape from defocus, in: ECCV. Springer, 2002, pp. 735–745.
[28] H. Fujiyoshi, A.J. Lipton, Real-time human motion analysis by image skeletonization, in: Applications of Computer Vision, 1998. WACV'98. Proceedings., Fourth IEEE Workshop on, IEEE, 1998, pp. 15–21.
[29] D. Gehrig, T. Schultz, Selecting relevant features for human motion recognition, in: ICPR, IEEE, 2008.
[30] N. Georgis, M. Petrou, J. Kittler, Error guided design of a 3d vision system, PAMI 20 (1998) 366–379.
[31] J.J. Gibson, The perception of the visual world, 1950.
[32] R. Grzeszcuk, G. Bradski, M.H. Chu, J.Y. Bouguet, Stereo based gesture recognition invariant to 3d pose and lighting, CVPR (2000) 826–833.
[33] M. Harville, D. Li, Fast, integrated person tracking and activity recognition with plan-view templates from a single stereo camera, in: CVPR, 2004.
[34] A. Hernández-Vela, M. Bautista, X. Perez-Sala, V. Ponce, X. Baró, O. Pujol, C. Angulo, S. Escalera, Bovdw: Bag-of-visual-and-depth-words for gesture recognition, in: ICPR, 2012.
[35] C. Hesher, A. Srivastava, G. Erlebacher, A novel technique for face recognition using range imaging, in: Signal processing and its applications, Seventh international symposium on, IEEE, 2003, pp. 201–204.
[36] M.B. Holte, T.B. Moeslund, P. Fihl, View-invariant gesture recognition using 3d optical flow and harmonic motion context, CVIU 114 (2010) 1353–1361.
[37] M.B. Holte, T.B. Moeslund, N., Nikolaidis, I. Pitas, 3d human action recognition for multi-view camera systems, in: 3DIMPVT, 2011, pp. 342–349.
[38] B.K. Horn, B.G. Schunck, Determining optical flow, Artif. Intell. 17 (1981) 185–203.
[39] F. Huguet, F. Devernay, A variational method for scene flow estimation from stereo sequences, in: ICCV, IEEE, 2007, pp. 1–7.
[40] A. Jalal, M.Z. Uddin, J.T. Kim, T.S. Kim, Recognition of human home activities via depth silhouettes and R transformation for smart homes, Indoor Built Environ. (2011) 1–7.
[41] H. Jin, P. Favaro, A variational approach to shape from defocus, in: ECCV. Springer, 2002, pp. 18–30.
[42] G. Johansson, Visual motion perception. Scientific American, 1975.
[43] T. Kanade, A. Yoshida, K. Oda, H. Kano, M. Tanaka, A stereo machine for video-rate dense depth mapping and its new applications, in: CVPR, IEEE. 1996, pp. 196–202.
[44] A. Klaser, M. Marszalek, A spatio-temporal descriptor based on 3d-gradients, in: BMVC, 2008.
[45] V.A. Kovalev, M. Petrou, Y.S. Bondar, Texture anisotropy in 3-d images, Image Proces. IEEE Trans. 8 (1999) 346–360.
[46] D. Kulić, W. Takano, Y. akamura, Incremental learning, clustering and hierarchy formation of whole body motion patterns using adaptive hidden markov chains, Int. J. Rob. Res. 27 (2008) 761–784.
[47] A. Kurakin, Z. Zhang, Z. Liu, A real time system for dynamic hand gesture recognition with a depth sensor, 2012.
[48] I. Laptev, On space-time interest points, Int. J. Comput. Vis. 64 (2005) 107–123.
[49] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, CVPR (2008) 1–8.
[50] J. Lei, X. Ren, D. Fox, Fine-grained kitchen activity recognition using rgb-d 2012.
[51] A. Letouzey, B. Petit, E. Boyer, M. Team, Scene flow from depth and color images, in: Jesse Hoey, Stephen McKenna and Emanuele Trucco, Proceedings of the British Machine Vision Conference, pages, 2011, pp. 46–1.
[52] W. Li, Z. Zhang, Z. Liu, Action recognition based on a bag of 3d points, CVPR Workshop 1 (2010) 9–14.
[53] A.M. Loh, The recovery of 3-D structure using visual texture patterns. University of Western Australia, 2006.
[54] B.D. Lucas, T. Kanade, et al., An iterative image registration technique with an application to stereo vision., in: IJCAI, 1981, pp. 674–679.
[55] F. Lv, R. Nevatia, Recognition and segmentation of 3-d human action using hmm and multi-class adaboost, ECCV (2006) 359–372.
[56] S. Malassiotis, N. Aifanti, M. Strintzis, A gesture recognition system using 3d data, in: 3D Data Processing Visualization and Transmission, 2002. Proceedings. First International Symposium on, IEEE, 2002, pp. 190–193.
[57] S.Z. Masood, C. Ellis, A. Nagaraja, M.F. Tappen, J. LaViola, R. Sukthankar, Measuring and reducing observational latency when recognizing actions, ICCV Workshops, IEEE (2011) 422–429.
[58] Y. Matsumoto, A. Zelinsky, An algorithm for real-time stereo vision implementation of head pose and gaze direction measurement, Autom. Face Gesture Recognit. IEEE (2000) 499–504.
[59] R. Muñoz-Salinas, E. Aguirre, M. García-Silvente, People detection and tracking using stereo vision and color. Image and Vision Computing 25, 2007, 995–1007.
[60] B. Ni, G. Wang, P. Moulin, Rgbd-hudaact: a color-depth video database for human daily activity recognition, ICCV Workshop (2011) 1147–1153.
[61] A. Ogale, A. Karapurkar, G. Guerra-Filho, Y. Aloimonos, View-invariant identification of pose sequences for action recognition, in: VACE, Citeseer, 2004.
[62] A. Oikonomopoulos, I. Patras, M. Pantic, Spatiotemporal salient points for visual recognition of human actions, Syst. Man Cybern. Part B: Cybern. IEEE Trans. 36 (2005) 710–719.
[63] V. Parameswaran, R. Chellappa, View invariance for human action recognition, IJCV 66 (2006) 83–101.
[64] R. Poppe, A survey on vision-based human action recognition, Image Vision Comput. 28 (2010) 976–990.
[65] E. Prados, O. Faugeras, Unifying approaches and removing unrealistic assumptions in shape from shading: mathematics can help, ECCV. Springer (2004) 141–154.
[66] E. Prados, O. Faugeras, Shape from shading: a well-posed problem?, CVPR, IEEE (2005) 870–877
[67] P. Pritchett, A. Zisserman, Wide baseline stereo matching, ICCV, IEEE (1998) 754–760.
[68] M.C. Roh, H.K. Shin, S.W. Lee, View-independent human action recognition with volume motion template on single stereo camera, Pattern Recognit. Lett. 31 (2010) 639–647.
[69] D.B. Russakoff, M. Herman, Head tracking using stereo, Mach. Vis. Appl. 13 (2002) 164–173.
[70] R. Rusu, N. Blodow, M. Beetz, Fast point feature histograms (fpfh) for 3d registration, ICRA (2009) 3212–3217.
[71] B. Sabata, J. Aggarwal, Estimation of motion from a pair of range images: A review, CVGIP: Image Understanding 54 (1991) 309–324.
[72] B. Sabata, J. Aggarwal, Surface correspondence and motion computation from a pair of range images, Comput. Vision Image Understanding 63 (1996) 232–250.
[73] H. Saito, M. Mori, Application of genetic algorithms to stereo matching of images, Pattern Recognit. Lett. 16 (1995) 815–821.
[74] P. Scovanner, S. Ali, M. Shah, A 3-dimensional sift descriptor and its application to action recognition, in: Proceedings of the 15th international conference on Multimedia, ACM, 2007, pp. 357–360.
[75] E. Seemann, K. Nickel, R. Stiefelhagen, Head pose estimation using stereo vision for human-robot interaction, Autom. Face Gesture Recognit. IEEE (2004) 626–631.
[76] S. Sempena, N. Maulidevi, P. Aryan, Human action recognition using dynamic time warping, ICEEI, IEEE (2011) 1–5.
[77] Y. Shen, H. Foroosh, View-invariant action recognition using fundamental ratios, in: CVPR, IEEE, 2008.
[78] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, A., Blake, Real-time human pose recognition in parts from a single depth image. CVPR, 2011.
[79] I. Stamos, P. Allen, 3-d model construction using range and image data, CVPR, IEEE (2000) 531–536.

[80] J. Sung, C. Ponce, B. Selman, A. Saxena, Human activity detection from RGBD images. PAIR, 2011.

[81] J. Sung, C. Ponce, B. Selman, A. Saxena, Unstructured human activity detection from rgbd images, ICRA, IEEE (2012) 842–849.

[82] A. Swadzba, N. Beuter, J. Schmidt, G. Sagerer, Tracking objects in 6d for reconstructing static scenes, CVPRW, IEEE (2008) 1–7.

[83] P. Turaga, R. Chellappa, V.S. Subrahmanian, O. Udrea, Machine recognition of human activities: A survey, Circuits Syst. Video Technol. IEEE Trans. 18 (2008) 1473–1488.

[84] M.Z. Uddin, N.D. Thang, J.T. Kim, T.S. Kim, Human activity recognition using body joint-angle features and hidden markov model, ETRI J. 33 (2011) 569–579.

[85] R. Urtasun, P. Fua, 3d tracking for gait characterization and recognition, in: Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on, IEEE, 2004, pp. 17–22.

[86] S. Vedula, S. Baker, P. Rander, R. Collins, T. Kanade, Three-dimensional scene flow, ICCV, IEEE (1999) 722–729.

[87] A. Vieira, E. Nascimento, G. Oliveira, Z. Liu, M. Campos, Stop: Space-time occupancy patterns for 3d action recognition from depth map sequences, Progress Pattern Recognit. Image Anal. Comput. Vision Appl. (2012) 252–259.

[88] H. Wang, M.M., Ullah, A. Klaser, I. Laptev, C. Schmid, et al., Evaluation of local spatio-temporal features for action recognition, in: BMVC, 2009.

[89] J. Wang, Z. Liu, J. Chorowski, Z. Chen, Y. Wu, Robust 3d action recognition with random occupancy patterns, in: ECCV. Springer, 2012a, pp. 872–885.

[90] J. Wang, Z. Liu, Y. Wu, J. Yuan, Mining actionlet ensemble for action recognition with depth cameras, CVPR (2012) 1290–1297.

[91] A. Wedel, T. Brox, T. Vaudrey, C. Rabe, U. Franke, D. Cremers, Stereoscopic scene flow computation for 3d motion understanding, IJCV 95 (2011) 29–51.

[92] D. Weinland, E. Boyer, R. Ronfard, Action recognition from arbitrary views using 3d exemplars, ICCV, IEEE (2007) 1–7.

[93] D. Weinland, M. Özuysal, P. Fua, Making action recognition robust to occlusions and viewpoint changes, ECCV. Springer (2010) 635–648.

[94] G. Willems, T. Tuytelaars, L. Van Gool, An efficient dense and scale-invariant spatio-temporal interest point detector, ECCV (2008) 650–663.

[95] C. Wolf, J. Mille, E. Lombardi, O. Celiktutan, M. B. Jiu, E. Dellandrea, C.E., Bichot, C. Garcia, B. Sankur, The liris human activities dataset and the icpr 2012 human activities recognition and localization competition. Technical Report RR-LIRIS-2012-004, LIRIS Laboratory.

[96] D. Wu, F. Zhu, L. Shao, One shot learning gesture recognition from rgbd images, CVPRW, IEEE (2012) 7–12.

[97] L. Xia, J. Aggarwal, Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera, in: CVPR, IEEE, 2013.

[98] L. Xia, C.C. Chen, J. Aggarwal, View invariant human action recognition using histograms of 3d joints, CVPRW, IEEE (2012) 20–27.

[99] C. Yang, G. Medioni, Object modelling by registration of multiple range images, Image Vision Comput. 10 (1992) 145–155.

[100] X. Yang, Y. Tian, Eigenjoints-based action recognition using naïve-bayes-nearest-neighbor, in: CVPRW, IEEE, 2012, pp. 14–19.

[101] X. Yang, C. Zhang, Y. Tian, Recognizing actions using depth motion maps-based histograms of oriented gradients, 2012.

[102] A. Yao, J. Gall, G. Fanelli, L. Van Gool, Does human action recognition benefit from pose estimation?, in: BMVC, 2011.

[103] A. Yao, J. Gall, L. Van Gool, A hough transform-based voting framework for action recognition, CVPR, IEEE (2010) 2061–2068.

[104] E. Yu, J. Aggarwal, Human action recognition with extremities as semantic posture representation, in: Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on, IEEE, 2009, pp. 1–8.

[105] Y.K. Yu, K.H. Wong, S.H. Or, M.M.Y. Chang, Robust 3-d motion tracking from stereo images: A model-less method, Instrum. Meas. IEEE Trans. 57 (2008) 622–630.

[106] K. Yun, J. Honorio, D. Chattopadhyay, T.L. Berg, D. Samaras, Two-person interaction detection using body-pose features and multiple instance learning, CVPRW, IEEE (2012) 28–35.

[107] C. Zhang, Y. Tian, Rgbd camera-based daily living activity recognition, 2012.

[108] H. Zhang, L. Parker, 4-dimensional local spatio-temporal features for human activity recognition, IROS (2011) 2044–2049.

[109] Y. Zhao, Z. Liu, L. Yang, H. Cheng, Combing rgb and depth map features for human activity recognition, APSIPA ASC, IEEE (2012) 1–4.

[110] Mitiche, Amar and Aggarwal, Jagdishkumar Keshoram, Computer Vision Analysis of Image Motion by Variational Methods, Springer, 2013.