Biologically Inspired Cognitive Architectures (2013) **5**, 42–51

Available at www.sciencedirect.com

SciVerse ScienceDirect

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

INVITED ARTICLE

3D flow estimation for human action recognition from colored point clouds

Matteo Munaro[*], Gioia Ballin, Stefano Michieletto, Emanuele Menegatti

Intelligent Autonomous Systems Laboratory, University of Padua, Via Gradenigo 6A, 35131 Padua, Italy

Received 12 March 2013; received in revised form 18 May 2013; accepted 18 May 2013

KEYWORDS
Action recognition;
3D motion flow;
IAS-Lab Action Dataset;
Colored point clouds;
RGB-D data;
Kinect

Abstract
Motion perception and classification are key elements exploited by humans for recognizing actions. The same principles can serve as a basis for building cognitive architectures which can recognize human actions, thus enhancing challenging applications such as human robot interaction, visual surveillance, content-based video analysis and motion capture. In this paper, we propose an autonomous system for real-time human action recognition based on 3D motion flow estimation. We exploit colored point cloud data acquired with a Microsoft Kinect and we summarize the motion information by means of a 3D grid-based descriptor. Finally, temporal sequences of descriptors are classified with the Nearest Neighbor technique. We also present a newly created public dataset for RGB-D human action recognition which contains 15 actions performed by 12 different people. Our overall system is tested on this dataset and on the dataset used in Ballin, Munaro, and Menegatti (2012), showing the effectiveness of the proposed approach in recognizing about 90% of the actions.

1. Introduction

The challenge of endowing robotic agents with human-like capabilities is a hot topic addressed by the cognitive robotics community. In that field, the aim is to create smart agents able to efficiently perform complex tasks in partially observable environments. In order to achieve real-world goals, a cognitive robot is equipped with a processing architecture that combines perception, cognition and action modules in the most effective way. Then, a cognitive robot is not only able to perceive complex stimuli from the environment, but also to reason about them and to act coherently. Furthermore, it should also be able to safely interact and cooperate with humans.

For a robust and fluid interaction, a robot is required to have high perception capabilities, in particular it should be endowed with a reliable human action recognition system able to recognize actions immediately after they are performed. In this context, we propose a method for real-time

[*] Corresponding author.
E-mail addresses: munaro@dei.unipd.it (M. Munaro), gioia.ballin@gmail.com (G. Ballin), michieletto@dei.unipd.it (S. Michieletto), emg@dei.unipd.it (E. Menegatti).

2212-683X/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.bica.2013.05.008

human action recognition which allows to recognize about 90% of actions at a medium frame rate of four frames per second. Our system exploits the Robot Operating System (Quigley et al., 2009) as a framework. We use the Microsoft Kinect sensor and the tracking system described in Basso, Munaro, Michieletto, and Menegatti (2012) and Munaro, Basso, and Menegatti (2012) to robustly detect and track people in the scene. Then, by taking inspiration from the way humans recognize actions, we estimate frame by frame 3D motion of person points from point cloud data only. The overall person motion is described with a grid-based descriptor and frame-wise descriptors are concatenated to compose sequence descriptors, which are then classified with a Nearest Neighbor classifier.

The main contributions of the paper are: a method for real time 3D flow estimation from point cloud data, a 3D grid-based descriptor which encodes the whole person motion and a newly created dataset which contains RGB-Depth (RGB-D) and skeleton data for 15 actions performed by 12 different people.

The remainder of the paper is organized as follows: at first, we provide a complete review about the recent advances in human action recognition systems and the existing datasets for 3D human action recognition. Then, we present the novel *IAS-Lab Action Dataset*. After that, the 3D motion flow estimation algorithm is described, together with the descriptors used for encoding person motion. Finally, we report experiments on the *IAS-Lab Action Dataset* and on the dataset used in Ballin et al. (2012), and we outline conclusions and future works.

2. Related work

Human action recognition is an active research area in computer vision. First investigations about this topic began in the seventies with pioneering studies accomplished by Johansson (1973). From then on, the interest in the field rapidly increased, motivated by a number of potential real-world applications such as video surveillance, human—computer interaction, content-based video analysis and retrieval. Moreover, in recent years, the task of recognizing human actions has gained increasingly popularity thanks to the emergence of modern applications such as motion capture and animation, video editing and service robotics. Most of the works on human action recognition rely on information extracted from 2D images and videos (Poppe, 2010). These approaches mostly differ in the features representation. We can distinguish between methods which exploit global traits of the human body and those which extract local information from the data. Popular global representations are edges (Carlsson & Sullivan, 2001), silhouettes of the human body (Blank, Gorelick, Shechtman, Irani, & Basri, 2005; Rusu, Bandouch, Meier, Essa, & Beetz, 2009; Yilmaz & Shah, 2005), 2D optical flow (Ali & Shah, 2010; Efros, Berg, Mori, & Malik, 2003; Yacoob & Black, 1998) and 3D spatio-temporal volumes (Blank et al., 2005; Ke, Sukthankar, & Hebert, 2005; Liu, Ali, & Shah, 2008; Rusu et al., 2009; Scovanner, Ali, & Shah, 2007; Yilmaz & Shah, 2005). Conversely, effective local representations mainly refer to Laptev and Lindeberg (2003), Laptev, Marszalek, Schmid, and Rozenfeld (2008), Dollar, Rabaud, Cottrell, and Belongie (2005), Schuldt, Laptev, and Caputo (2004), Niebles, Wang, and Fei-Fei (2008), Kläser, Marszałek, and Schmid (2008), and Weinland, Ronfard, and Boyer (2006).

The recent spread of inexpensive RGB-D sensors has paved the way to new studies in this direction. These 3D recognition systems are inherently related to the acquisition of a more informative input content with respect to their 2D counterparts and thus it is expected they outperform the traditional approaches. The first RGB-D related work is signed by Microsoft Research (Li, Zhang, & Liu, 2010). In Li et al. (2010), a sequence of depth maps is given as input to the system. Next, the relevant postures for each action are extracted and represented as a bag of 3D points. The motion dynamics are modeled by means of an action graph and a Gaussian Mixture Model is used to robustly capture the statistical distribution of the points. Subsequent studies mainly refer to the use of three different technologies: Time of Flight cameras (Holte & Moeslund, 2008; Holte, Moeslund, Nikolaidis, & Pitas, 2011), motion capture systems (Korner & Denzler, 2012; Ofli, Chaudhry, Kurillo, Vidal, & Bajcsy, 2012; Thuc, Tuan, & Hwang, 2012) and active matricial triangulation systems (i.e.: Kinect-style cameras) (Bloom, Makris, & Argyriou, 2012; Lei, Ren, & Fox, 2012; Ming, Ruan, & Hauptmann, 2012; Ni, Wang, & Moulin, 2011; Popa, Koc, Rothkrantz, Shan, & Wiggers, 2011; Sung, Ponce, Selman, & Saxena, 2011, 2012; Xia, Chen, & Aggarwal, 2012; Yang & Tian, 2012; Zhang & Parker, 2011; Zhao, Liu, Yang, & Cheng, 2012). We can further distinguish these works by means of the features used during the recognition process. The most used features are related to the extraction of the skeleton body joints (Bloom et al., 2012; Ofli et al., 2012; Sung et al., 2011, Sung, Ponce, Selman, & Saxena, 2012; Thuc et al., 2012; Xia et al., 2012; Yang & Tian, 2012). Usually, these approaches first collect raw information about the body joints (e.g.: spatial coordinates, angle measurements). Next, they summarize the raw data into features through specific computations, in order to characterize the posture of the observed human body. Differently from the other joints-related publications, Ofli et al. (2012) distinguishes itself by computing features which carry a physical meaning. Indeed, in Ofli et al. (2012), a *Sequence of Most Informative Joints* (SMIJ) is computed based on measures like the mean of joint angles, the variance of joint angles and the maximum angular velocity of body joints.

Other popular features are the result of the extension to the third dimension of typical 2D representations. Within this feature category, we shall perform a further distinction between local representations and global representations. Features in (Ming et al., 2012; Ni et al., 2011; Zhang & Parker, 2011; Zhao et al., 2012) are actually local representations since they aim to exploit *Space-Time Interest Points* (STIPs) (Laptev & Lindeberg, 2003; Laptev, Caputo, Schüldt, & Lindeberg, 2007; Laptev et al., 2008) by extending it with the depth information. Examples of global representations in the 3D field can be found in Popa et al. (2011), Holte and Moeslund (2008), and Holte et al. (2011). In Popa et al. (2011), Popa et al. propose a Kinect-based system able to continuously analyze customers' shopping behaviors in malls. Silhouettes for each person in the scene are extracted and then summarized by computing moment invariants. In Holte and Moeslund (2008) and Holte et al. (2011), a 3D extension of 2D optical flow is exploited for the gesture recognition task. Holte et al. compute optical flow in the image using the traditional Lukas—Kanade (Lukas & Kanade, 1981) method and then extend the 2D velocity vectors to incorporate also the depth dimension. At the end of this process, the 3D velocity vectors are used to create an annotated velocity cloud. 3D Motion Context and Harmonic Motion Context serve the task of representing the extracted motion vector field in a view-invariant way. With regard to the classification task, Holte and Moeslund (2008) and Holte et al. (2011) do not follow a learning-based approach, instead a probabilistic distance classifier is proposed in order to identify which gesture best describes a string of primitives. Note that Holte et al. (2011) differs from Holte and Moeslund (2008) because the optical flow is estimated from each view of a multi-camera system and is then combined into a unique 3D motion vector field.

Finally, works in which trajectory features are exploited (Lei et al., 2012; Korner & Denzler, 2012) emerged recently. While in Lei et al. (2012) trajectory gradients are computed and summarized, in Korner and Denzler (2012) an action is represented as a set of subspaces and a mean shape.

From the application point of view Sung et al. (2011, 2012) and Ni et al. (2011) are targeted to applications in the personal robotics field, while Li et al. (2010) and Yang and Tian (2012) are addressed to human—computer interaction and gaming applications. Finally, Zhang and Parker (2011) and Popa et al. (2011) are primarily addressed to applications in the field of video surveillance.

Unlike Holte and Moeslund (2008) and Holte et al. (2011), which compute 2D optical flow and then extend it to 3D, we proposed in Ballin et al. (2012) a method to compute the motion flow directly on 3D points with color. From the estimated 3D velocity vectors, a motion descriptor is derived and a sequence of descriptors is concatenated and classified by means of Nearest Neighbor. Tests are reported on a dataset of six actions performed by six different actors.

3. Dataset

The rapid dissemination of inexpensive RGB-D sensors, such as Microsoft Kinect (Kinect), boosted the research on 3D action recognition. At the same time a new need arose: the acquisition of new datasets in which the RGB stream is aligned with the depth stream. Currently, the following datasets have been released: RGBD-HuDaAct Database (Zhang & Parker, 2011), Indoor Activity Database (Sung et al., 2012), MSR-Action3D Dataset (Wanqing et al., 2010), MSR-DailyActivity3D Dataset (Jiang Wang, Zicheng Liu, & Yuan, 2012), LIRIS Human Activities Dataset (Wolf et al., 2012) and Berkeley MHAD (Ofli, Chaudhry, Kurillo, Vidal, & Bajcsy, 2013). All these datasets are targeted to recognition tasks in indoor environments. The first two are thought for personal or service robotics applications, while the two from MSR are also targeted to gaming and human—computer interaction. The LIRIS dataset concerns actions performed from both single persons and groups, acquired in different scenarios and changing the point of view. The last one was acquired using a multimodal system (mocap, video, depth, acceleration, audio) to provide a very controlled set of actions to test algorithms across multiple modalities.

3.1. IAS-Lab Action Dataset

Two key features of a good dataset are size and variability. Moreover, it should allow to compare as many different algorithms as possible. For the RGB-D action recognition task, that means that there should be enough different actions, many different people performing them and RGB and depth synchronization and registration. Moreover, the 3D skeleton of the actors should be saved, given that it is easily available and many recent techniques rely on it. Hovever, we noticed the lack of a dataset having all these features, thus we acquired the *IAS-Lab Action Dataset*,[1] which contains 15 different actions performed by 12 different people. Each person repeats each action three times, thus leading to 540 video samples. All these samples are provided as ROS `bags` containing synchronized and registered RGB images, depth images and point clouds and ROS `tf` for every skeleton joint as they are estimated by the NITE middleware (NITE middleware). Unlike Ofli et al. (2013), we preferred NITE's skeletal tracker to a motion capture technology in order to test our algorithms on data that could be easily available on a mobile robot and, unlike Wolf et al. (2012), we asked the subjects to perform well defined actions, because, beyond a certain level, variability could bias the evaluation of an algorithm performance.

In Table 1, the *IAS-Lab Action Dataset* is compared to the already mentioned datasets, while in Fig. 1 an example image for every action is reported.

4. 3D motion flow

Optical flow is a powerful cue to be used for a variety of applications, from motion segmentation to structure-from-motion passing by video stabilization. As reported in the introduction section, some researchers proved its usefulness also for the task of action recognition (Ali & Shah, 2010; Efros et al., 2003; Yacoob & Black, 1998). The most famous algorithm for optical flow estimation was proposed by Lukas and Kanade (1981). The main drawbacks of this approach were that it only works for highly textured image patches and, if repeated for every pixel of an image, it results to be highly computational expensive. Moreover, 2D motion estimation in general has the limitation to be dependent on the viewpoint and closer objects appear to move faster because they appear bigger in the image.

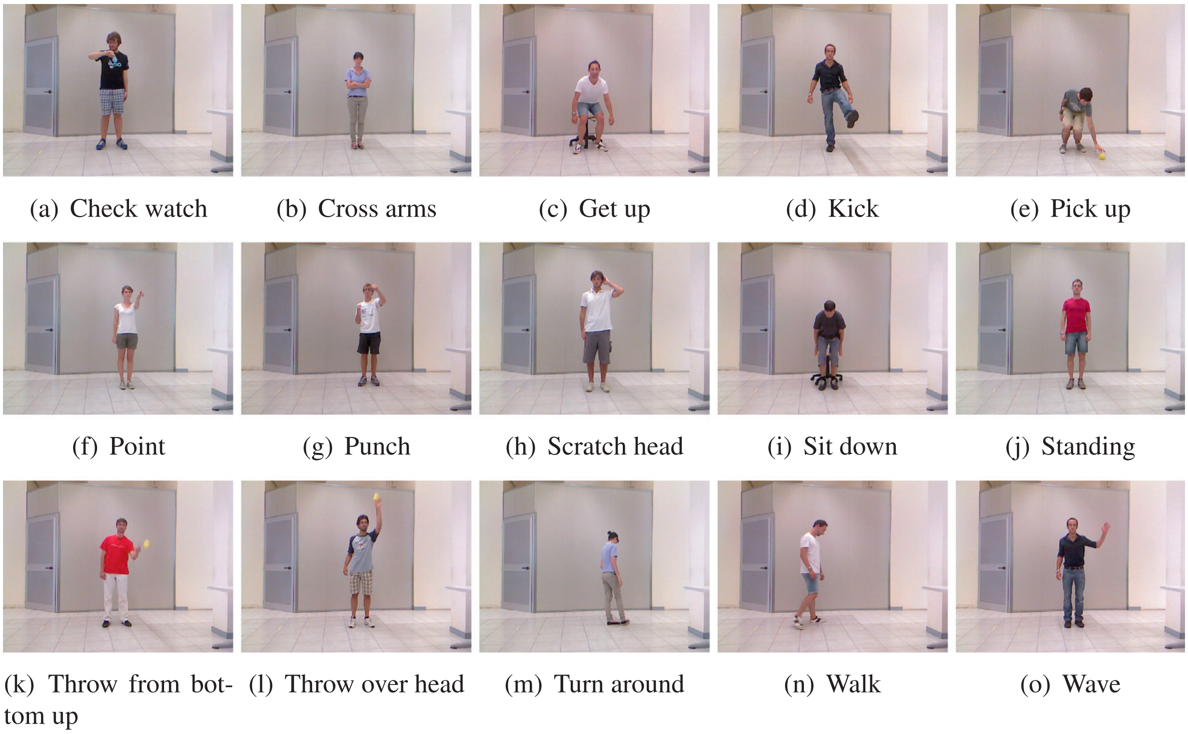
When depth data are available and registered to the RGB/intensity image, the optical flow computed in the image can be extended to 3D by looking at the corresponding points in the depth image or point cloud (Holte & Moeslund, 2008; Holte et al., 2011). This procedure allows to compute 3D velocity vectors, thus overcoming some of the limitations of 2D-only approaches, such as viewpoint and scale dependence. However, the motion estimation process is still executed on the RGB image, and it does not exploit the available 3D information for obtaining a better estimate. Moreover, the computational burden is still high.

[1] http://robotics.dei.unipd.it/actions.

Table 1 Datasets for 3D human action recognition.

	#actions	#people	#samples	RGB	skel
Zhang and Parker (2011)	6	1	198	Yes	No
Sung et al. (2012)	12	4	48	Yes	Yes
Wanqing et al. (2010)	20	10	567	No[a]	Yes
Jiang Wang et al. (2012)	16	10	320	No	Yes
Wolf et al. (2012)[b]	10	21	461	Yes[c]	No
Ofli et al. (2013)	11	12	660	Yes	Yes[d]
Ours	15	12	540	Yes	Yes

[a] The RGB images are provided, but they are not synchronized with the depth images.
[b] Only the set provided with depth information was considered.
[c] The RGB information has been converted to grayscale.
[d] Obtained from motion capture data.


(a) Check watch (b) Cross arms (c) Get up (d) Kick (e) Pick up

(f) Point (g) Punch (h) Scratch head (i) Sit down (j) Standing

(k) Throw from bottom up (l) Throw over head (m) Turn around (n) Walk (o) Wave

Fig. 1 Examples of images for the 15 actions present in the *IAS-Lab Action Dataset*.

In this work, we introduce a novel technique for computing 3D motion of points in the 3D-color space directly. This method consists in estimating correspondences between points of clouds belonging to consecutive frames. Our approach is fast and able to overcome some singularities of optical flow estimation in images by relying also on 3D points coordinates. Moreover, it is applicable to any point cloud containing XYZ and RGB information, and not only to those derived from a 2D matrix of depth data (projectable point clouds).

4.1. 3D flow estimation pipeline

Given two point clouds (called *source* and *target*) containing 3D coordinates and RGB/HSV color values of an object of interest (in this work, a person), the following pipeline is applied:

1. correspondence finding: for every point of the target point cloud, we select K nearest neighbors in the source point cloud in terms of Euclidean distance in the XYZ space; among the resulting points, we select the nearest neighbor in terms of HSV coordinates. We preferred HSV to RGB because it is more perceptually uniform. If $\mathcal{N}_{\mathbf{p}_i^{target}}$ is the set of K nearest neighbors in the source point cloud to the point $\mathbf{p}_i$ in the target point cloud, then $\mathbf{p}_i^{target}$ is said to match with

$$\mathbf{p}_*^{source} = \mathrm{argmin}_{\mathbf{p}_i^{source} \in \mathcal{N}_{\mathbf{p}_i^{target}}} d_{HSV}\left(\mathbf{p}_i^{target}, \mathbf{p}_i^{source}\right), \qquad (1)$$

where d_{HSV} is the distance operator in the HSV space. The number of neighbors K is a function of the point cloud density. In this work, we filter the point clouds to have a voxel size of 0.02 m and we set K to 50. An illustration of this method with $K = 3$ is reported in Fig. 2.

2. outlier rejection by means of reciprocal correspondences: this method consists in estimating correspondences from target to source and from source to target. Then, points which match in both directions are kept.
3. computation of 3D velocity vectors $\mathbf{v}_i$ for every match i as spatial displacement over temporal displacement of corresponding 3D points $\mathbf{p}_i$ from target and source:

$$\mathbf{v}_i = \left(\mathbf{p}_i^{target} - \mathbf{p}_i^{source}\right) / \left(t_i^{target} - t_i^{source}\right) \quad (2)$$

4. unlike in Ballin et al. (2012), we perform an additional outlier rejection: points with 3D velocity magnitude $\|\mathbf{v}_i\|$ below a threshold are discarded. Isolated moving points (not near to other moving points) are also deleted. In particular, points moving faster than 0.3 m/s are retained and a moving point is considered to be isolated if none of its neighbors moves faster than 0.75 m/s.

The reciprocal correspondence technique for outlier rejection can be considered as a 3D extension of the *Template Inverse Matching* method (Liu, Li, Yuan, & He, 2008b), which has been widely used to estimate the goodness of 2D optical flow estimation. The constraints we apply on the flow magnitude and on the proximity to other moving points are thought to remove spurious estimates which can be generated from the noise inherent in the depth values.

In this work, we segment persons point clouds from the rest of the scene by means of the people detection and tracking method for RGB-D data described in Munaro et al. (2012) and then we apply the flow estimation algorithm to the detected persons clusters.

In Fig. 3, we report two consecutive RGB frames of a person performing the *Check Watch* action. Green arrows show magnitude and direction of the estimated flow when reprojected to the image. It can be noticed how outlier rejection manages to remove most of the noisy measurements, while preserving the real motion at the right arm position.

5. Feature descriptors

In this section, we describe the frame-wise and sequence-wise descriptors we extract for describing actions.

5.1. Gridded flow descriptor

In order to compute a descriptor accounting for direction and magnitude of motion of every body part, we center a 3D grid of suitable dimensions around a person point cloud. This grid divides the space around the person into a number of cubes. In Fig. 4, a person point cloud is reported, together with the 3D grid which divides its points into different clusters represented with different colors. The size of the grid is proportional to the person height in order to contain the whole limbs motion and to make the flow descriptor person-independent. For a person 1.75 m tall, the grid results to be 2 m width, 2.3 m tall and 1.8 m deep.

5.2. MEANFLOW and SUMFLOW

For every cube of the grid, we extract flow information from all the points inside the cube. In this work, we compare two kinds of descriptors: the former is similar to the one proposed in Ballin et al. (2012) and computes the mean 3D motion vector from every cube, while the latter computes the sum of the motion vectors of the current cube. We will refer to these descriptors with the name *MEANFLOW* and *SUMFLOW*, respectively. For both of them, the resulting vectors for all the cubes are concatenated into a single descriptor which is then L2-normalized for making the descriptor invariant to the speed at which an action is performed. In this work, the grid is divided into four parts in every dimension, thus the total number of cubes is $C = 64$.

5.3. Sequence descriptors

Since an action actually represents a sequence of movements over time, the use of multiple frames can provide more discriminant information to the recognition task with respect to approaches in which only a single-frame classification is performed. For this reason, we compose a single descriptor from every sequence of frames to be classified. In particular, we select a fixed number of frames evenly spaced in time from every pre-segmented sequence and we concatenate the single-frame descriptors to form a single sequence descriptor. Thanks to this approach, we not only exploit the motion information, but we also take into

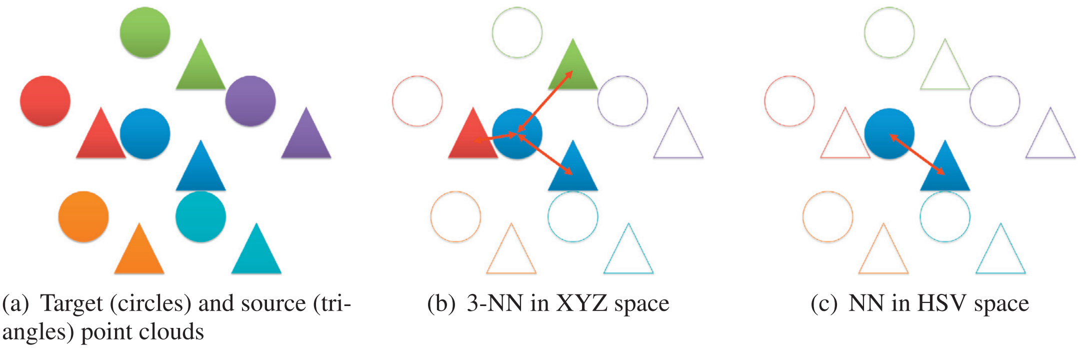

(a) Target (circles) and source (triangles) point clouds
(b) 3-NN in XYZ space
(c) NN in HSV space

Fig. 2 Illustration of the matching process between points of two point clouds represented by circles and triangles, respectively. In particular, a point of the target point cloud (blue circle) is matched with the corresponding point of the source point cloud (blue triangle) after (b) K-Nearest Neighbor (K-NN) in the XYZ space with $K = 3$ and (c) Nearest Neighbor (NN) in the HSV space among the points obtained at the K-NN stage.

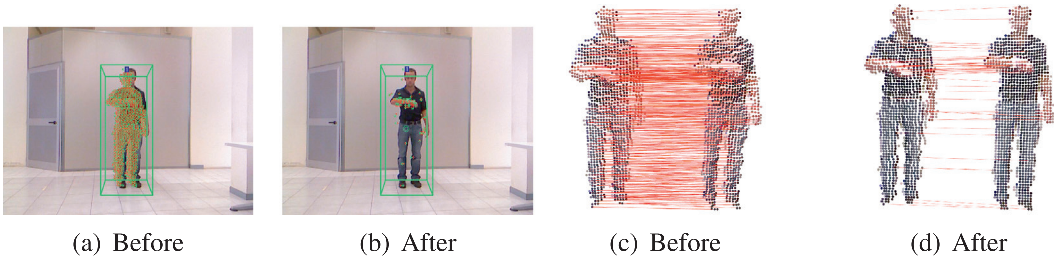

(a) Before (b) After (c) Before (d) After

Fig. 3 Example of 3D flow estimation results reprojected to the image (a and b) for action *Check watch*. Flow is visualized as green arrows in the image, before (a) and after (b) outlier removal. Also correspondences between point clouds are visualized without (c) and with (d) outlier removal.

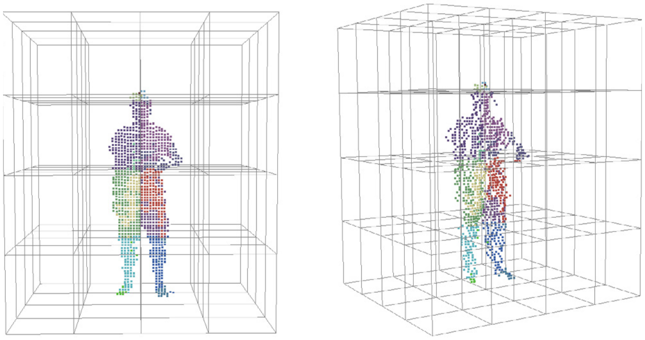


Fig. 4 Two different views of the computed 3D grid: four portions along the x, y and z axis are used.

account the temporal order in which this motion occurs. As it will be highlighted in the experiments section, we obtained the best results by choosing 30 frames for composing sequence descriptors. That is, the total dimension of these descriptors is $F * C * 3 = 30 * 64 * 3 = 5760$, where F is the number of frames extracted from a sequence, C is the number of cubes of the descriptor grid and 3 is the dimension of every cube descriptor.

6. Nearest Neighbor classification

At a training stage, we learn how to recognize actions by storing descriptors computed from labeled frames sequences containing one action each. For classifying a new sequence, we compare its sequence descriptor with those of the training set by means of the Euclidean distance and we assign to it the action label of the nearest training descriptor. It is worth noting that the whole recognition procedure remains the same if we substitute Nearest Neighbor with a Support Vector Machine classifier, which scales better when increasing the number of training examples.

7. Experiments

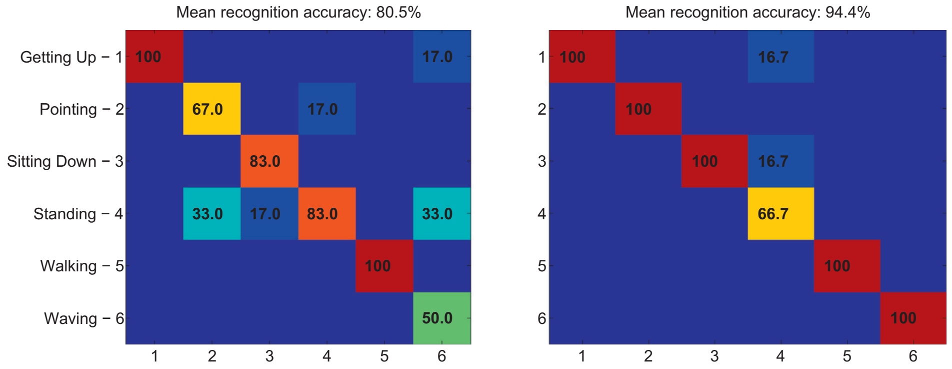
In this section, we report the human action recognition experiments we performed by exploiting the 3D motion flow estimation presented in this work. We used the people detection and tracking algorithm described in (Munaro et al., 2012) for segmenting people point clouds out from raw Kinect data. That method also performs a voxel grid filtering of the whole point cloud in order to reduce the number of points that should be handled. It is worth noting that a voxel size of 0.06 m proved to be ideal for people tracking purposes, but it was found to be insufficient for capturing local movements of the human body. For this reason, we chose it to be of 0.02 m.

7.1. Results

The first tests we report are based on the action dataset used in Ballin et al. (2012). That dataset contains six types of human actions: *getting up, pointing, sitting down, standing, walking, waving*. Each action is performed once by six different actors and recorded from the same point of view. Every action is already segmented out into a video containing only one action. Each of the segmented video samples spans from about 1—7 s.

For assigning an action label to every test sequence, we performed Nearest Neighbor classification with a *leave-one-person-out* approach, that is we trained the actions classifiers on the videos of all the persons except one and we tested on the video containing the remaining person. Then, we repeated this procedure for all the people and we computed the mean of all the rounds for obtaining the mean recognition accuracy. In Fig. 5b, we report the confusion matrix obtained on this dataset with our approach based on the *SUMFLOW* descriptor when using 10 frames evenly spaced in time for composing the sequence descriptor. The mean accuracy is 94.4% and the only errors occur for recognizing the *standing* action, which is sometimes confused with the *getting up* and *sitting down* actions. The accuracy we obtained in this work is considerably higher than that obtained in Ballin et al. (2012), which was of 80% (Fig. 5a). This improvement is due to: The outlier rejection performed after the motion flow estimation, the use of the *SUMFLOW* descriptor, which is less sensitive to noise than the *MEANFLOW* proposed in Ballin et al. (2012), and the choice of the frames which are concatenated to compose the sequence descriptor. In fact, we select frames evenly spaced within a sequence, while, in Ballin et al. (2012), the central frames of every sequence were chosen, thus encoding only the central part of an action.

The single contributions of this paper have been also evaluated on the *IAS-Lab Action Dataset*. In Fig. 6, we show an example of 3D flow estimation for some key frames of the *Throw from bottom up* action. We adopted the same



(a) (Ballin et al., 2012) (b) This work

Fig. 5 Confusion matrix obtained on the dataset presented in Ballin et al. (2012).

leave-one-person-out approach described above for computing the recognition accuracy.

In the confusion matrices shown in Fig. 7, 30 frames have been selected from every action sequence to compute the sequence descriptor. It can be noticed how the *MEANFLOW* descriptor (Fig. 7a) reaches considerably lower recognition accuracy with respect to the *SUMFLOW* (Fig. 7c), 58% against 85.2%. We also report the results obtained when the outlier rejection described in the pipeline section is not performed (Fig. 7b), thus leading to a drop of 3.3% in perfomance.

A further improvement can be obtained by performing a Principal Component Analysis (PCA) projection of the frame descriptors. The descriptor length reduced from 192 to 61 elements when retaining the 99% of information. This reduction allowed a faster comparison between descriptors and a reduction of the noise influence. We show in Fig. 7d the confusion matrix obtained when performing this PCA projection. The overall accuracy increased by 2.2%, reaching 87.4%, which can be considered as a very good score given that the people used as test set were not present in the training set. These results prove that 3D local motion is highly discriminative for the action recognition task. We can notice that most of errors occurred for the action *Point* and *Wave*, and in particular that actions with little motion can be confused with the *Standing* action. It is worth noting that, even if the *Standing* action is not included in all the datasets we reported in the datasets section, it is very important for the task of action detection: an algorithm able to reliably distinguish this action from the rest could be easily extended to detect actions from an online stream, rather than needing pre-segmented sequences.

In Fig. 8, we report the mean recognition accuracy obtainable on the *IAS-Lab Action Dataset* when using the *SUMFLOW* frame-wise descriptor and varying the number of frames used for composing the sequence descriptor. It can be noticed how the accuracy rapidly increases until five frames per sequence and continues to considerably improve until 30 frames are used. By comparing the curves with and without PCA projection, we can also observe that the accuracy obtainable without PCA and 30 frames per sequence can be obtained with PCA and half (15) of the frames, thus allowing faster comparison between sequence descriptors.

7.2. Runtime performance

In Table 2, the computation times needed by every step of our algorithm for processing one frame are shown. These timings are measured on a notebook with an Intel i7-620M 2.67 Ghz processor and 4 GB of RAM. The most demanding operation is the matching between the previous and current point clouds, that is the search for correspondences, which takes 0.015 s for initializing the octree used for the search and 0.23 s for the actual matching. On the contrary, the outlier removal and the frame descriptor computation are very fast operations. The overall runtime is then of about 0.25 s, meaning a framerate of four frames per second. The nearest neighbor classification is also rapidly performed in 0.0015 s if we consider the case where the descriptors are reduced in length by means of PCA projection.

The runtime of our algorithm is highly dependent on the number of points belonging to the person cluster cloud. In this paper, we filtered the Kinect point cloud with a voxel grid filter with voxel size of 0.02 m, thus obtaining person clusters of about 1000 points. If we use a voxel size of

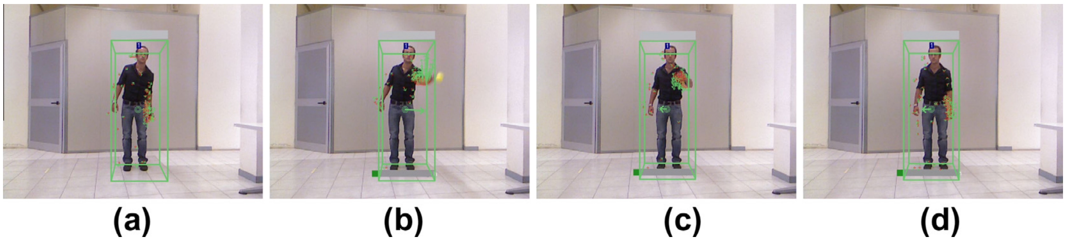
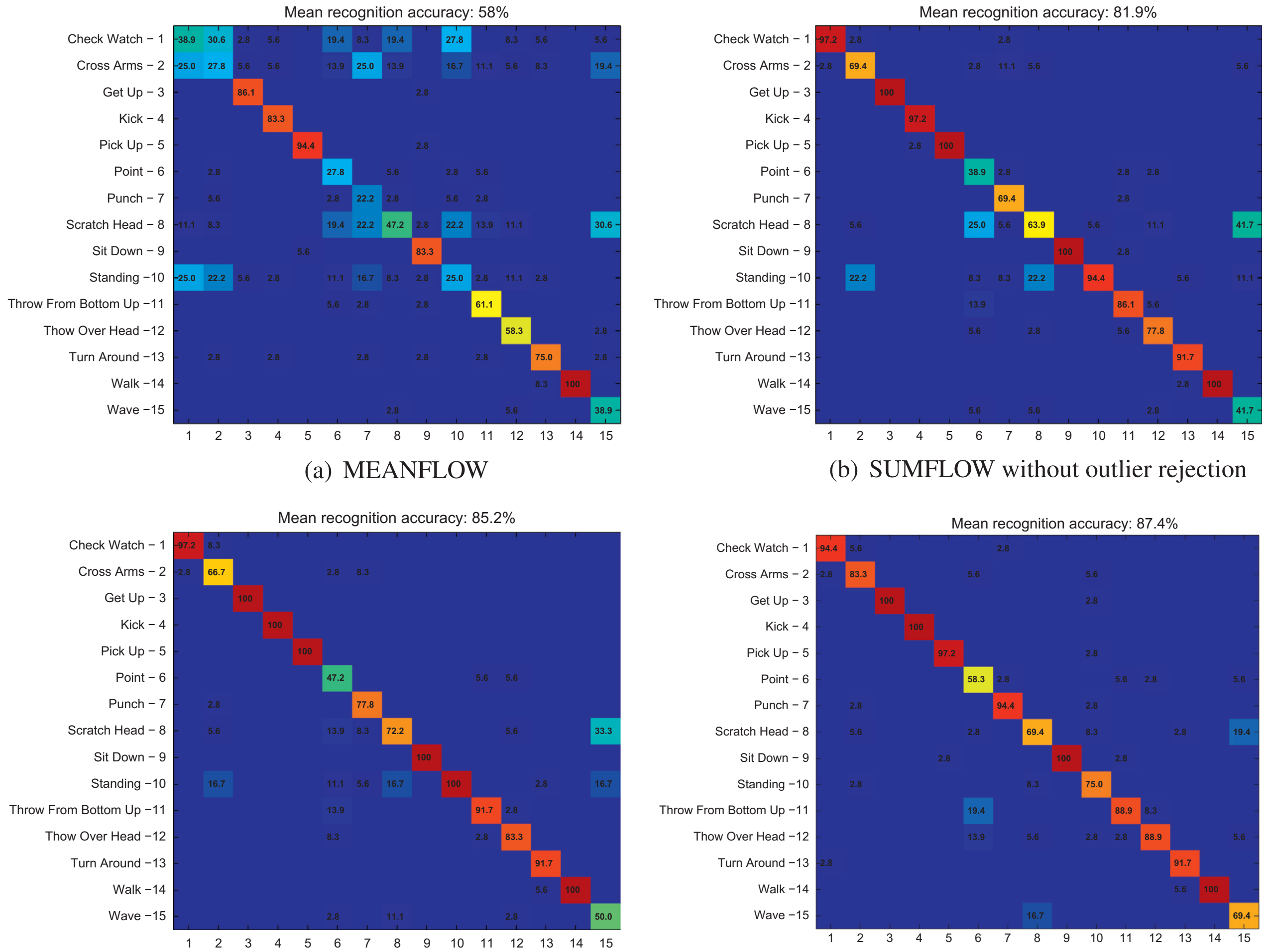


Fig. 6 Example of 3D flow estimation for some key frames of the *Throw from bottom up* action of the *IAS-Lab Action Dataset*.



(a) MEANFLOW

(b) SUMFLOW without outlier rejection

(c) SUMFLOW with outlier rejection

(d) SUMFLOW after PCA projection

Fig. 7 Confusion matrix obtained with the MEANFLOW and some variants of the SUMFLOW descriptor: (a) MEANFLOW, (b) SUMFLOW without outlier rejection, (c) SUMFLOW with outlier rejection, and (d) SUMFLOW after projection on a PCA subspace.

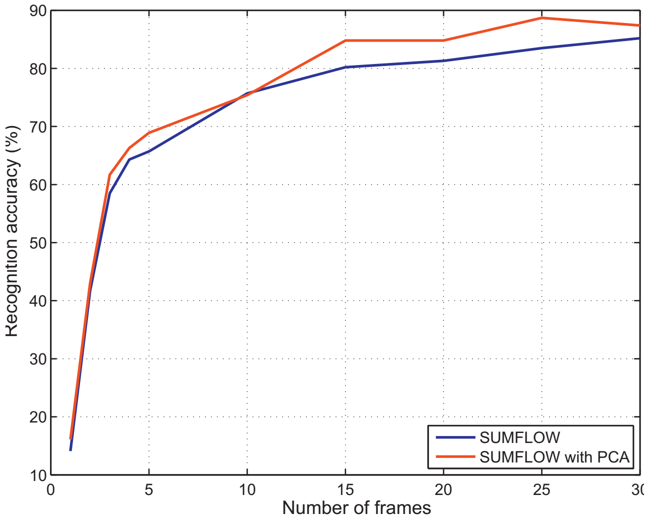


Fig. 8 Mean recognition accuracy when varying the number of frames used for composing the sequence descriptor.

Table 2 Runtime for every step of our algorithm.

Octree initialization	0.0150 s
Point clouds matching	0.2300 s
Flow vectors computation	0.0001 s
Outlier removal	0.0004 s
Descriptor computation	0.0003 s
Total	**0.2458 s**

0.06 m, the medium person cluster size is of 400 points and our whole algorithm runs at 20 fps. However, for achieving a good accuracy in finding correspondences between point clouds, the voxel size had to be reduced with respect to that chosen in Munaro et al. (2012) for people tracking purposes.

8. Conclusions

In this paper, we presented a novel method for real-time estimation of 3D motion flow from colored point clouds and a complete system for human action recognition which exploits this motion information. In particular, we also proposed two 3D grid-based descriptors as frame-wise motion descriptors and a sequence descriptor which is then classified with the Nearest Neighbor classifier for performing action recognition. We also presented a new dataset with high variability in the number of people performing the actions and providing both RGB-D data and skeleton pose for every frame. The tested 3D flow technique reported very good results in classifying all the actions of the dataset, reaching 87.4% of accuracy.

As future works, we plan to test new types of descriptors which encode the motion inside each grid with an histogram and exploit SVM classifiers in place of Nearest Neighbor. We will also implement an automatic method for online segmentation of action sequences. Moreover, we will exploit skeleton information or other person orientation estimation methods for making our motion-based action recognition system be invariant to person orientation.

References

Ali, S., & Shah, M. (2010). Human action recognition in videos using kinematic features and multiple instance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 32*, 288–303. http://dx.doi.org/10.1109/TPAMI.2008.284.

[Online], M.K. (a). *Microsoft kinect for windows* <http://www.microsoft.com/en-us/kinectforwindows/>.

[Online], N.M. (b). *Nite middleware* <http://www.primesense.com/solutions/nite-middleware>.

Ballin, G., Munaro, M., & Menegatti, E. (2012). Human action recognition from RGB-D frames based on real-time 3D optical flow estimation. In A. Chella, R. Pirrone, R. Sorbello, & K. R. Jóhannsdóttir (Eds.), *Biologically Inspired Cognitive Architectures 2012. Advances in Intelligent Systems and Computing* (vol. 196, pp. 65–74). Berlin Heidelberg: Springer.

Basso, F., Munaro, M., Michieletto, S., & Menegatti, E. (2012). Fast and robust multi-people tracking from rgb-d data for a mobile robot. In *Proceedings of the 12th intelligent autonomous systems (IAS) Conference*. Jeju Island (Korea).

Blank, M., Gorelick, L., Shechtman, E., Irani, M., & Basri, R. (2005). Actions as space-time shapes. In *Proceedings of the Tenth IEEE international conference on computer vision ICCV 2005* (Vol. 2, pp. 1395–1402). doi:http://dx.doi.org/10.1109/ICCV.2005.28.

Bloom, V., Makris, D., & Argyriou, V. (2012). G3d: A gaming action dataset and real time action recognition evaluation framework. In *IEEE computer society conference on computer vision and pattern recognition workshops (CVPRW)* (pp. 7–12).

Carlsson, S., & Sullivan, J. (2001). Action recognition by shape matching to key frames. In *IEEE computer society workshop on models versus exemplars in computer vision*.

Dollar, P., Rabaud, V., Cottrell, G., & Belongie, S. (2005). Behavior recognition via sparse spatio-temporal features. In *Proceeding of the 2nd joint IEEE international visual surveillance and performance evaluation of tracking and surveillance workshop* (pp. 65–72). doi:http://dx.doi.org/10.1109/VSPETS.2005.1570899.

Efros, A., Berg, A., Mori, G., & Malik, J. (2003). Recognizing action at a distance. In *Ninth IEEE international conference on proceedings of computer vision, 2003* (Vol. 2, pp. 726–733) doi:http://dx.doi.org/10.1109/ICCV.2003.1238420.

Holte, M., & Moeslund, T. (2008). View invariant gesture recognition using 3d motion primitives. In *IEEE international conference on acoustics, speech and signal processing, 2008. ICASSP 2008* (pp. 797–800) doi:http://dx.doi.org/10.1109/ICASSP.2008.4517730.

Holte, M., Moeslund, T., Nikolaidis, N., & Pitas, I. (2011). 3d human action recognition for multi-view camera systems. In 3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), 2011 International Conference on (pp. 342 –349). doi:10.1109/3DIMPVT.2011.50.

Jiang Wang, Y.W., Zicheng Liu, & Yuan, J. (2012). Mining actionlet ensemble for action recognition with depth cameras. In *IEEE conference on computer vision and pattern recognition (CVPR 2012)*. Providence, Rhode Island.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Attention, Perception, & Psychophysics, 14*, 201–211, doi:http://dx.doi.org/10.3758/BF03212378.

Ke, Y., Sukthankar, R., & Hebert, M. (2005). Efficient visual event detection using volumetric features. In *Proceedings of the tenth IEEE international conference on computer vision ICCV 2005* (Vol. 1, pp. 166–173), doi:http://dx.doi.org/10.1109/ICCV.2005.85.

Kläser, A., Marszałek, M., & Schmid, C. (2008). A spatio-temporal descriptor based on 3d-gradients. In *British machine vision conference* (pp. 995–1004) <http://lear.inrialpes.fr/pubs/2008/KMS08>.

Korner, M., & Denzler, J. (2012). Analyzing the subspaces obtained by dimensionality reduction for human action recognition from 3d data. In *IEEE ninth international conference on advanced video and signal-based surveillance (AVSS), 2012* (pp. 130–135).

Laptev, I., & Lindeberg, T. (2003). Space-time interest points. In *Proceedings of the ninth IEEE international computer vision conference* (pp. 432–439), doi:http://dx.doi.org/10.1109/ICCV.2003.1238378.

Laptev, I., Caputo, B., Schüldt, C., & Lindeberg, T. (2007). Local velocity-adapted motion events for spatio-temporal recognition. *Computer Vision and Image Understanding, 108*, 207–229.

Laptev, I., Marszalek, M., Schmid, C., & Rozenfeld, B. (2008). Learning realistic human actions from movies. In *Proceedings of the IEEE conference computer vision and pattern recognition CVPR 2008* (pp. 1–8), doi:http://dx.doi.org/10.1109/CVPR.2008.4587756.

Lei, J., Ren, X., & Fox, D. (2012). Fine-grained kitchen activity recognition using rgb-d. In *Proceedings of the 2012 ACM conference on ubiquitous computing UbiComp '12* (pp. 208–211). New York, NY, USA: ACM.

Li, W., Zhang, Z., & Liu, Z. (2010). Action recognition based on a bag of 3d points. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on (pp. 9 –14). doi:10.1109/CVPRW.2010.5543273.

Liu, J., Ali, S., & Shah, M. (2008). Recognizing human actions using multiple features. In *IEEE conference on computer vision and pattern recognition, 2008. CVPR 2008* (pp. 1–8), doi:http://dx.doi.org/10.1109/CVPR.2008.4587527.

Liu, R., Li, S.Z., Yuan, X., & He, R. (2008). Online determination of track loss using template inverse matching. In *The eighth international workshop on visual surveillance – VS2008*. Marseille, France: Graeme Jones and Tieniu Tan and Steve Maybank and Dimitrios Makris <http://hal.inria.fr/inria-00325656>.

Lukas, B., & Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. In *IJCAI '81* (pp. 674–679).

Ming, Y., Ruan, Q., & Hauptmann, A. (2012). Activity recognition from rgb-d camera with 3d local spatio-temporal features. In *IEEE international conference on multimedia and expo (ICME), 2012* (pp. 344–349), doi:http://dx.doi.org/10.1109/ICME.2012.8.

Munaro, M., Basso, F., & Menegatti, E. (2012). Tracking people withing groups with rgb-d data. In *Proceedings of the international conference on intelligent robots and systems (IROS)*. Vilamoura (Portugal).

Niebles, J. C., Wang, H., & Fei-Fei, L. (2008). Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision, 79*, 299–318.

http://dx.doi.org/10.1007/s11263-007-0122-4, <http://portal.acm.org/citation.cfm?id=1380728.1380729>.

Ni, B., Wang, G., & Moulin, P. (2011). Rgbd-hudaact: a color-depth video database for human daily activity recognition. In *IEEE international conference on computer vision workshops (ICCV Workshops), 2011* (pp. 1147–1153), doi:http://dx.doi.org/10.1109/ICCVW.2011.6130379.

Ofli, F., Chaudhry, R., Kurillo, G., Vidal, R., & Bajcsy, R. (2012). Sequence of the most informative joints (smij): A new representation for human skeletal action recognition. In *IEEE computer society conference on computer vision and pattern recognition workshops (CVPRW), 2012* (pp. 8–13).

Ofli, F., Chaudhry, R., Kurillo, G., Vidal, R., & Bajcsy, R. (2013). Berkeley MHAD: A comprehensive multimodal human action database. In *Proceedings of WACV*.

Popa, M., Koc, A., Rothkrantz, L., Shan, C., & Wiggers, P. (2011). Kinect sensing of shopping related actions. In Undefined, K. Van Laerhoven, & J. Gelissen (Eds.), *Constructing ambient intelligence: AmI 2011 workshops*. Amsterdam, Netherlands.

Poppe, R. (2010). A survey on vision-based human action recognition. *Image Vision Computation, 28*, 976–990.

Quigley, M., Gerkey, B., Conley, K., Faust, J., Foote, T., Leibs, J., Berger, E., Wheeler, R., & Ng, A. (2009). Ros: an open-source robot operating system. In *Proceedings of the IEEE international conference on robotics and automation (ICRA)*.

Rusu, R. B., Bandouch, J., Meier, F., Essa, I. A., & Beetz, M. (2009). Human action recognition using global point feature histograms and action shapes. *Advanced Robotics, 23*, 1873–1908.

Schuldt, C., Laptev, I., & Caputo, B. (2004). Recognizing human actions: a local svm approach. In *Proceedings of the 17th international conference on pattern recognition ICPR 2004* (Vol. 3, pp. 32–36), doi:http://dx.doi.org/10.1109/ICPR.2004.1334462.

Scovanner, P., Ali, S., & Shah, M. (2007). A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the 15th international conference on Multimedia MULTIMEDIA '07* (pp. 357–360). New York, NY, USA: ACM, http://doi.acm.org/10.1145/1291233.1291311.

Sung, J., Ponce, C., Selman, B., & Saxena, A. (2011). Human activity detection from rgbd images. In *Plan, activity, and intent recognition. AAAI volume WS-11-16 of AAAI Workshops*.

Sung, J., Ponce, C., Selman, B., & Saxena, A. (2012). Unstructured human activity detection from rgbd images. In *International conference on robotics and automation, ICRA*.

Thuc, H.L.U., Tuan, P.V., & Hwang, J.-N. (2012). An effective 3d geometric relational feature descriptor for human action recognition. In *IEEE RIVF international conference on computing and communication technologies, research, innovation, and vision for the future (RIVF), 2012* (pp. 1–6).

Wanqing Li, Z.L., Zhengyou Zhang (2010). Action recognition based on a bag of 3d points. In *IEEE international workshop on CVPR for human communicative behavior analysis (in conjunction with CVPR 2010)*. San Francisco, CA.

Weinland, D., Ronfard, R., & Boyer, E. (2006). Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding, 104*, 249–257. http://dx.doi.org/10.1016/j.cviu.2006.07.013, http://dx.doi.org/10.1016/j.cviu.2006.07.013.

Wolf, C., Mille, J., Lombardi, E., Celiktutan, O., Jiu, M., Baccouche, M., et al. (2012). *The LIRIS Human activities dataset and the ICPR 2012 human activities recognition and localization competition*. Technical Report.

Xia, L., Chen, C.-C., & Aggarwal, J. (2012). View invariant human action recognition using histograms of 3d joints. In *IEEE computer society conference on computer vision and pattern recognition workshops (CVPRW), 2012* (pp. 20–27).

Yacoob, Y., & Black, M. (1998). Parameterized modeling and recognition of activities. In *Sixth international conference on computer vision, 1998* (pp. 120–127), doi:http://dx.doi.org/10.1109/ICCV.1998.710709.

Yang, X., & Tian, Y. (2012). Eigenjoints-based action recognition using Naive–Bayes-nearest-neighbor. In *IEEE workshop on CVPR for human activity understanding from 3D data*.

Yilmaz, A., & Shah, M. (2005). Actions sketch: a novel action representation. In *IEEE computer society conference on computer vision and pattern recognition, 2005. CVPR 2005* (Vol. 1, pp. 984–989), doi:http://dx.doi.org/10.1109/CVPR.2005.58.

Zhang, H., & Parker, L.E. (2011). 4-dimensional local spatio-temporal features for human activity recognition. In *IEEE/RSJ international conference on intelligent robots and systems (IROS), 2011* (pp. 2044–2049), doi:http://dx.doi.org/10.1109/IROS.2011.6094489.

Zhao, Y., Liu, Z., Yang, L., & Cheng, H. (2012). Combing rgb and depth map features for human activity recognition. In *Signal information processing association annual summit and conference (APSIPA ASC), 2012 Asia–Pacific* (pp. 1–4).