Contents lists available at ScienceDirect

Pattern Recognition

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

Human activity recognition using multi-features and multiple kernel learning

Salah Althloothi [a], Mohammad H. Mahoor [a,*], Xiao Zhang [a], Richard M. Voyles [b]

[a] Department of Electrical and Computer Engineering, University of Denver, CO 80208, USA
[b] College of Technology, Purdue University, West Lafayette, IN 47907, USA

ARTICLE INFO

Article history:
Received 17 April 2013
Received in revised form
8 November 2013
Accepted 30 November 2013
Available online 17 December 2013

Keywords:
Human activity recognition
Multiple kernel learning
Spherical harmonics coefficients
Distal limb segments
Support vector machines

ABSTRACT

This paper presents two sets of features, shape representation and kinematic structure, for human activity recognition using a sequence of RGB-D images. The shape features are extracted using the depth information in the frequency domain via spherical harmonics representation. The other features include the motion of the 3D joint positions (i.e. the end points of the distal limb segments) in the human body. Both sets of features are fused using the Multiple Kernel Learning (MKL) technique at the kernel level for human activity recognition. Our experiments on three publicly available datasets demonstrate that the proposed features are robust for human activity recognition and particularly when there are similarities among the actions.

1. Introduction

Human activity recognition has remained as an interesting and challenging topic in the field of computer vision and pattern recognition. This research topic is motivated by many applications such as surveillance systems, video browsing, and human–computer interfaces (HCI) design. In the past two decades, a significant amount of research has been done in the area of human activity recognition using a sequence of 2D images. Most published research is based on either shape features or motion features. Recently, researchers have paid more attention to using 3D spatio-temporal features for describing and recognizing human activities [1–5] due to easy access to depth information via new consumer technologies such as Microsoft's Kinect sensor.

In general, 3D spatio-temporal features look at the changes in the human body shape based on dominant motions in the human limbs [1,3]. The variations in the body shape can be detected and represented with 3D spatio-temporal features as space-time volumes. Those features mainly focus on the representation of the shape and motion as a function of time. The main idea behind the methods that utilize the spatio-temporal features is to recognize human activity by detecting/describing the changes in human limbs either by describing the motion of human limbs or through measuring the similarities among different space-time volumes.

Recently, the developed commodity depth sensors such as Kinect [6] have opened up new possibilities of dealing with 3D data. The Kinect sensor has given the computer vision community the opportunity to acquire RGB images as well as depth maps simultaneously at a good frame rate with a good resolution. As we can see in Fig. 1, the depth map provides additional information as 3D data which is expected to be helpful in distinguishing different poses of silhouettes. Furthermore, compared with RGB images, the depth map increases the amount of information that can be used to detect 3D joint positions.

The research in human activity recognition based on a sequence of depth maps has been motivated with the release of the Kinect Windows SDK, which is utilized to estimate the 3D joint positions of the human body. Although Kinect produces better quality 3D motion than those estimated from regular RGB sensors (e.g., Stereo vision systems for 3D estimation), the estimated 3D joint positions are still noisy and fail when there are occlusions among human limbs such as two limbs crossing each other. Furthermore, the motion of 3D joint positions alone is insufficient to distinguish similar activities such as eating and drinking. Therefore, extra information needs to be included in the feature level to enhance the classification performance. In this context, we need to develop a method to fuse multiple types of features in order to discriminate similar activities and to enhance the recognition rate of the system. For instance, the motion features of human limbs, such as forearms and shins, may be augmented with the shape features that describe the silhouette structure to improve the accuracy of the action classification.

Consequently, fusion techniques can be used to enhance the classification performance of human activity recognition. In this

* Corresponding author.
E-mail addresses: salthloo@du.edu (S. Althloothi), mmahoor@du.edu (M.H. Mahoor), xzhang62@du.edu (X. Zhang), rvoyles@purdue.edu (R.M. Voyles).

0031-3203/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.11.032

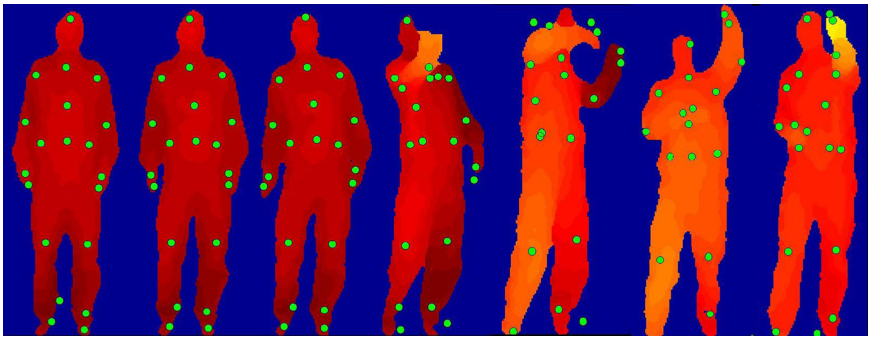

Fig. 1. Sample of depth maps with 3D joint positions for Tennis Serve action.

context, some researchers have conducted fusion at the feature level such as [7–9] where they combined multiple types of features extracted from 2D images into a fused feature vector and then used a single classifier for action recognition. In particular, Liu et al. [7] and Wang et al. [8] combined spatio-temporal volume features and a deformation of the human silhouette obtained from sequences of 2D images to derive action descriptors. In another work, Liu et al. [9] fused local spatio-temporal volumes and statistical models of interest points (Cuboids and 2D SIFT) obtained from 2D images for action recognition using hypersphere multi-class Support Vector Machines (SVM). Fusion can also be performed at the classifier level. For instance [10,11] designed multiple classifiers for two types of features extracted from 2D images, and their final decision was made by taking into account the complementaries among classifiers. In our work, two sets of features are extracted from a depth map (3D data) and are fused at the kernel level instead of the feature level in order to select useful features based on the weights using the MKL technique.

Recently, MKL techniques [12,47] have been proposed for feature fusion within kernel-based classifiers. The works presented in [13–15] show that the MKL technique can enhance the discrimination power and improve the performance of classifiers. The idea behind MKL is to optimally combine different kernel matrices calculated from multiple types of features with multiple kernel functions. Within this framework, the problem of multi-feature representation with a single kernel function in the canonical SVM is transferred to set the optimal value of kernel combination weights for multiple kernel matrices. These works empirically show that the MKL-based multiclass SVM outperforms the canonical multiclass SVM.

This paper presents a method to recognize human activities using a sequence of RGB-D data. The basic idea of our method is illustrated in Fig. 2. Based on the surface representation and the kinematic structure of the human body, we propose a method that can characterize shapes and motions. In our approach the shape features, extracted from the depth map using spherical harmonics representation, are used to describe the 3D silhouette structure. The motion features, extracted from the estimated 3D joint positions, are used to describe the movement of the human body. The distal limb segments of the human body are utilized in our method to describe the motion because we believe that segments such as forearms and shins provide sufficient and compact information for human activity recognition. Therefore, each distal limb segment is described by the orientation and translation distance with respect to the initial frame in order to create motion features. Both sets of features are fused using the MKL technique [16] to produce an optimally combined kernel matrix within SVM for activity classification. This kernel matrix has more discriminating power than a single kernel function due to the utility of multiple features within different kernel functions.

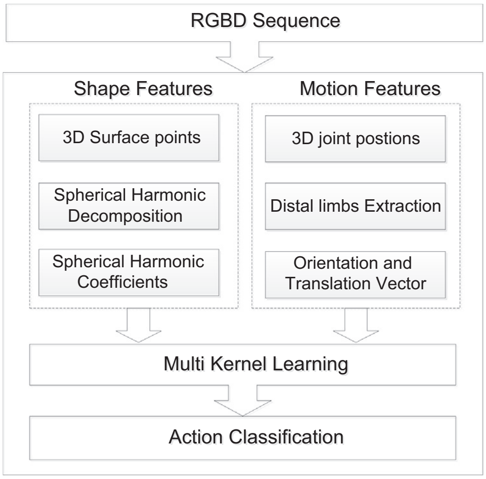

Fig. 2. Our proposed method for human activity recognition using multiple types of features.

Compared with the aforementioned 2D-based approaches for multi-feature fusion, our work is based on features extracted from 3D data (depth map). Also, our approach is based on multiple kernel functions and multiple features which have more advantages over single kernel function with multiple features. In fact, a single kernel cannot perform well when the nature of the features are different and incompatible. Furthermore, combining multiple features into one feature vector introduces the curse of dimensionality problem.

In summary, the contributions of this paper are summarized as follows: (1) A novel 3D shape feature using spherical harmonics transformation to represent the body silhouette is proposed. (2) The human body motions (i.e. kinematic structure) are described using only the distal limb segments. (3) These two types of features are fused at the kernel level as a novel methodology in order to differentiate similar activities and enhance the classification rate of the system.

The remainder of this paper is organized as follows. A brief review of related work is presented in Section 2. Section 3 explains our proposed frame work for activity recognition using 3D spatio-temporal features and feature fusion using the simple MKL approach. Our experimental results are presented in Section 4. Finally, Section 5 concludes the paper.

2. Related work

In the last decade, the shape-based methods using 2D images (captured by a regular RGB camera) have been widely used for action recognition. There are several different shape-based methods which are based on silhouettes for statistical shape analysis as a function of time. The shape analysis approaches aim to describe and locate the changes in the human body shape. For instance, Blank et al. [17] analyzed 3D shapes of silhouettes in a space-time volume to create 3D spatio-temporal features for human activity recognition. Cohen and Li [18] presented a 3D spatio-temporal feature for classifying and identifying human posture using SVM. They proposed global shape features that are invariant to rotation and translation. The main advantage of their approach is in its ability to capture human shape variations – allowing for the identification of body postures – but it needs to use multiple cameras to create 3D spatio-temporal features. Another example is the work of Yilmaz and Shah [19] which represented the action as a spatio-temporal feature defined on structure of contour across successive video frames. The action features are described by analyzing the differential geometry properties of the surface volume. The limitation of this method is the exhaustive search required for finding the correspondences between two volumes.

The motion-based methods, which are based on motion trajectory, recognize the action by employing either human limb positions or interest points on the human body. For instance, Chen et al. [20] and Fujiyoshi and Lipton [21] proposed a star skeleton representation based on the shape geometry to recognize human action. They detected the extremities of the silhouette with respect to the centroid and assumed that these points represent the head, hands, and feet. Chun et al. [22] proposed a 3D star spatio-temporal pattern based on the shape boundary information of the human posture using eight projection maps from different views. They detected the extremities of the silhouette with respect to the centroid as a shape feature. They assumed that these extremities represent the head, hands, and feet. They accumulated the motion history of these features in order to create a spatio-temporal pattern. Although the star shape representation is simple and fast for computation, its accuracy for detecting limbs needs further improvement. In the method proposed by Althloothi et al. [23,24], skeleton models were fitted to silhouettes to capture the positions of distal limb segments (i.e. arms, legs and head). The positions of distal limb segments were used as features. Then, Gaussian mixture models were used to model the spatio-temporal distribution of the distal limb segments over the period of an action. Sun et al. [25] extracted trajectories through pairwise Scale Invariant Feature Transform (SIFT) correspondences between two consecutive frames. The stationary distribution of a Markov chain model was then used to compute motion features. Tian et al. [26] employed the Harris detector and local HOG descriptor on Motion History Images (MHI) to perform action recognition and detection. The fundamental difference between the aforementioned and recent methods described in the next paragraphs is that all the extracted features are captured by regular RGB cameras, instead of an RGB-D sensor which represents the depth map of the human body shape.

Recently, with the release of the Microsoft Kinect sensor, research on human activity recognition based on a sequence of depth maps and kinematic structure has resurged. Wang et al. [1] proposed a model for human actions, called the Actionlet Ensemble Model which is learned using the MKL technique to represent each action and to capture the intra-class variance. Wang et al. [1] added a temporal pyramid to capture the temporal in order to improve the accuracy. Xia et al. [3] developed a method based on 3D joint positions estimated from a depth map to create a histogram of 3D joint positions using a spherical coordinate system. Then, they modeled the temporal evolutions of 3D joint positions by discrete hidden Markov models in order to train/classify the action. Li et al. [2] proposed a Bag-of-3D-Points model for action recognition. They first sampled 2D points at equal distances along the contours of projections formed by mapping the depth map onto three orthogonal Cartesian planes, i.e. XY, XZ, and YZ planes. Then, the sampled 2D points were used to characterize the posture in each frame.

Similarly, Yang et al. [4] generated the Depth Motion Maps (DMM) from three orthogonal planes and accumulated global activities through entire video sequences. Then, Histograms of Oriented Gradients (HOG) are computed from the DMM as the representation of an action model. Yang et al. [4] collapsed the whole sequence of frames into one image DMM, which eliminates the temporal order of shape/motion cues. Sung et al. [27] compute a set of features based on human pose and motion of the 3D joint positions provided by Prime Sense with Kinect. They proposed a hierarchical maximum entropy Markov model (MEMM), which considers the human activity as composed of two-layered graph structure.

Yang and Tian [5] proposed a method based on position differences of 3D joint positions. They applied Principal Component Analysis (PCA) to joint differences to obtain EigenJoints by reducing redundancy and noise. Then, they employed the Naive Bayes Nearest Neighbor (NBNN) classifier for multi-class action classification. Zhang and Parker [28] proposed a 4D local spatial-temporal descriptor that combines both intensity and depth information. The proposed descriptor computes and concatenates the intensity and depth gradients within a 4D hyper cuboid, which is centered at the detected feature point, as a feature. Zhao et al. [29] extract the feature vector of each video clip by combining the RGB-based descriptor and depth-map based descriptor. They used three types of features; Local Depth Pattern (LDP), which describes the local region of interest points in the depth map, HOG and HOF features in RGB data. All these features are concatenated in one vector in order to classify human action as multi-class classification using LibSVM [30]. Omar et al. [31] presented the HON4D approach that can describe a sequence of depth maps using a histogram capturing the distribution of the surface normal orientation in a 4D space of time, depth, and spatial coordinates. This process needs to create 4D projectors, which quantize the 4D space and represent the possible directions for the 4D normal. The limitation of the HON4D approach is the quantization process used to build the histogram. Holte et al. [32] proposed an approach for view-invariant gesture recognition. They focused only on arm gestures by segmenting the arms (when they move) using optical flow to represent the motion context which represents the velocity of the arms by using the location of motion, the amount of motion, and its direction. Also, they utilized spherical harmonics representation to make the motion context to be rotation invariant through the use of a depth map. Compared to our method, they did not use spherical harmonics as shape features for human action recognition.

From the literature review, we have observed that the motion and shape features alone have their own limitations in representing human activities. The motion features are not robust in capturing the change in velocity among the frames. While the shape features can capture some pose information of the human body, but without motion features the capability of describing human activity is limited. Also, the depth map captured from an RGB-D sensor contains rich shape information of the human body compared with the RGB images. In addition, the distal limb segments such as forearms and shins provide sufficient and compact features for human action recognition. The movements of distal limb segments are important cues for estimating the 3D motion of the human limbs. These observations form the core concept of the method proposed in this paper. Therefore, the main idea of our work is to fuse the motion features of the distal limb segments with the shape features extracted from the depth map in a novel way to enhance the classification performance.

3. Proposed method

In our work, we utilized spherical harmonics representation to extract 3D shape features of the silhouette and the kinematic structure of the human body to extract the 3D motion features. These features are fused using the SimpleMKL algorithm [16] with different kernel functions where both the kernel combination weights and discriminate hyper planes of multiclass-SVMs are optimized. In the end, human action classification based on the learned kernel weights and discriminated hyper planes of multiclass-SVM are used to address the issues of recognizing complex activities. The general framework of our proposed method is demonstrated in Fig. 2.

3.1. *Spatio-temporal features*

This section gives a detailed description of the two proposed 3D spatio-temporal features for human activity representation and classification.

3.1.1. *Motion features*

The distal limb segments are employed to extract the motion features of the human movement. In fact, the positions of the distal limb segments (four limbs) provided by a Kinect sensor, as illustrated in Fig. 3, are utilized to extract the motion features which represent the orientation and the translation distance of the distal limb segments. Our key observation is that the change in the positions of the distal limb segments provide sufficient information to represent the human body movement as discriminative features to classify the actions.

In our approach, the end points of the distal limb segments are used to characterize the motion features. Thus, each distal limb segment L_k is described by its end points as a 3D vector with respect to the body center (hip) in order to create the 3D unit vector which represents the orientation of the distal limb segment. The orientation of the 3D vector is defined by a unit vector $\overrightarrow{U}_{ab}$ for frame F_t. Therefore, the 3D coordinates (x_i, y_i, z_i) of end points $J(t)$ of distal segments L_k in frame F_t is $J_i^{L_k}(t) = \{j_a, j_b\}$, where k is a distal limb number and N is the number of distal limb segments.

$$\overrightarrow{V}_{ab}(t) = \{J_b^{L_k}(t) - J_a^{L_k}(t) \mid k = 1, 2, \ldots, N; a \neq b\} \tag{1}$$

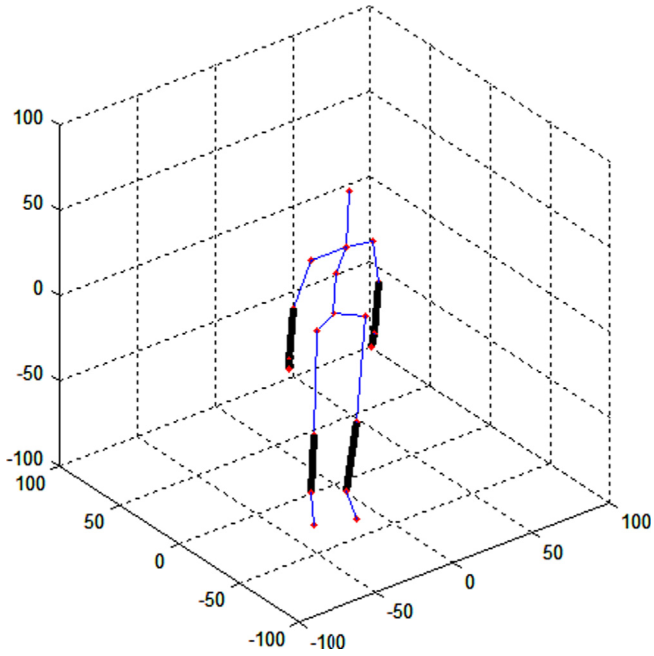

Fig. 3. The 3D distal limb segments (four limbs) are used to extract the motion features.

$$\overrightarrow{V}_{ab}(t) = (x_b - x_a)\hat{i} + (y_b - y_a)\hat{j} + (z_b - z_a)\hat{k} \tag{2}$$

$$\overrightarrow{U}_{ab}(t) = \frac{\overrightarrow{V}_{ab}(t)}{\| \overrightarrow{V}_{ab}(t) \|} = A_x\hat{i} + A_y\hat{j} + A_z\hat{k} \tag{3}$$

Since all the measurements are relative to the center of the body in one frame, it is necessary to add more information about the motion of the human limbs between the current frame F_t and the initial frame F_{t_0}, which represents to the neutral human pose. Thus, the translation, which gives the difference in position for the distal limb segments between two frames, is computed in order to create a 3D spatio-temporal feature and to make the classifier discriminate between the actions that have similar orientation but different positions. Practically, each human subject has four distal limb segments which are tracked by the skeleton tracker [33], each distal limb L_k is represented by 3D unit vectors $\overrightarrow{U}_{ab}$ and translation vectors D_{tt_0} in each frame.

$$D_{tt_0} = \{ \| J_a^t - J_a^{t_0} \| \mid J_a^t \varepsilon F_t ; J_a^{t_0} \varepsilon F_{t_0} \} \tag{4}$$

The 3D unit vector $\overrightarrow{U}_{ab}$ and translation vector D_{tt_0} with respect to the initial frame F_{t_0} represent motion features in one frame. In order to represent the motion features that include t number of frames, our motion features for all the frames are concatenated together to build a spatio-temporal feature vector for motion features. This yields to a 3-D spatio-temporal feature with precise orientation and translation data for each human limb in all the frames. These motion features define the movement of distal limbs in the video. In addition, all of the vectors which represent the 3D distal limb segments were normalized to reduce intra-class variations among subjects and to be invariant to the body size. We use a linear normalization scheme to scale the features in the range $[-1$ to $+1]$.

3.1.2. *Shape features*

Due to the motion similarity in some activities, it is insufficient to only use the motion features to fully describe and model an activity. We must create another feature that can describe the human body shape in order to increase the accuracy of the classifier. We use spherical harmonics coefficients [34] as shape features for representation and recognition of human actions from a sequence of depth maps. Over the last decade, spherical harmonics coefficients have been applied to several computer vision applications such as 3-D shape descriptors [35] as well as 3-D model retrieval [36], medical image analysis [34] and rotation estimation [37]. According to [34,38], spherical harmonics coefficients are suitable for shape comparison because it can deal with protrusions and intrusions. Also, it can be used as an abstract features that can characterize the 3D object with different resolution depending on the spherical harmonics band. Consequently, a 3D surface object with thousands of vertices can be represented using spherical harmonics coefficients up to a user specified maximum band L_{max}.

In our approach, the body silhouette is described by a large number of 3D geometric cloud points obtained from the depth map. These points are down-sampled to reduce the number of mesh points and to pare down the number of vertices that can represent the human body. Spherical harmonics decomposition is then applied to these vertices and a set of spherical harmonics coefficients are extracted. These coefficients are used to build spatio-temporal features that describe the human body shape in the spherical harmonics domain.

Spherical harmonics decomposition was originally used as a type of parametric surface representation for radial surfaces and later extended to more general shapes by representing the shape surface using a spherical function $S(\theta, \phi) = (S_x(\theta, \phi), S_y(\theta, \phi),$

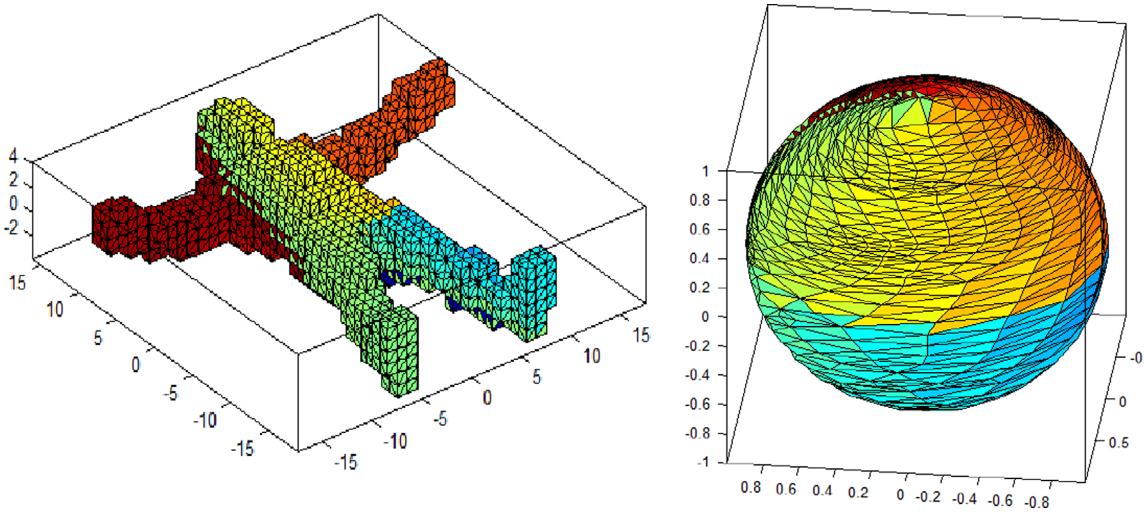

Fig. 4. Example of the voxel surface and its spherical parametrization.

$S_z(\theta,\phi)$) with θ as the polar angle and ϕ as the azimuth angle [39]. In the spherical harmonics decomposition a spherical parametrization algorithm [40] is first used to establish a one-to-one mapping between 3D points on a human body surface (vertices) and 3D points on the unit sphere. The result of the spherical parametrization process is a bijective mapping between each 3D point on a human body surface and a pair of spherical coordinates, θ and ϕ. Therefore, the human body surface is represented as a spherical function: $S(\theta,\phi)$ which specifies the distance from a specified origin to each point on the sphere surface for all three coordinates $x=R\cos\phi\sin\theta, y=R\sin\phi\sin\theta$, and $z=R\cos\theta$. Fig. 4 shows a voxel representation of the human body and its mapping onto a unit sphere.

After spherical parametrization, the spherical harmonic expansion [34] can be used to expand the human surface represented as a spherical function into a complete set of spherical harmonics coefficients. Spherical harmonics expansion is essentially a Fourier transform technique that defines a 3D surface using a spherical function and transforms them into a set of spherical harmonics coefficients $(c_{l_x}^m, c_{l_y}^m, c_{l_z}^m)$ in the frequency domain. These coefficients can be calculated up to a user-desired band L_{max} by solving a system of linear equations.

Thus, given spherical function $S(\theta,\phi)$ defined on the surface of the unit sphere, we can transform it into a set of spherical harmonics coefficients $(c_l^m = c_{l_x}^m, c_{l_y}^m, c_{l_z}^m)$ using the following equations:

$$S(\theta,\phi) = \sum_{l=0}^{L_{max}} \sum_{m=-l}^{l} c_l^m Y_l^m(\theta,\phi) \tag{5}$$

where Y_l^m denotes the orthogonal spherical harmonics base of degree l and order m, and L_{max} is the band-width of the spherical harmonics, l and m are integers with $0 \le l < L_{max}$ and $|m| \le l$:

$$Y_l^m(\theta,\phi) = \sqrt[2]{\left[\frac{(2l+1)(l-m)!}{(4\pi(l+m)!)}\right]} P_l^m(\cos\theta) e^{(im\phi)} \tag{6}$$

$P_l^m(x)$ is the associated Legendre polynomial defined by the differential equation:

$$P_l^m(x) = \frac{(-1)^m}{(2^l l!)}(1+x^2)^{m/2} \frac{d^{l+m}}{dx^{l+m}}(x^2-1)^l \tag{7}$$

Then, the spherical harmonics coefficients $(c_{lx}^m, c_{l_y}^m, c_{l_z}^m)$, which are related to $S(\theta,\phi)$ can be independently decomposed in terms (S_x, S_y, S_z) of the spherical harmonics.

Clearly, calculating the spherical harmonics coefficients linearly depends on the spherical harmonics band L_{max} and number of surface points. In particular, these coefficients are 3D vectors which are computed up to a user specified maximum band to construct the 3D shape features. Therefore, based on spherical harmonics decomposition, we can extract our 3D shape features with different frequency harmonics. These 3D shape features define the body shape in one frame. In order to represent an action that includes t number of frames, our 3D spatio-temporal features for all the frames are concatenated together to build a single vector for all shape features.

In our work, we modified the algorithm in [34] that works with 3D closed surface (genus-zero surface) for modeling and enhancing 3D complex morphological structures. First, we convert the mesh surface into a voxel surface which has uniform vertex sampling in order to fill all the holes on the surface. Then, we adapt only the initial mapping method to the voxel surface in order to accelerate the process of the spherical parametrization and to establish a one-to-one mapping on the unit sphere. Afterwards, spherical harmonics expansion is used to calculate spherical harmonics coefficients that represent the human body surface in the frequency domain. Our algorithm is used only to calculate the spherical harmonics coefficients from a 3D surface without any smoothing, enhancement or optimization process on the surfaces compared with the Shen et al. [34] algorithm.

3.2. Multi-feature fusion using SimpleMKL

The MKL algorithms have recently received great attention in the field of computer vision and pattern recognition. The idea behind MKL is to optimally combine and utilize multiple kernels and features instead of using a single kernel in learning kernel based classifiers. In this work, SimpleMKL algorithm [16] based SVM is employed to fuse multiple types of features with multiple kernel functions to create discriminative weights d_m, $m=1,\dots,M$ for generalizing a multiple kernel matrix K in multiclass-SVM based classifiers. Furthermore, combining multiple types of features in MKL helps the classifier to achieve a high recognition rate since different features reflect different pieces of useful information. Practically, two types of features are extracted from each frame t: the motion features x_i from 3D distal limbs segments, and the shape features z_i from spherical harmonics coefficients.

Let us assume that we are given a set of N training samples $\{(x_i, z_i, y_i)\}_{i=1}^{N}$, where x_i is a feature vector of the *ith* sample in the training set $\mathcal{X}$, $x_i \in R^{D_1}$, and $z_i \in R^{D_2}$ is another feature vector of that sample in the set $\mathcal{Z}$, and $y_i \in \{-1, +1\}$ is their corresponding class label. $k(\cdot,\cdot)$ is a kernel function that maps two feature vectors to be a positive scalar. The SimpleMKL algorithm was proposed to address the MKL-based SVM problem by solving the convex problem defined as

$$\min_{d} \max_{\beta} \quad L(d,\beta) = -\frac{1}{2}\beta^T\left(\sum_{m=1}^{M} d_m K^m\right)\beta + y^T\beta$$
$$\text{s.t.} \quad \sum_{i=1}^{N} \beta_m = 0, \quad 0 \le \beta_i y_i \le C, \quad i = 1,2,\ldots,N$$
$$\sum_{i=1}^{M} d_m = 1, \quad d_m \ge 0, \quad m = 1,2,\ldots,M \tag{8}$$

Here, $y = (y_1, y_2, \ldots, y_N)^T$, and $\beta = (\beta_1, \beta_2, \ldots, \beta_N)^T$ is known as the vector of Lagrangian coefficients in SVM. The multiple kernel matrix K is generated as $K = \sum_{m=1}^{M} d_m K^m$, where K^m is the kernel matrix calculated based on single type of features with single kernel function. $d = (d_1, d_2, \ldots, d_M)^T$ is the kernel combination weight vector, and M is the total number of kernel functions in use.

In our work, suppose we are given a pair of samples (e.g., the *ith* and *jth* registered images), the fusion of the extracted shape $(\{x_i, x_j\})$ and motion features $(\{z_i, z_j\})$ at kernel level within the SimpleMKL framework is handled as follows:

$$K_{i,j} = \sum_{m=1}^{M} (d_m k_m(x_i, x_j) + d_{m+M} k_m(z_i, z_j)) \tag{9}$$

Hence in this fusion work, the dimensionality of kernel combination weight vector d is $2M$.

In the SimpleMKL algorithm, the Two-Step method [41] is used to solve the optimization problem defined in Eq. (8), where two nested iterative loops are set to optimize both the classifier and kernel combination weights. In the inner iteration a solver of SVM is implemented by fixing the vector of kernel combination weights while in the outer iteration a reduced gradient descent algorithm [42], along with the golden section search method [43], is used to update the combination weights with the fixed parameters of the SVM classifier. Therefore, once given a new test sample, its label (determined by y_0) can be calculated according to the following function:

$$y_0 = \text{sgn}\left[\sum_{i=1}^{N}\sum_{m=1}^{M} \beta_i (d_m k_m(x_i, x_0) + d_{m+M} k_m(z_i, z_0))\right] \tag{10}$$

3.3. SimpleMKL based multiclass-SVM

In our work, we employed SimpleMKL framework for multiclass classification. Suppose we want to classify U classes using binary classifiers. Two techniques are commonly used in the literature: one-against-one and one-against-rest. In the one-against-one technique, $U(U-1)/2$ binary classifiers are built from all pairs of distinct classes, whereas in the one-against-rest technique U binary classifiers are built for each class of data.

The authors of [16] presented a structure of MKL based multiclass-SVM using the SimpleMKL algorithm in the outer iteration of the Two-Step method [41]. In their structure, a single kernel combination weight vector is jointly learned for all binary classifiers in the multiclass-SVM, and the general objective function $L(d)$ is defined as

$$L(d) = \sum_{u \in \Phi} L_u(d) \tag{11}$$

where Φ is the set of all pairs of distinct classes considered in the multiclass-SVM, and $L_u(d)$ is the object function of a binary MKL-based SVM defined in Eq. (8) with fixed β for each binary classifier.

By this definition, the inner loop of the SimpleMKL based multiclass-SVM is meant to solve the multiclass-SVM while in the outer loop a single kernel weight vector is learned to minimize the summation of the objective functions from all binary classifiers. Therefore, the learned optimal kernel weight vector can be used for all binary classifiers, which generally increases the recognition result of multiclass-SVM. In our work, the one-against-one technique is used based on the experiences of the work in [44], and the classification of novel samples is done by a max-wins voting strategy.

4. Experimental results

In our experiments, we used three challenging publicly available datasets for human activity recognition, MSR-Action 3D dataset [2], MSR-Daily Activity 3D dataset [1], and 3D Action-Pairs dataset [31], to evaluate our proposed method. The content of each dataset and the experimental results are described in the following subsections.

Also, for each type of feature two kernel functions are used: Gaussian function and polynomial function, each with different parameters are linearly combined to classify the action using a multiclass-SVM classifier. We used the following configuration for kernel functions to fuse multiple types of features with different kernel function parameters based on the SimpleMKL framework as defined in the following equations:

$$k_{Gaussian}(x,y) = e^{-\|x-y\|^2/\sigma^2} \tag{12}$$

$$k_{poly}(x,y) = \langle x,y \rangle^d, d \in \mathbb{N} \tag{13}$$

where σ is the kernel parameters of the Gaussian function and d is the parameter for varying the order of the polynomial function. In all experiments, we set different values of parameters for the two kernel functions with the criterion that they fill a proper range of the defined domain. For the Gaussian function, we set $\sigma^2 \in \{0.01, 0.1, 1, 5, 10, 60, 500\}$, and for the polynomial function, we set $d \in \{1,2,3\}$. Thus, we obtained 10 alternatives for parameterizing the two defined kernel functions. Hence, given any pair of samples (e.g., the *ith* and *jth*), the fusion of extracted shape features $(\{x_i, x_j\})$ and motion features $(\{z_i, z_j\})$ at the kernel level within our framework is handled as follows:

$$K_{i,j} = \sum_{m=1}^{10} [d_m k_m(x_i, x_j) + d_{m+10} k_m(z_i, z_j)] \tag{14}$$

where $k_m(\cdot,\cdot)$ is one of the 10 optional kernel functions, and $d = (d_1, d_2, \ldots, d_{20})^T$ $(\|d\|_p = 1, p \ge 1)$ is the kernel combination vector to be optimized during MKL.

Within this frame, several experiments were conducted using different number of training samples in order to evaluate the performance of our proposed activity recognition method. The empirical results show that our proposed method is comparable with the state-of-the-art-methods.

4.1. MSR-Action 3D dataset

The MSR-Action 3D dataset [2] consists of depth map sequences captured by a depth camera (RGB-D sensor). It contains different human actions: *high arm wave, horizontal arm wave, hammer, hand catch, forward punch, high throw, draw x, draw tick, draw circle, hand clap, two hand wave, side boxing, bend, forward kick, side kick, jogging, tennis swing, tennis serve, golf swing, pick up and throw*. It includes 20 actions performed by 10 subjects. Each action was performed 2 or 3 times by each subject. The size of the depth map is 320 × 240. All the subjects were facing the camera during the performance, and were given a freedom to perform the actions at their own place in front of the camera. For instance, Fig. 5 shows a sequence of depth maps for Tennis Serve action.

Fig. 5. A sequence of depth maps for Tennis Serve action.

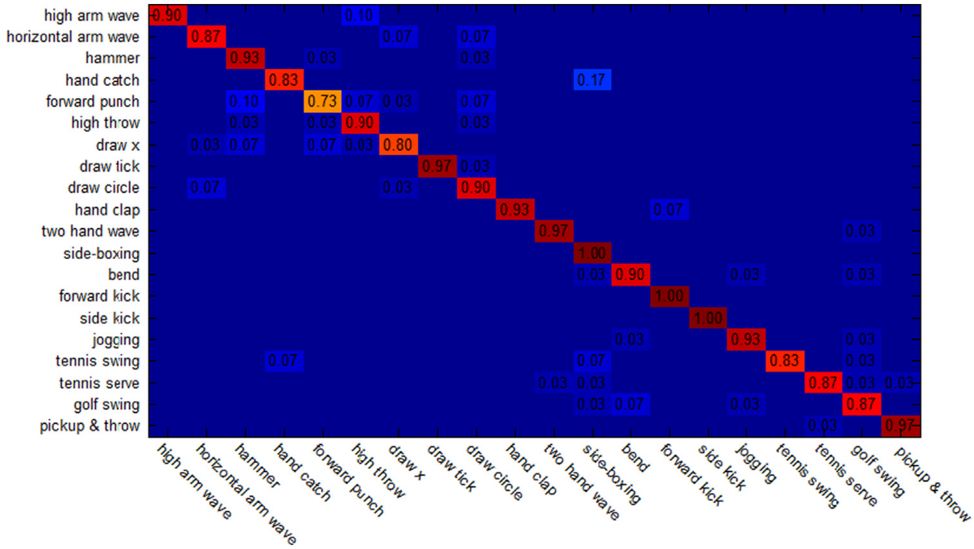

Fig. 6. The confusion matrix for MSR-Action 3D dataset (Classifier C_3).

Furthermore, most of the actions in this dataset involve only the movement of limbs (i.e. arm and leg) in one place, which makes most of the actions highly similar to each other. In fact, this dataset is more challenging compared with the previous datasets such as the Weizmann (Blank et al. [17]) and KTH (Schuldt et al. [45]) datasets in human activity recognition.

In the first experiment, we used the end points of distal limb segments to calculate the orientation as a unit vector and the translation distance with respect to the initial frame as explained in Section 3. In addition, the surface points of the silhouette are employed to calculate the spherical harmonics coefficients. We trained two thirds of samples in total and the other third of the samples were used to test the recognition rate of our proposed method in classifying the performed actions. We repeated this experiment three times, each time we change the training and testing samples, and the results of classifying actions were averaged. Fig. 6 illustrates the result of action classification in the form of a confusion matrix. From Fig. 6, we note that the proposed method achieves an average recognition rate of 90.7% for all actions together. Classification errors occur if there is a high rate of similarity among the actions, such as ”forward punch” and ”draw x” or if the occlusion occurs among human limbs making the skeleton tracker fail as in the tennis swing action. Therefore, since the skeleton tracker sometimes fails and because of the high rate of similarity among the actions, we considered the 90.7% recognition rate for 20 actions a success comparable with other methods [2,3].

In addition, we compared the performance of our proposed method with the state-of-the-art methods which are evaluated on the same dataset (MSR-Action 3D dataset [2]). The first approach was developed by Xia et al. [3], where they modeled the histograms of 3D skeleton joint locations by discrete hidden Markov models in order to train/classify the action. The second approach proposed by Li et al. [2] is a Bag-of-3D-Points model for action recognition.

Table 1
The three subsets of actions used in comparison.

Action set 1 (AS1)	Action set 2 (AS2)	Action set 3 (AS3)
Horizontal wave	High wave	High throw
Hammer	Hand catch	Forward kick
Forward punch	Draw X	Side kick
High throw	Draw tick	Jogging
Hand clap	Draw circle	Tennis swing
Bend	Hands wave	Tennis serve
Tennis serve	Forward kick	Golf swing
Pickup throw	Side boxing	Pickup throw

In order to make a fair comparison, we follow the same experimental settings as [3,2] by dividing the twenty actions into three subsets, each subset contains eight actions, as listed in Table 1. We compare our method with the state-of-the-art methods in two tests, where different number of training samples are chosen. In the first test (Test-1), since each subject has three samples, two third of the samples were used as training samples and the rest as testing samples. In the second test (Test-2), which is a cross subject test, half of the subjects were used for training and the rest subjects were used for testing (all the samples were used in Test-2). Then, the results of action classification were averaged over three different training/testing subsets. Table 2 reports the average recognition rates for the state-of-the-art methods.

As Table 2 shows, in the second test (Test-2), the overall accuracies are lower than Test-1 for all the methods because some of the actions are very similar (e.g. forward punch, high throw, draw x, draw tick, and draw circle) and also the number of training samples are lower than the number of samples in Test-1. Furthermore, Test-2 is conducted across subjects, whereas in Test-1 the same subjects may be in both training/testing (2/3 of the samples were used as training and 1/3 of the samples were used as testing). Overall, our approach outperforms [2] in both tests. However, compared with [3], it gains 1% in Test-2 but degrades by 2.7% in Test-1. This degradation is due to the fact that [3] requires building a large posture vocabulary set and manually encoding the action sequences which improve the accuracy. Whereas, all the processes of our method are done automatically starting from depth map and ending with the action classification.

Table 2
Recognition accuracy comparison using MSR-Action3-D dataset.

Subset	Xia et al. [3]		Li et al. [2]		Our method	
	Test-1	Test-2	Test-1	Test-2	Test-1	Test-2
AS1	0.986	0.879	0.934	0.729	**0.932**	**0.743**
AS2	0.979	0.854	0.926	0.719	**0.945**	**0.768**
AS3	0.949	0.634	0.963	0.792	**0.956**	**0.867**
Overall	0.971	0.789	0.941	0.747	**0.944**	**0.797**

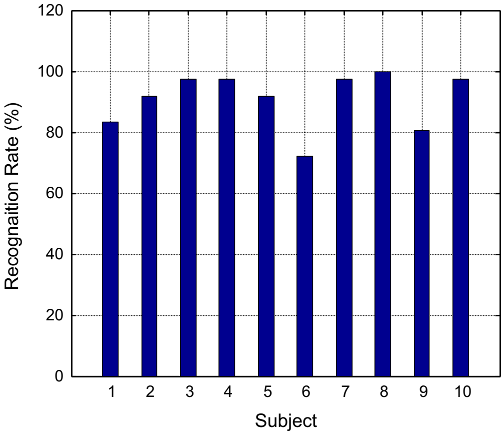

Fig. 9. The accuracy of each subject for 3D Action-Pairs dataset.

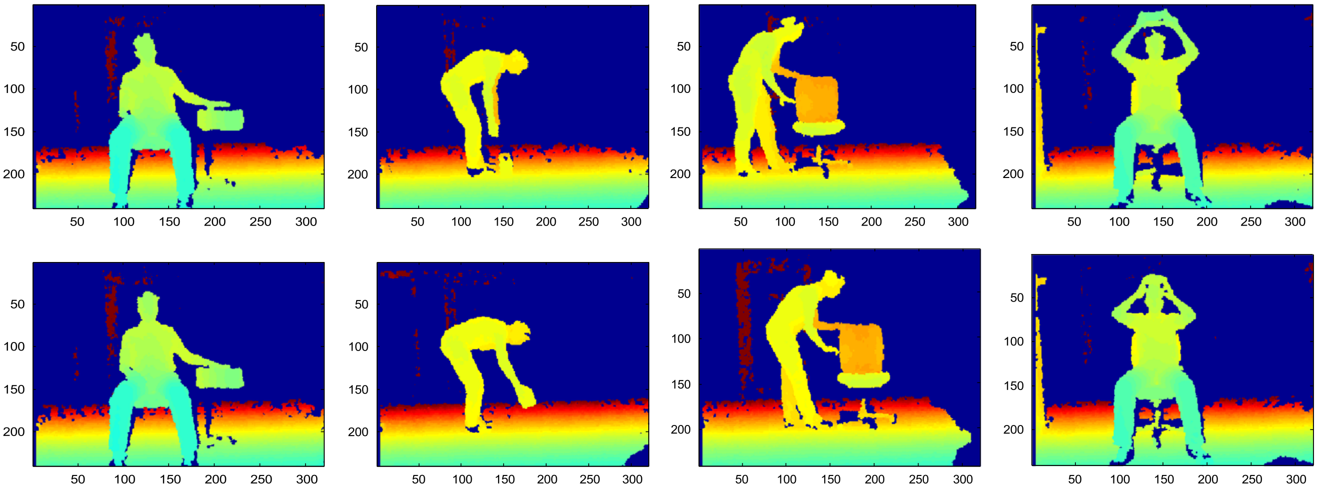

Fig. 7. Example frames for four pairs from 3D Action Pairs dataset; each column shows two frames from a pair of activities. The activities from left to right: “Pick up a box/Put down a box”, “Lift a box/Place a box”, “Push a chair/Pull a chair”, and “Wear a hat/Take off a hat”.

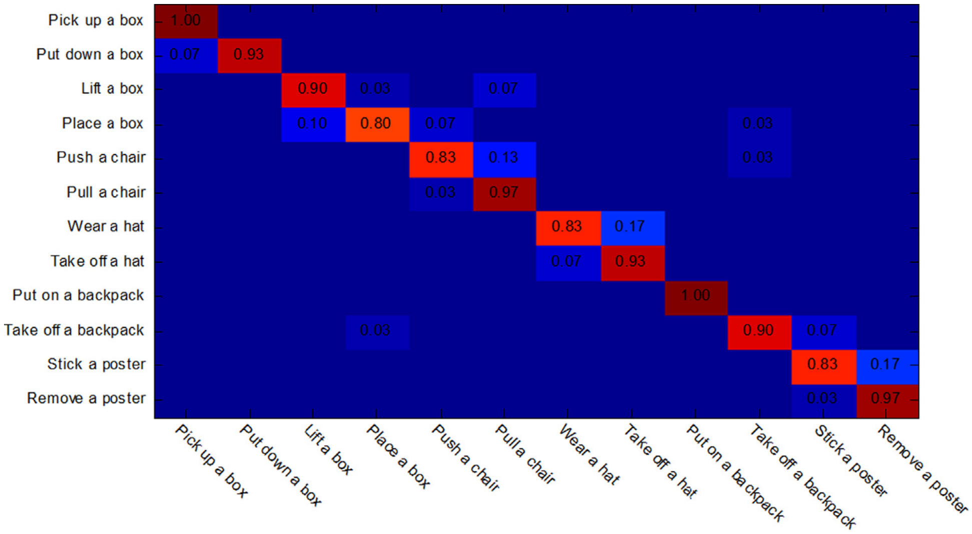

Fig. 8. The confusion matrix for 3D Action-Pairs dataset.

4.2. 3D action-pairs dataset

The 3D Action-Pairs dataset [31] contains new styles of activities which are selected in pairs such that the two activities of each pair are similar in motion (have similar trajectories) and shape (have similar objects). For instance, "Pick up" and "Put down" actions have similar motion and shape. This dataset has six pairs of activities: *"Pick up a box/Put down a box", "Lift a box/Place a box", "Push a chair/Pull a chair", "Wear a hat/Take off a hat", "Put on a backpack/Take off a backpack"*, and *"Stick a poster/Remove a poster"*. A few samples of the 3D Action Pairs Dataset are shown in Fig. 7. The dataset includes 12 activities performed by 10 different subjects. Each activity was performed three times by each subject. We used this dataset in order to emphasize two points: First, to evaluate the performance of our proposed method in the case of activities that have similar trajectories and objects. Second, to show the advantage of using the feature fusion at the kernel level to enhance the recognition rate.

Table 3
Recognition accuracy comparison using 3D Action-Pairs dataset [31].

Method	Accuracy
(Skeleton + LOP + Pyramid) CVPR 2012 (Wang et al. [1])	0.823
HON4D CVPR 2103 (Omar et al. [31])	0.930
DMM-HOG ACM 2012 (Yang et al. [4])	0.661
Our method	**0.908**

In order to verify the performance of our method, Leave-One-Subject-Out (LOSO) cross validation was applied. In the LOSO test, one subject is removed from the training set and the other subjects were utilized to train the multiclass-SVM classifier. The excluded subject is used to test the accuracy of our method in classifying the performed activities. This process is repeated for all the subjects, and the results are averaged for all test subjects. Fig. 8 illustrates the classification results in terms of a confusion matrix. From Fig. 8 we note that the proposed method achieves an average recognition rate of 90.8%. Also, we observe that most of the classification errors occur in the activity pairs because some subjects have a low accuracy due to the noisy skeleton data. Fig. 9 illustrates the variation in the accuracy among the 10 subjects. We can observe that most of the subjects have a high accuracy (over than 80%) except subject 6 (the accuracy is 70%) due to the noisy skeleton data for this subject. By excluding subject 6 from the experiment, the recognition rate of our method is increased from 90.8% to 93%. Overall, these results show a significant improvement in the recognition rate when multiple types of features are used for activity classification.

In Table 3, we compare the performance of our method with the state-of-the-art methods which used the 3D Action-Pairs dataset [31]). The first method [1] utilizes three types of features: the skeleton-

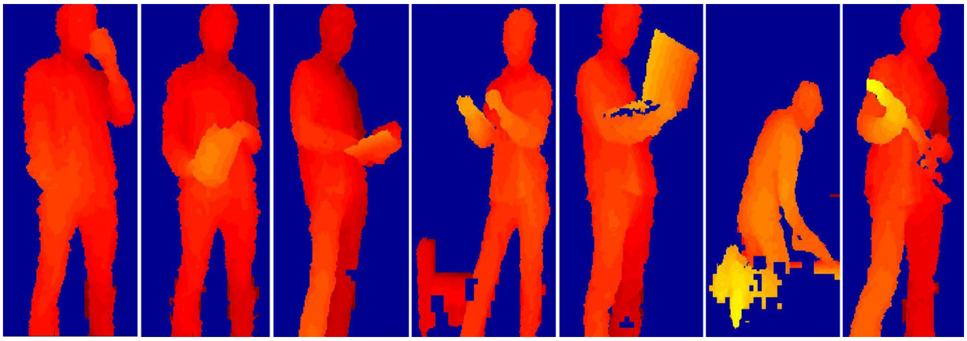

Fig. 10. A few sample frames of MSR-Daily Activity 3D dataset; the activities from left to right: drink, eat, read book, play game, use laptop, use vacuum cleaner, and play guitar.

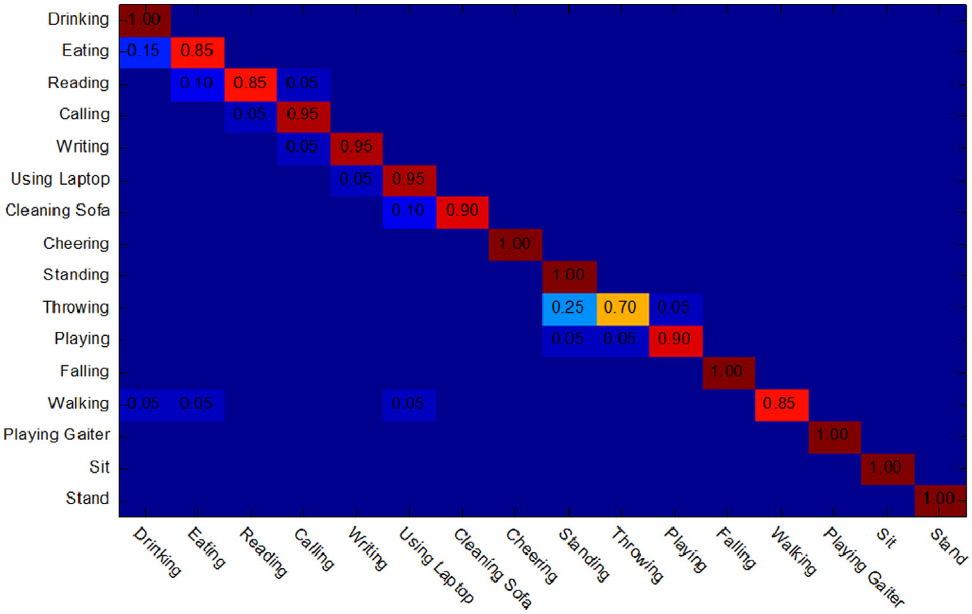

Fig. 11. The confusion matrix for MSR-Daily Activity 3D dataset.

based pair-wise feature, Local Occupancy Pattern (LOP) feature, and a temporal pyramid feature. The other works are the motion map method (DMM-HOG) proposed by Yang et al. [4] and HON4D method presented by Omar et al. [31]. As Table 3 shows, our method significantly outperforms [1] and [4] but is comparable with the HON4D method [31]. The HON4D approach [31] uses only the surface normals (shape features) extracted from the depth map and requires to create a 4D projector in order to quantize the 4D space. Although, this process is complex, it is superior to other motion-based methods in describing activities with similar silhouette shapes. However, it may fail in describing shapes with different poses such as the actions in the MSR-Daily Activity dataset (see the discussion in 4.3).

4.3. *MSR-Daily Activity 3D dataset*

In another experiment, the MSR-Daily Activity 3D dataset [1] was utilized to evaluate the performance of our method for recognizing different human activities. The MSR-Daily Activity 3D dataset contains 16 different human activities: *drink, eat, read book, call cellphone, write on a paper, use laptop, vacuum cleaner use, cheer up, sit still, toss paper, play game, lay down on sofa, walk, play guitar, stand-up, sit-down*, and each subject performs an activity in two different poses: a standing pose and a sitting on sofa pose. Each pose has 160 total samples, with each subject performing one sample per activity in each pose. A few samples of our MSR-Daily Activity 3D dataset are shown in Fig. 10. This dataset is created to cover daily activities and human–object interactions in the living room. These tests are more challenging than the other datasets because of frequent human–object interactions. Furthermore, the segmentation of the dataset is not clean and the skeleton tracker is more noisy in the sitting pose.

Table 4
Recognition accuracy comparison using MSR-Daily Activity dataset.

Method	Accuracy
Restricted graph-based genetic programming (RGGP) IJCAI 2013 (Liu et al. [46])	0.904
Actionlet Ensemble Model CVPR 2012 (Wang et al. [1])	0.857
HON4D CVPR 2103 (Omar et al. [31])	0.800
Our method	**0.931**

We used both the standing and sitting poses in our experiments. Since the silhouette shape is different in the sitting pose, we restricted our activity recognition to motion features only. Shape features had to be avoided because of the differences in the silhouette shape between the sitting pose and the standing pose. In this context, the end points of distal limb segments were used to calculate the orientation and the translation distance of the distal limbs as motion features.

Since this dataset has one sample per subject, LOSO cross validation was applied in our experiments. The average recognition rate is 93.1%. From the confusion matrix in Fig. 11, we can observe that there are confusions among some activities such as *write on a paper, use laptop*, and *read book*.

Table 4 presents the accuracy of our proposed method and the state-of-art-methods which used the MSR-Daily Activity 3D dataset [1]. These methods are the actionlet ensemble model [1], the restricted Graph-based Genetic Programming (RGGP) [46], the Fourier Temporal Pyramid method [1], and the HON4D method [31]. It is clear that our proposed method significantly outperforms the state-of-art methods that rely on 3D joint positions. In fact, in MSR-Daily Activity dataset, each subject performs an activity in two different poses: a standing pose and a sitting on sofa pose. In particular, the method in [31] utilizes only the surface normals of the silhouette shape. These shape features do not have the same characteristics between the two poses while motion-based features such as ours can better discriminate between the standing pose and sitting pose.

4.4. *Discussion*

In order to evaluate the performance of using multiple types of features with a SimpleMKL based multiclass-SVM classifier for human activity recognition, three independent SVM classifiers are adopted in our experiments. Two classifiers C_1 and C_2 are used as C-SVM classifiers with a single kernel function and a group of kernel parameters to classify the activities using either shape features or motion features separately. Classifier C_1 is used for spherical harmonics coefficients as shape features and classifier C_2 is used for the motion features. The other classifier C_3 is built based on a SimpleMKL based multiclass-SVM classifier used for the fused features. Classifier C_3 is designed with two kernel functions (Gaussian and polynomial) and a group of kernel parameters. This classifier C_3 is used to create one kernel weight vector for activity classification.

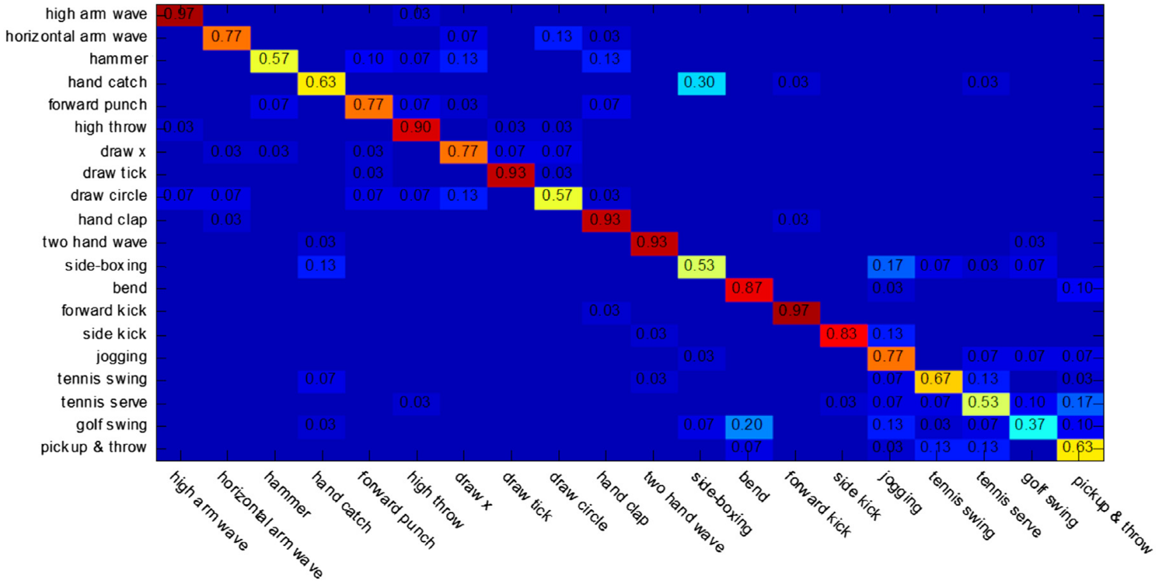

Fig. 12. Confusion matrix of classifier C_1 using only the shape features with single kernel function (average recognition rate: 74.1%).

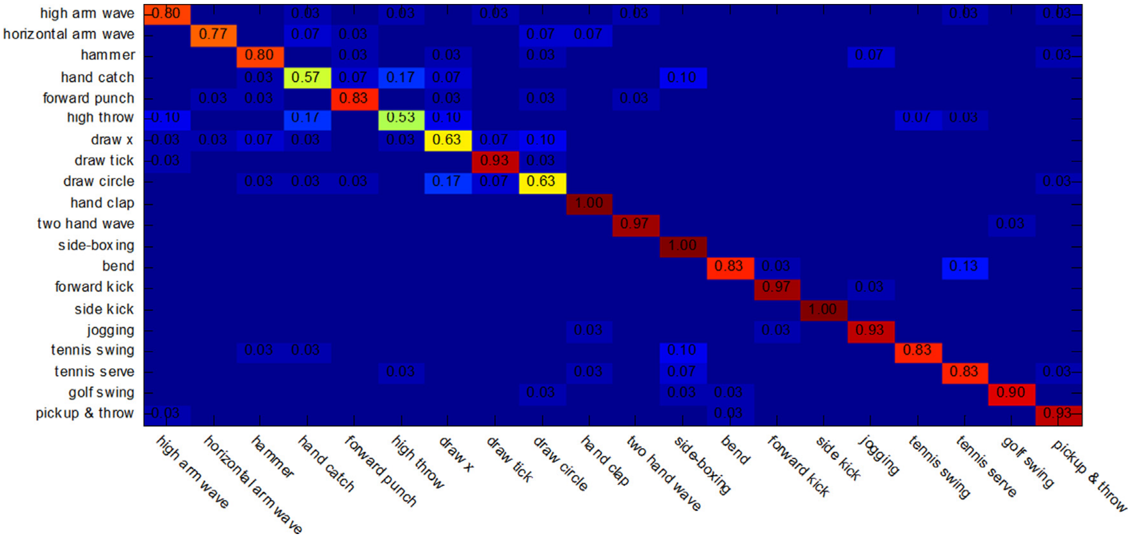

Fig. 13. Confusion matrix of classifier C_2 using only the motion features with single kernel function (average recognition rate: 83.7%).

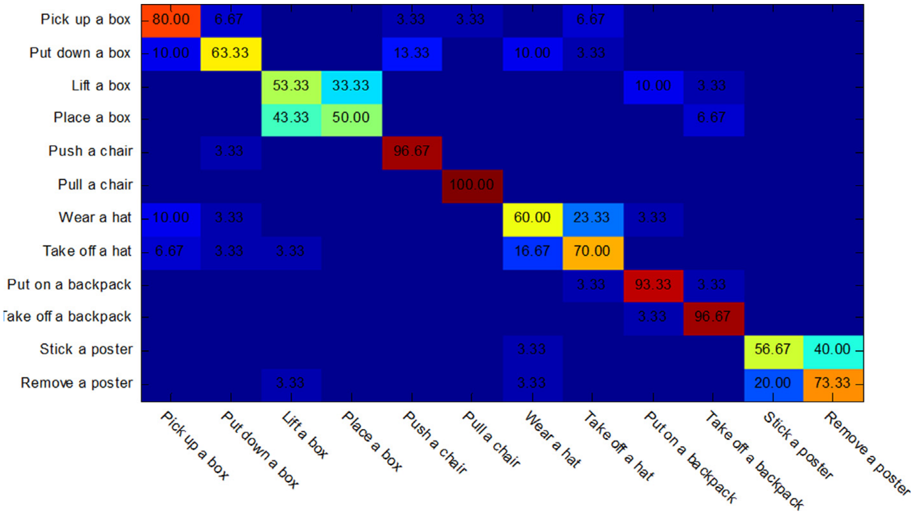

Fig. 14. Confusion matrix of classifier C_1 using only the shape features with single kernel function (average recognition rate: 74.4%).

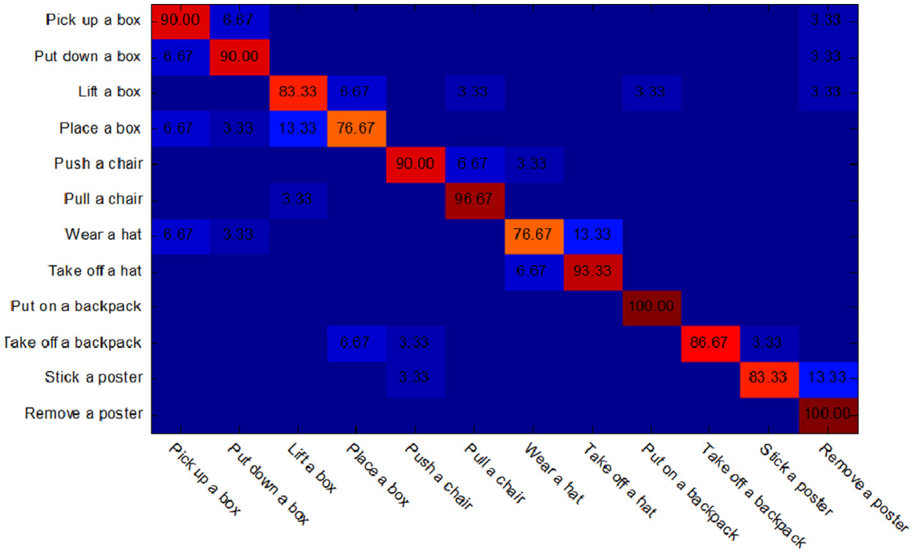

Fig. 15. Confusion matrix of classifier C_2 using only the motion features with single kernel function (average recognition rate: 88.8%).

In our work, we tuned 20 alternative kernel function parameters for Gaussian kernel functions and polynomial kernel functions. By fusing shape features and motion features, our proposed method can automatically learn the optimal kernel combination weights. This advantage comes from utilizing multiple types of features and different kernel functions, which enhances the performance of the classifier system in the kernel level.

Figs. 12, 13 and 8 illustrate the calculated confusion matrix for the three classifiers C_1, C_2 and C_3 using MSR-Action 3D dataset [2]. By comparing the results, we can observe that our proposed method using a SimpleMKL-based multiclass-SVM can generally boost the recognition rate of human activity recognition by fusing multiple types of features with different kernel functions and parameters as shown in Fig. 6. Specifically, the recognition rate for most actions have been increased from 74.1% for classifier C_1, and 83.7% for classifier C_2 to 90.7% using classifier C_3 which represents the fused features using the MKL technique.

By comparing the results in the three Figs. 12, 13 and 6, we can observe that the feature fusion technique at the kernel level enhances the recognition rate especially if there is a big difference in the accuracy between the two sets of features (Classifiers C_1 and C_2). In other words, there are some actions such as side boxing, side kick, tennis serve, golf serve, pick up and throw that are better described and recognized using kinematic structure than the shape features and vice versa. Hence, combining those two sets of features can enhance the classification performance of similar activities.

Figs. 14, 15 and 8 show the confusion matrix for each classifier C_1, C_2 and C_3 with shape features, motion features and fused features respectively using the 3D Action-Pairs dataset [31]. By comparing the results in the confusion matrices, we can observe that there is an improvement in the recognition rate by fusing two types of features with two kernel functions. Specifically, the recognition rate for most activities has been increased from 74.4% for classifier C_1, and 88.8% for classifier C_2 to 90.8% for classifier C_3 which represents the fused features. In addition, we note that the MKL technique works well in improving the performance if the two sets of features have a big difference in the recognition rate for each activity. Otherwise, the enhancement will be small if the difference in the recognition rate is small between the two sets of features (Classifiers C_1 and C_2). Based on the comparison among the experimental results of the three classifiers, we observed that the SimpleMKL-based multiclass-SVM (MKL technique) enhances the classification performance and outperforms single kernel function.

5. Conclusion and future work

We presented two sets of features obtained from RGB-D data acquired using a Kinect sensor for human activity recognition. The features are (1) a novel 3D spatio-temporal feature obtained from the end points of the distal limb segments and (2) a novel shape feature using spherical harmonics coefficients extracted from the surface points. These features were fused via the SimpleMKL algorithm and multiclass-SVM using two kernel functions (Gaussian and polynomial functions) in order to enhance the classification performance for similar activities. Our experimental results on three challenging public datasets have shown the significance of the presented features in human activity recognition. In the future, we will exploit the efficiency of 3D data for view invariant action recognition and in dealing with more complex human object interactions.

Conflict of interest statement

None declared.

References

[1] J. Wang, Z. Liu, Y. Wu, J. Yuan, Mining actionlet ensemble for action recognition with depth cameras, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2012, pp. 1290–1297.

[2] W. Li, Z. Zhang, Z. Liu, Action recognition based on a bag of 3d points, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, 2010, pp. 9–14.

[3] L. Xia, C.-C. Chen, J. Aggarwal, View invariant human action recognition using histograms of 3d joints, in: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), June 2012, pp. 20–27.

[4] X. Yang, C. Zhang, Y. Tian, Recognizing actions using depth motion maps-based histograms of oriented gradients, in: Proceedings of the 20th ACM international conference on Multimedia, MM '12, New York, NY, USA, ACM, 2012, pp. 1057–1060.

[5] X. Yang, Y. Tian, Eigenjoints-based action recognition using Naive–Bayes-nearest-neighbor, in: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), June 2012, pp. 14–19.

[6] M. Corporation, Kinect for windows. ⟨http://www.microsoft.com/en-us/kinectforwindows/⟩, cited April 2013.

[7] J. Liu, S. Ali, M. Shah, Recognizing human actions using multiple features, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, 2008, pp. 1–8.

[8] L. Wang, H. Zhou, S.-C. Low, C. Leckie, Action recognition via multi-feature fusion and gaussian process classification, in: 2009 Workshop on Applications of Computer Vision (WACV), IEEE, 2009, pp. 1–6.

[9] J. Liu, J. Yang, Y. Zhang, X. He, Action recognition by multiple features and hyper-sphere multi-class SVM, in:20th International Conference on Pattern Recognition (ICPR) 2010, IEEE, 2010, pp. 3744–3747.

[10] R. Benmokhtar, Robust human action recognition scheme based on high-level feature fusion, Multimed. Tools Appl. (2012) 1–23.

[11] K. Tran, I.A. Kakadiaris, S.K. Shah, Fusion of human posture features for continuous action recognition, in: Proceedings of the European Conference on Computer Vision Workshop on Sign, Gesture, and Activity, Crete, Greece, 2010.

[12] M. Gönen, E. Alpaydın, Multiple kernel learning algorithms, J. Mach. Learn. Res. 12 (2011) 2211–2268.

[13] A. Noguchi, K. Yanai, A surf-based spatio-temporal feature for feature-fusion-based action recognition, in: Proceedings of ECCV Workshop on Human Motion: Understanding Modeling Capture and Animation, 2010.

[14] N. Ikizler-Cinbis, S. Sclaroff, Object, scene and actions: combining multiple features for human action recognition, Comput. Vis.–ECCV 2010 (2010) 494–507.

[15] M. Bregonzio, T. Xiang, S. Gong, Fusing appearance and distribution information of interest points for action recognition, Pattern Recognit. 45 (3) (2012) 1220–1234.

[16] A. Rakotomamonjy, F. Bach, S. Canu, Y. Grandvalet, Simplemkl, J. Mach. Learn. Res. 9 (2008) 2491–2521.

[17] M. Blank, L. Gorelick, E. Shechtman, M. Irani, R. Basri, Actions as space-time shapes, in:Tenth IEEE International Conference on Computer Vision, 2005. ICCV 2005., vol. 2, IEEE, 2005, pp. 1395–1402.

[18] I. Cohen, H. Li, Inference of human postures by classification of 3d human body shape, in: IEEE International Workshop on Analysis and Modeling of Faces and Gestures, 2003, AMFG 2003, IEEE, 2003, pp. 74–81.

[19] A. Yilmaz, M. Shah, A differential geometric approach to representing the human actions, Comput. Vis. Image Underst. 109 (3) (2008) 335–351.

[20] H.-S. Chen, H.-T. Chen, Y.-W. Chen, S.-Y. Lee, Human action recognition using star skeleton, in: Proceedings of the 4th ACM International Workshop on Video Surveillance and Sensor Networks, VSSN '06, New York, NY, USA, ACM, 2006, pp. 171–178.

[21] H. Fujiyoshi, A.J. Lipton, Real-time human motion analysis by image skeletonization, in: In Proceedings of IEEE WACV98, 1998, pp. 15–21.

[22] S. Chun, K. Hong, K. Jung, 3d star skeleton for fast human posture representation, in: World Academy of Science and Engineering, vol. 34, 2008, pp. 273–282.

[23] S.R. Althloothi, M.H. Mahoor, R.M. Voyles, Fitting distal limb segments for accurate skeletonization in human action recognition, J. Ambient Intell. Smart Environ. 4 (2) (2012) 107–121.

[24] S. Althloothi, R. Voyles, M. Mahoor, G. Wu, 2d human skeleton model from monocular video for human activity recognition, in: The 2010 International Conference on Image Processing, Computer Vision and Pattern Recognition, 2010.

[25] J. Sun, X. Wu, S. Yan, L.-F. Cheong, T.-S. Chua, J. Li, Hierarchical spatio-temporal context modeling for action recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, June 2009, pp. 2004–2011.

[26] Y. Tian, L. Cao, Z. Liu, Z. Zhang, Hierarchical filtered motion for action recognition in crowded videos, IEEE Trans. Syst. Man Cybern. Part C: Appl. Rev. 42 (3) (2012) 313–323.

[27] J. Sung, C. Ponce, B. Selman, A. Saxena, Unstructured human activity detection from RGBD images, in: 2012 IEEE International Conference on Robotics and Automation (ICRA), May 2012, pp. 842–849.

[28] H. Zhang, L. Parker, 4-dimensional local spatio-temporal features for human activity recognition, in: 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), September 2011, pp. 2044–2049.

[29] Y. Zhao, Z. Liu, L. Yang, H. Cheng, Combing RGB and depth map features for human activity recognition, in: 2012 Asia-Pacific Signal Information Processing Association Annual Summit and Conference (APSIPA ASC), December 2012, pp. 1–4.

[30] C.-C. Chang, C.-J. Lin, Libsvm: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (3) (2011) 27.

[31] O. Oreifej, Z. Liu, W. Redmond, Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences, in: Computer Vision and Pattern Recognition (CVPR), 2013.

[32] M.B. Holte, T.B. Moeslund, P. Fihl, View-invariant gesture recognition using 3d optical flow and harmonic motion context, Comput. Vis. Image Underst. 114 (12) (2010) 1353–1361.

[33] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, A. Blake, Real-time human pose recognition in parts from single depth images, in: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2011, pp. 1297–1304.

[34] L. Shen, H. Farid, M.A. McPeek, Modeling three-dimensional morphological structures using spherical harmonics, Evolution 63 (4) (2009) 1003–1016.

[35] D.V. Vranic, An improvement of rotation invariant 3d-shape based on functions on concentric spheres, in: Proceedings of 2003 International Conference on Image Processing, ICIP 2003, vol. 3, IEEE, 2003, pp. III–757.

[36] D. Saupe, D.V. Vranić, 3d model retrieval with spherical harmonics and moments, in: Pattern Recognition, Springer, 2001, pp. 392–397.

[37] S. Althloothi, M. Mahoor, R. Voyles, A robust method for rotation estimation using spherical harmonics representation, IEEE Trans. Image Process. 22 (6) (2013) 2306–2316.

[38] L. Shen, S. Kim, A.J. Saykin, Fourier method for large-scale surface modeling and registration, Comput. Graph. 33 (3) (2009) 299–311.

[39] C. Brechbühler, G. Gerig, O. Kübler, Parametrization of closed surfaces for 3-d shape description, Comput. Vis. Image Underst. 61 (2) (1995) 154–170.

[40] L. Shen, F. Makedon, Spherical mapping for processing of 3d closed surfaces, Image Vis. Comput. 24 (7) (2006) 743–761.

[41] O. Chapelle, A. Rakotomamonjy, Second order optimization of kernel parameters, in: Proceedings of the NIPS Workshop on Kernel Learning: Automatic Selection of Optimal Kernels, 2008.

[42] D.G. Luenberger, Y. Ye, Linear and Nonlinear Programming, Springer, 2008, vol. 116.

[43] J. Nocedal, S.J. Wright, Numerical Optimization, Springer Verlag, 1999.

[44] C.-W. Hsu, C.-J. Lin, A comparison of methods for multiclass support vector machines, IEEE Trans. Neural Netw. 13 (2) (2002) 415–425.

[45] C. Schuldt, I. Laptev, B. Caputo, Recognizing human actions: a local SVM approach, in: Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004, vol. 3, IEEE, 2004, pp. 32–36.

[46] L. Liu, L. Shao, Learning discriminative representations from RGB-D video data, in: Proceedings of International Joint Conference on Artificial Intelligence (IJCAI), 2013.

[47] Xiao Zhang, Mahoor, M.H., Voyles, R.M. Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on Facial expression recognition using HessianMKL based multiclass-SVM, 2013, pp. 1–6.

Salah Althloothi received the B.S. degree in electrical and computer engineering from Tripoli University, Tripoli, Libya, in 1993, the M.S. degree in computer systems engineering from University Putra Malaysia, Malaysia, in 2003. He is currently pursuing the Ph.D. degree with the University of Denver, Denver, CO, USA. His current research interests include image processing, pattern classification and human activity recognition.

Mohammad H. Mahoor (S'03-M'07) received the B.S. degree in Electronics from Abadan Institute of Technology, Iran, in 1996, the M.S. degree in biomedical engineering from Sharif University of Technology, Iran, in 1998, and the Ph.D. in electrical and computer engineering from University of Miami, Florida, in 2007. He joined the University of Denver (DU) as assistant professor of computer engineering in September 2008. He has authored or co-authored over 60 refereed research publications. He is the director of image processing and computer vision laboratory at DU. His research interests include affective computing and particularly developing automated systems for facial expression recognition.

Xiao Zhang received the B.S. degree from the Beijing Institute of Petrochemical Technology, Beijing, China, in 2008 and the M.S. degree from the Beijing Institute of Technology, Beijing, China, in 2010. He is currently a Ph.D. candidate and graduate research assistant in the Department of Electrical and Computer Engineering, University of Denver, Denver. His research interests include automatic analysis of facial expression and action units, pattern recognition, and machine learning.

Richard Voyles received the B.S. degree in electrical engineering from Purdue University, West Lafayette, IN, USA, in 1983, the M.S. degree in manufacturing systems engineering from the Department of Mechanical Engineering, Stanford University, Stanford, CA, USA, in 1989, and the Ph.D. degree in robotics from the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA, in 1997. His current research interests include robotics and artificial intelligence. Specifically, he is interested in the development of small resource-constrained robots and robot teams for urban search and rescue and surveillance. He has additional expertise in sensors and sensor calibration, particularly haptic and force sensors, and real-time control. His industrial experience includes Dart Controls, IBM Corp., Integrated Systems, Inc., and Avanti Optics.