IET Computer Vision

*Special Issue on Unsupervised Feature Learning for Vision*

IET Journals
The Institution of Engineering and Technology

# Discriminating features learning in hand gesture classification

ISSN 1751-9632
Received on 28th November 2014
Revised on 7th May 2015
Accepted on 15th May 2015
doi: 10.1049/iet-cvi.2014.0426
www.ietdl.org

*Feng Jiang[1], Cuihua Wang[2] ✉, Yang Gao[1], Shen Wu[1], Debin Zhao[1]*

[1]*School of Computer, Harbin Institute of Technology, Harbin, People's Republic of China*
[2]*School of Computer, Harbin Institute of Technology, Weihai, People's Republic of China*
✉ *E-mail: cuihuawanghit@163.com*

**Abstract:** The advent and popularity of Kinect provides a new choice and opportunity for hand gesture recognition (HGR) research. In this study, the authors propose a discriminating features extraction for HGR, in which features from red, green and blue (RGB) images and depth images are both explored. More specifically, histogram of oriented gradient feature, local binary pattern feature, structure feature and three-dimensional voxel feature are first extracted from RGB images and depth images, then these features are further reduced with a novel deflation orthogonal discriminant analysis, which enhances the discriminative ability of the features with supervised subspace projection. The extensive experimental results show that the proposed method improves the HGR performance significantly.

## 1 Introduction

Hand gestures, an unsaid body language, play very important roles in daily communication. They are considered as the most natural expressive means of communication between humans and computers [1]. For the purpose of improving humans' interaction with computers, considerable scholarly work has been undertaken on hand gesture recognition (HGR), whose extensive applications include sign language recognition [2], socially assistive robotics [3], directional indication through pointing [4] and so on [5].

### 1.1 Hand gesture recognition

Hands have highly articulated structures with some 27° of freedom (DOF). Owing to these high DOF of the human hand, HGR is indeed an extremely challenging task. Many researchers have tried with different instruments and equipment to measure hand movements such as sensors or wires. For sensor-based methods, Kadous [6] demonstrated a system based on Powergloves to recognise isolated Australian sign language. Fels and Hinton [7] developed a system using a data glove with a Polhemus tracker as input devices. Kim *et al.* [8] used fuzzy min–max neural network to recognise manual alphabets and Korean signs based on data gloves. Fang used fuzzy decision trees and synthetic data generation technique for large vocabulary Chinese sign recognition [9, 10].

The sensor-based methods have high recognition accuracy because they can precisely catch the movement of hands [9, 10]. However, due to cumbersome and expensiveness of devices, these techniques make less sense in practical usage [11]. Consequently, computer vision-based HGR system, which can perform recognition as natural as human-to-human interaction, is considered to be the more promising method, as local features such as histogram of oriented gradient (HOG) [12] and local binary patterns (LBPs) [13] extracted from red, green and blue (RGB) pictures can be efficient used in computer vision-based HGR system.

Computer vision-based methods depend on direct registration of hand gestures with 2D image features and many promising vision-based system have been developed [14]. Skin colour is one of the most important clues commonly used in computer vision-based HGR [15]. In [16], scale-space colour features were used to recognise hand gestures in user independence conditions. In [17], a clear-cut and integrated hand contour was first obtained and then used to compute the curvature of each point on the contour for recognition. In [18], a view-based approach is proposed for continuous American SLR. They used single camera to extract 2D features and the extracted features were then taken as the input of hidden Markov model (HMM). Recently, there have been some research efforts focusing on local invariant features [19–21]. In [19], AdaBoost learning algorithm and scale-invariant feature transform (SIFT) features were used to achieve in-plane rotation invariant hand detection.

Although these existing vision-based methods have achieved great success, they are still facing challenging problems caused by the complex nature of static and dynamic hand gestures, cluttered backgrounds, transformations, lighting changes, and occlusions. These problems are quite difficult to solve by the current feature descriptors and classifiers based on RGB camera.

### 1.2 Kinect and its application in HGR

One difference between human vision system and ordinary camera is the ability to interpret three-dimensional (3D) information. In an ordinary camera-based system, there is loss in information whenever a 3D image is projected to a 2D plane. Kinect [22] is an infrared light range-sensing camera, a motion sensing input device by Microsoft for the Xbox 360 video game console and Windows personal computers. In June 2011, Microsoft released a Kinect Software Development Kit for extracting scene depth and object masks and subsequently building a skeleton model in real time. Although its resolution and accuracy is relatively low, the low-cost promises to make Kinect the primary 3D measuring devices in human–computer interfaces.

3D information provided by Kinect is an important supplementary to the traditional vision-based HGR especially in the condition that the background clutters or the hands may have different textures. Until now, several works have been done related to Kinect-based gesture recognition and other similar area [23, 24]. In this paper, we propose hand gesture classification based on Kinect, in which two kinds of complemented features on RGB images and depth images are first extracted, more specifically, HOG feature, LBP feature, and structure feature are extracted. It is noted that it is the first time to extract such features from depth image for SLR. The main contributions of this paper are summarised as follows:

- One of the crucial challenges in HGR, how to capture the most meaningful information of gestures is addressed. Instead of

extracting features from the whole frame, the dynamic regions, which convey the most meaningful information of gestures, are effectively extracted.

- HOG features, LBP features, and structure features are extended to the depth domain. We verify that RGB and depth information collaborate with each other in the HGR task, and with the adoption of these features, the HGR performance can be improved significantly.
- On the basis of the subspace projection, the concatenate features are reduced with the proposed deflation orthogonal discriminant analysis to enhance the discriminative ability, in which the feature vectors are mapped to a low-dimensional space where the class separability is maximised with respect to Fisher discriminant criteria. Compared with existing techniques, the proposed method does not depend on the number of classes which determines the rank of the between-class-scatter matrix.

The remainder of this paper is organised as follows. Related work is reviewed in Section 2. The feature explorations in RGB images and depth images are presented in Section 3. Section 4 elaborates the proposed supervised subspace projection methods. Extensive experiment results are reported in Section 5. Section 6 concludes this paper.

## 2 Related work

Vision-based HGR encompasses two main categories: 3D model-based methods and appearance-based methods. The former computes a geometrical representation of a hand configuration using the joint angles of a 3D articulated structure recovered from a hand gesture sequence, which provides a rich description that permits a wide range of hand gestures. However, the computation of 3D model has high computational complexity [23]. In contrast, appearance-based methods extract appearance features from a hand gesture sequence and then construct a classifier to recognise different hand gestures, which have been widely used in vision-based HGR [25].

The well-known features used to locate human hands and recognise hand gestures are colour [26, 27], shapes [28, 29] and motion [30, 31]. In early work, colour information is widely used to segment the hands. To simplify the colour-based segmentation, the users are required to wear single or differently coloured gloves [29, 32]. The skin colour models are also used [33, 34] where a typical restriction is wearing of long sleeved clothes. When it is difficult to exploit colour information to segment the hands from an image, motion information extracted from two consecutive frames is used for HGR. Agrawal and Chaudhuri [35] explore the correspondences between patches in adjacent frames and uses 2D motion histogram to model the motion information. Shao and Ji [36] compute optical flow from each frame and then use different combinations of the magnitude and direction of optical flow to compute a motion histogram. Zahedi *et al.* [37] combine skin colour features and different first- and second-order derivative features to recognise sign language. Wong *et al.* [38] use principal component analysis (PCA) on motion gradient images of a sequence to obtain features for a Bayesian classifier. To extract motion features, Cooper *et al.* [39] extend Haar-like features from spatial domain to spatiotemporal domain and propose volumetric Haar-like features.

The features introduced above are usually extracted from RGB images captured by a traditional optical camera. Owing to the nature of optical sensing, the quality of the captured images is sensitive to lighting conditions and cluttered backgrounds, thus the extracted features from RGB images are not robust. In contrast, depth information from a calibrated camera pair [40] or direct depth sensors such as light detection and ranging is more robust to noises and illumination changes. More importantly, depth information is useful for discovering the distance between the hands and body orthogonal to the image plane, which is an important cue for distinguishing some ambiguous hand gestures. Since the direct depth sensors are expensive, inexpensive depth cameras, for example, Microsoft's Kinect, have been recently used in HGR [24]. Although the skeleton information offered by Kinect is more effective in the expression of human actions than pure depth data, there are some cases that skeleton cannot be extracted correctly, such as interaction between human body and other objects or micro-movement in close distance. Besides, when an action is too fast or hands occlude each other, the positions of hands are difficult to locate. To extract more robust features from Kinect depth images for HGR, Ren *et al.* [41] propose the part-based finger shape features, which do not depend on the accurate segmentation of the hands. Wan *et al.* [42] extends SIFT to spatiotemporal domain and proposes 3D EMoSIFT to extract features from RGB and depth information, which is invariant to scale and rotation, and has more compact and richer visual representations. On the basis of 3D histogram of flow (3DHOF) and global HOG (GHOG), Fanello *et al.* [43] apply adaptive sparse coding to capture high-level feature patterns. Mahbub *et al.* [31] propose a space–time descriptor and apply motion history imaging (MHI) techniques to track the motion flow in consecutive frames.

In many real-world applications, the feature dimension (i.e. the number of features or attributes in an input vector) could easily be as high as tens of thousands. Such extreme dimensionality could be very detrimental to data analysis and processing. As a solution to the curse of dimensionality, feature reduction [44] for classification has been a popular topic for decades. There are many reasons for caring about the dimensionality. (i) Overfitting is inevitable for high-dimensional feature space, which might ruin the generalisation ability of the classifier. (ii) When the number of variables is too large, high storage capacity is required, and computational complexity is yet another issue. (iii) In many cases, high dimensionality causes computational instability and singularity [45]. (iv) Class separability is very likely to be enhanced by eliminating redundant information.

There are two types of feature reduction techniques: feature selection and feature extraction. Feature selection [46] is to select a subset of the variables with respect to some criteria. On the other hand, feature extraction attempts to find a function $f(x)$: $R^m \rightarrow R^k$, which transforms data from the original space $R^m$ to a low-dimensional feature space $R^k$ where $m > k$. Statistically, a training dataset is commonly modelled as multivariate stochastic observations with a Gaussian distribution. In this case, the optimal subspace can be obtained via PCA, which exploits the statistical dependence and inherent redundancy embedded in the multivariate training dataset to obtain a compact description of the data. Pearson [47] proposed PCA in 1901 as a methodology for fitting planes in the least squares sense. Subsequently, it was Hotelling [48] who adopted PCA for the analysis of the correlation structure between many random variables. Some interesting applications may be found in some recent books, for example [49, 50]. Assuming Gaussian distributed data, PCA is well known to be optimal under both mean-square-error and maximum-entropy criteria. PCA can be computed by several numerically stable algorithms, including eigenvalue decomposition and singular value decomposition [51, 52]. Moreover, the optimal performances achieved by PCA can be expressed in closed form. Consequently, PCA is commonly adopted as a convenient tool for feature extraction and visualisation.

Linear discriminant analysis (LDA) has a very long history. The underlying idea is based on Fisher criteria for maximising the class separability. Given class label $c \in \{+, -\}$ and training data $Xc = \{X_1^c, \ldots, X_{N_c}^C\}$, the class separability is measured using the 'between-class scatter matrix' $\boldsymbol{S}_B$ and the 'within-class scatter matrix' $\boldsymbol{S}_W$. In LDA, a vector $\boldsymbol{w}$ is estimated to maximise the following Fisher score:

$$J = \frac{W^T \boldsymbol{S}_B W}{W^T \boldsymbol{S}_W W} \tag{1}$$

The solution is the generalised eigenvector corresponding to the largest eigenvalue of the problem $\boldsymbol{S}_B \boldsymbol{w} = \boldsymbol{S}_W \boldsymbol{w} \lambda$. However,


*IET Comput. Vis.*, 2015, Vol. 9, Iss. 5, pp. 673–680

problems occur when the within matrix $\boldsymbol{S}_W$ is singular. There are many ways of tackling this problem [53, 54] and one way is to compute the eigenvector of the Fisher matrix $\boldsymbol{F} = \boldsymbol{S}_W^{\dagger}\boldsymbol{S}_B$. The objective of LDA is to find a projection vector such that the projected values of data from both classes have maximum class separability. In LDA, only one such vector can be found due to the rank deficiency.

## 3 Features exploration in RGB images and depth images

In the gesture sequence, dynamic regions in each frame contain the most meaningful location information, which are illustrated in Fig. 1 and it should be segmented and represented. We assume that there is a home pose between a gesture and another one in a multi-gesture sequence. The proposed segmentation is illustrated in Fig. 1.

In Fig. 1, threshold segmentations of the foreground and background are applied to obtain the gesturer's foreground on the depth image. The threshold used is maximal depth minus 100. The segmented gesturer's foreground in the first depth image indicates the initial position of the gesture, which is then denoised by median filter and dilated ten times. A swing region is obtained by the subtraction of the denoised foreground from the dilated foreground. The swing region covers the slight swing of gesturer's trunk and can be used to eliminate the influence of body swing. For depth frame $I_t$, define region $\Xi$ as:

$$\Xi = \left\{(m,\ n)|F_1(m,\ n) - F_t(m,\ n) \geq \text{Threshold}_{\text{QOM}}\right\} \quad (2)$$

where $F_1$ and $F_t$ are the foregrounds of the first depth image and the depth image $I_t$, respectively. In the process of gesture communication, the gesturer's trunk has a slight back-and-forth movement due to breathing and keeping balance. This kind of automatic movement should be excluded from the segmentation of dynamic region. On the basis of this consideration and the range of depth information, $\text{Threshold}_{\text{QOM}}$ is set to 60 empirically in this paper. For each connected region in $\Xi$, only if the number of pixels in this region exceeds $N_p$ and the proportion overlapped with swing region is less than $r$ of the connected region, it is regarded as a dynamic region. Here $N_p = 500$ is a threshold used to remove the meaningless connected regions in the difference frame. If a connected region has less than $N_p$ pixels, we think this region should not be a good dynamic region for extracting location features. This parameter can be set intuitively. The parameter $r = 50\%$ is also a threshold used to complement with $N_p$ to remove the meaningless connected regions in the difference frame. After using $N_p$ to remove some connected regions, there may be a retained connected region which has more than $N_p$ pixels, but it may still not be a meaningful dynamic region for extracting position features if the connected region is caused by the large movement of the body. Obviously, we can exploit the swing region to remove such a region. To do this, we first compute the overlap rate between this region and the swing region. If the overlap rate is larger than $r$, it is reasonable to think of this region is mainly produced by the movement of the body. Therefore, it should be further removed.

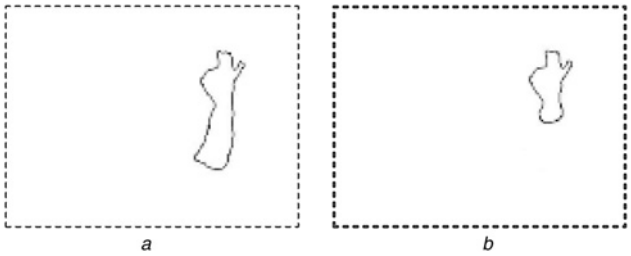


**Fig. 2** *Hand shape segmentation*
*a* C1
*b* C2

Once the dynamic regions of a frame have been obtained, the largest dynamic region D is used for hand shape segmentation. Although hand shapes are complex, they do not have robust texture and structured appearance. In most cases, hand shapes can be distinguished with their shape contour. The contour points of D are extracted using the Canny algorithm. The obtained contour point set is defined as C1. K-means is adopted to cluster the points in D into two clusters using Euclidean distance based on the image coordinates and depth of each point, because if a gesturer faces the camera, the cluster with smaller depth value contains most information for identifying the hand shape component. Canny algorithm is used again to extract contour points of the cluster with smaller depth value. The obtained closed contour point set is defined as C2 (Fig. 2).

### 3.1 Feature extraction

Having obtained the dynamic region, two kinds of complemented features are extracted from the corresponding RGB image and depth image. First, the dynamic region is resized according to the maximal boundary dissentience in horizontal and vertical, with this procedure, a mask image is obtained. Then HOG feature, LBP feature, structure feature and 3D-voxel features are extracted in the dynamic region and concentrated accordingly.

*3.1.1 HOG feature in RGB and depth images:* In the proposed method, HOG features [12] are extracted from both RGB images and depth images according to the obtained mask. For each image of size 48 48, we finally made 14 blocks. Each block is divided into four cells, and for all the pixels in one cell, the gradient is calculated and the histogram of that cell is obtained

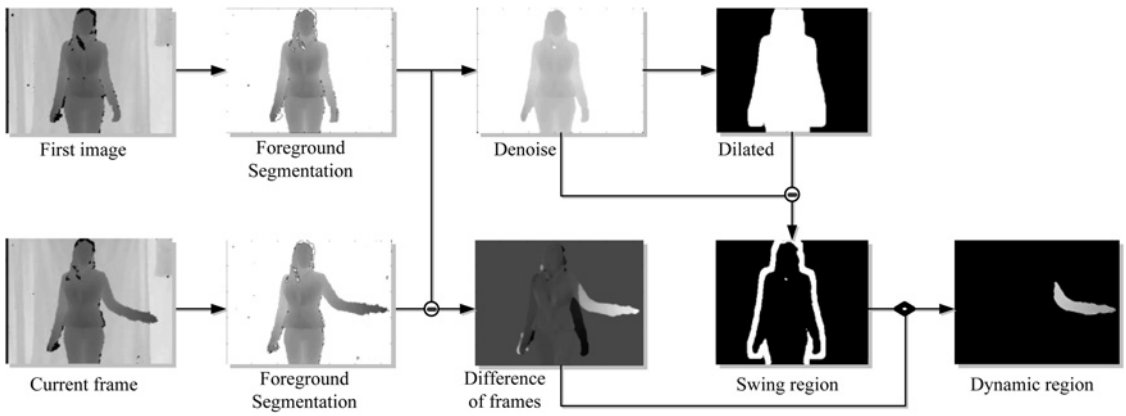


**Fig. 1** *Dynamic region segmentation*

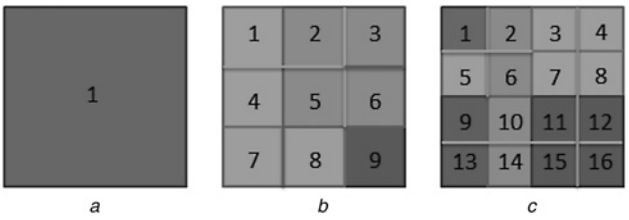


**Fig. 3** *HOG feature in RGB and depth images*

*a* One block in the first layer
*b* Four blocks in the second layer: 1, 2, 4, 5/2, 3, 5, 6/4, 5, 7, 8/5, 6, 8, 9
*c* Nine blocks in the third layer: 1, 2, 5, 6/2, 3, 6, 7/3, 4, 7, 8/5, 6, 9, 10/6, 7, 10, 11/7, 8, 11, 12/9, 10, 13, 14/10, 11, 14, 15/ 11, 12, 15, 16

accordingly, which contains nine bins and each bin corresponding to 40°.

The 14 blocks are arranged into three layers. In the first layer, one block of size is adopted, containing the global information. In the second layer, four blocks of sizes are used, each containing a 16-pixel overlap region with the neighbouring block. In the third layer, we use nine blocks and there is a 12-pixel overlap region between two blocks, as illustrated in Fig. 3. Finally, a 2016 (4 images 14 blocks 4 cells 9 bins) dimension vector is obtained as the HOG feature of one group images.

*3.1.2 LBP feature in RGB and depth images:* LBP proposed by Ojala [13] is a powerful and effective texture description descriptor, keeping invariance to light by measuring and extracting the inner texture information of an image (Fig. 4).

We extract LBP feature from RGB images and depth images. For one mask image of size $48 \times 48$, it is divided into $4 \times 4$ blocks; ignoring the edge of a picture, each block has a size of $14 \times 14$, the starting and ending point is 2–15/12–25/22–35/32–45, so there are four pixels overlapping between two blocks. Finally, an 1888 (2 images × 14 blocks × 4 cells × 9 bins) dimension vector is obtained as the LBP feature of one group images.

*3.1.3 Structure feature:* Structure feature is used to descript the structure character of an obtained mask image. First, obtain the Canny edge of the mask image, and then randomly choose a point P inside the edge enclosed region. Considering the tradeoff between the time complexity and recognition accuracy, for each P, 18 rays are launched, 20° per ray. Accordingly, 18 bins can be obtained, each bin corresponding to a fan of 40° with 20° overlapping. Then the statistics for each bin, mean and variance of the corresponding RGB image and depth image are estimated. To keep the invariance to hand movement, P can be chosen randomly 50 times in a range $\{(x, y), 20 < x, y < 28\}$, and then finally taking the mean vector as the structure feature (Fig. 5).

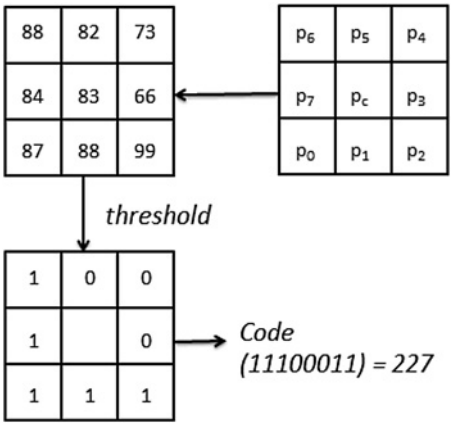


**Fig. 4** *LBP code calculating. If pi < pc, code+ = (1 << i)*

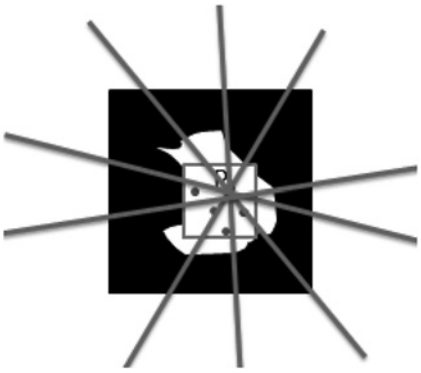

**Fig. 5** *Structure feature*

Centre point P and the rays

*3.1.4 3D-voxel feature:* 3D-voxel feature describes the hand skeletal coordinates of centre which is collected by Kinect. For all the voxel coordinates that have a $z$ value of 1, meaning that point is belonging to hand, save all the distance between that point and the hand joint coordinates, and then find the maximum distance. Normalise all the distance with maximum distance, and a statistical histogram with the set of distance can be obtained, by dividing the range [0, 1] into 16 bins. Here, we use the mean, variance, symmetry as statistical information and finally a 51D feature is obtained as 3D-voxel feature, describing the space information of hands.

## 4 Supervised subspace projection methods

LDA is the most popular supervised feature extraction method. In this method, the between-class scatter matrix $\boldsymbol{S}_{\mathrm{B}}$ is maximised and simultaneously the within-class scatter matrix $\boldsymbol{S}_{\mathrm{W}}$ is minimised. There are some difficulties with LDA. The within-class scatter matrix becomes singular in small sample size situation. Moreover, the rank of between-class scatter matrix is limited. Therefore, LDA can extract maximum $c$–1 features (where $c$ is the number of classes). In LDA, only one such vector can be found due to the rank deficiency for binary classification problems. Having obtained a supervised training dataset $[X, Y]$, the subspace projection is to find an optimal $M \times m$ matrix: $\boldsymbol{W} = [\boldsymbol{w}_1 \ldots \boldsymbol{w}_m]$ and the subspace vector $\boldsymbol{x} \rightarrow \boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}$ presents a lower $m$-dimensional description of the original $M$-dimensional vector $\boldsymbol{x}$ for better HGR performance. Supposing $L$ is the number of classes, the between-class scatter matrix $\boldsymbol{S}_{\mathrm{B}}$ can be defined as

$$\boldsymbol{S}_{\mathrm{B}} = \sum_{l=1}^{L} N_l[\boldsymbol{\mu}^{(l)} - \boldsymbol{\mu}][\boldsymbol{\mu}^{(l)} - \boldsymbol{\mu}]^{\mathrm{T}} \tag{3}$$

and a multi-class scatter matrix $\boldsymbol{S}_{\mathrm{W}}$ is further defined as

$$\boldsymbol{S}_{\mathrm{W}} = \sum_{l=1}^{L} \sum_{j=1}^{N_l} [x_j^{(l)} - \boldsymbol{\mu}^{(l)}] N_l [x_j^{(l)} - \boldsymbol{\mu}^{(l)}]^{\mathrm{T}} \tag{4}$$

where $N_l$ is the samples' number of the $l$th class. The centre-adjusted scatter matrix is represented as

$$\bar{S} = \bar{X}\bar{X}^{\mathrm{T}} = \sum_{i=1}^{N} [x_i - \boldsymbol{\mu}][x_i - \boldsymbol{\mu}] \tag{5}$$

### 4.1 Deflation-based orthogonal discriminant analysis

Assume $m$ components are expected to be extracted, the signal-to-noise ratio (SNR) pertaining to the $i$th component can be


*IET Comput. Vis.*, 2015, Vol. 9, Iss. 5, pp. 673–680

**Algorithm 1**

**Input:** training dataset [X, Y] containing the positive and the negative samples

**Output:** M × m transformation matrix: $\mathbf{W} = [\mathbf{w_1} \ldots \mathbf{w}_m]$, and the subspace vector $\mathbf{x} \rightarrow \mathbf{W}^{\mathrm{T}}\mathbf{x}$ presents a lower $m$-dimensional description of the original $M$-dimensional vector $\mathbf{x}$

1. Compute the between-class scatter matrix $\boldsymbol{S_B}$ **and the** multi-class scatter matrix $\mathbf{S}_W$ **with Eq. 3 and Eq. 4;**
2. Initialize $\mathbf{Q}^{(1)} \equiv (\mathbf{S}_W)^+$, $\mathrm{w}_i = \vec{o}$;

**For** i = 1: m

3. Compute $\mathrm{w}_i = v_i/\|v_i\|$, where $v_i = Q^{(i)}\,\Delta^{(i)}$;
4. Compute the deflation matrix: $\mathrm{D}^{(i)} = I - w_i w_i^T$
5. Update $\mathrm{Q}^{(i+1)} = D^{(i)}Q^{(i)}D^{(i)}$;
6. Update $\Delta^{(i+1)} = \Delta^{(i)} - w_i w_i^T\,\Delta^{(i)}$;

**End for**

**Return** $\mathbf{W} = [\mathbf{w_1} \ldots \mathbf{w}_m]$.

**Fig. 6** *Deflation-based orthogonal discriminant analysis*

defined as

$$\mathrm{SNR}_i = \frac{w_i^{\mathrm{T}}\boldsymbol{S}_{\mathrm{B}}\boldsymbol{w}_i}{w_i^{\mathrm{T}}\boldsymbol{S}_{\mathrm{W}}\boldsymbol{w}_i}, \quad i = 1, \ \ldots, \ m \tag{6}$$

Ideally, we would prefer maximising the sum-of-SNRs of all the components. More precisely, such optimiser aims at finding a solution for $\boldsymbol{W}_{\mathrm{SoSNR}}$

$$\boldsymbol{W}_{\mathrm{SoSNR}} = \underset{\boldsymbol{W}=[w_1 \ldots w_m]:W^{\mathrm{T}}\boldsymbol{W}=I}{\arg\max} \left( \sum_{i=1}^{m} \frac{w_i^{\mathrm{T}}\boldsymbol{S}_{\mathrm{B}}\boldsymbol{w}_i}{w_i^{\mathrm{T}}\boldsymbol{S}_{\mathrm{W}}\boldsymbol{w}_i} \right) \tag{7}$$

Unfortunately, this is an NP complete problem and the complexity for computing the SoSNR-optimal solution is heavy. For numerical efficiency, an approximated variant, the deflation-based orthogonal discriminant analysis aims at sequentially finding $\boldsymbol{w}_1$, $\boldsymbol{w}_2$, …, $\boldsymbol{w}_m$ such that

$$\text{maximise}_{\boldsymbol{w}_i} \frac{w_i^{\mathrm{T}}\boldsymbol{S}_{\mathrm{B}}\boldsymbol{w}_i}{w_i^{\mathrm{T}}\boldsymbol{S}_{\mathrm{W}}\boldsymbol{w}_i} \tag{8}$$

subject to $\boldsymbol{w}_i \perp \boldsymbol{w}_1, \ \ldots, \ i-1$

$$w_i^{\mathrm{T}}\boldsymbol{w}_i = 1, \quad \boldsymbol{w}_i \in \mathrm{range}(X)$$

The proposed method can be applied to either binary or multi-class scenario. Let us now focus on the proposed procedure for binary classification, that is, $L = 2$. Denote $\Delta = \boldsymbol{\mu}_+ - \boldsymbol{\mu}_-$, that is, the difference between the positive and negative centroids in the original space, the initial pseudo-inverse as $\boldsymbol{Q}^{(1)} \circ (\boldsymbol{S}_{\mathrm{W}})^+$. In this case, the optimiser in (8) can be efficiently solved by the following algorithm (see Fig. 6).

It is noted that this procedure enjoys an important merit in that no eigenvalue decomposition is required, making it computationally simpler than PCA and many others. Verifying that the procedure indeed yields an optimal solution for the criterion given in (8).

When $i = 1$, the first principle vector $\boldsymbol{w}_1$ can be computed. Then the deflation operator (step 4) removes $\boldsymbol{w}_1$ component from $\boldsymbol{Q}^{(1)}$, forcing Range $(Q^{(2)})$ to become orthogonal to $\boldsymbol{w}_1$. At the iteration $i = 2$, note that $\boldsymbol{w}_2$ is orthogonal to $\boldsymbol{w}_1$ because $\boldsymbol{w}_2 \in \mathrm{Range}(\mathrm{Q}^{(2)})$ Now, in this iteration, a new deflation operator (Eq. (11)) will further remove component from $Q^{(2)}$, forcing Range $(Q^{(3)})$ to become orthogonal to both $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$. By induction, it can be shown that $\boldsymbol{w}_1 \perp \boldsymbol{w}_2 \perp \ \cdot \ \perp \boldsymbol{w}_m$ and thus the proof. When SW is non-singular, such that

$$\boldsymbol{w}_1 \propto \boldsymbol{v}_1 = \left(S_{\mathrm{W}}^{(1)}\right)^+ \Delta^{(1)} = S_{\mathrm{W}}^{-1}\Delta^{(1)} \tag{9}$$

It is noted that Fisher discriminant analysis is the same as the first vector obtained with the proposed method. Compared with the recently proposed feature space discriminant analysis method SODA [55], our proposed method has the advantage both in computation complexity and discriminative ability. In each interaction of SODA, eigenvalue decomposition is used to obtain $\boldsymbol{w}_1$ and pseudo-inverse estimation is adopted in each deflation step. It means when dealing with high-dimension features, the computational complexity will be high and beyond tolerance in practice. To overcome this problem, we adopted a different strategy to obtain $\boldsymbol{w}_i$. First, the difference between the positive and negative centroids in the original space $\Delta$ is estimated; then in each interaction, we compute $\boldsymbol{v}_i$ with the updated $Q^{(i)}$, and finally obtain the normalised $\boldsymbol{w}_i$. It is noted that no eigenvalue decomposition is required in the proposed method. To enhance the discriminative ability of the transformation matrix $\boldsymbol{W}$, the difference between the positive and negative centroids in the original space $\Delta$ is updated in each interaction. In estimating the current $\boldsymbol{w}_i$, the pre-estimated discriminative component is removed from the original $\Delta$, eliminating the possible negative effects of the having obtained discriminative components.

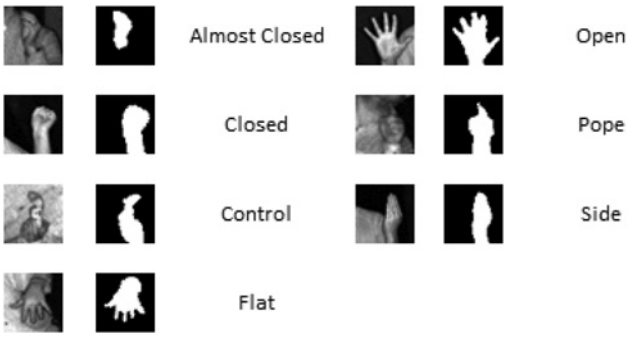


**Fig. 7** *Seven commonly used static hand gestures performed by the gesturers*

**Table 1** Recognition accuracy with RGB features (HOG, LBP and structure features)

| | Original | PCA | KPCA | LDA | KLDA | SODA [55] | MMLDA [56] | Proposed |
|---|---|---|---|---|---|---|---|---|
| AC | 0.53 | 0.56 | 0.59 | 0.48 | 0.5 | 0.57 | 0.57 | 0.61 |
| Close | 0.31 | 0.35 | 0.37 | 0.27 | 0.29 | 0.35 | 0.33 | 0.39 |
| Ctrl | 0.6 | 0.59 | 0.61 | 0.53 | 0.55 | 0.59 | 0.58 | 0.64 |
| Flat | 0.33 | 0.36 | 0.38 | 0.31 | 0.32 | 0.37 | 0.37 | 0.39 |
| Open | 0.34 | 0.35 | 0.38 | 0.32 | 0.35 | 0.39 | 0.36 | 0.42 |
| Pope | 0.4 | 0.42 | 0.41 | 0.35 | 0.37 | 0.44 | 0.42 | 0.45 |
| Side | 0.36 | 0.37 | 0.4 | 0.31 | 0.33 | 0.36 | 0.33 | 0.37 |
| Nega | 0.67 | 0.65 | 0.66 | 0.61 | 0.63 | 0.65 | 0.66 | 0.68 |
| Ave. | 0.44 | 0.46 | 0.48 | 0.4 | 0.43 | 0.47 | 0.45 | 0.49 |

## 5 Experiment result

In this section, extensive experimental results are presented to evaluate the proposed discriminative features learning in HGR. To analyse this problem in detail, both static and dynamic gesture recognitions are performed. All the experiments are performed in MATLAB 7.12.0 on a Dell OPTIPLEX computer with Intel(R) Core(TM) 2 Duo central processing unit E8400 processor (3.00 GHz), 3.25 G memory, and Windows 7 operating system. We first evaluate the performance on standard gesture recognition and further extend our method to compare with other state-of-the-art methods. Our experiments reveal that the proposed method gives superior recognition performance than many exiting approaches.

### 5.1 Recognition performance on standard static gestures

The experiment data contain 700 samples of seven commonly used static hand gestures performed 20 times by five performers. These gestures are performed in front of Kinect as shown in Fig. 7. In our training data, we classify gestures into seven gesture categories: almost closed, closed, control, flat, open, pope, and side. In the experiment, cross-validation is adopted, the data of four people being used for training, the data of the left one used as unregistered test set.

We compare the classification results with original space, PCA, kernel PCA, successively orthogonal discriminant analysis (SODA), and median–mean line-based discriminant analysis (MMLDA). The MMLDA is proposed in [56], alleviating the negative effect on the class mean caused by outliers with introducing the median–mean line as an adaptive class-prototype. The classifier we used is support vector machine (SVM) [57] with rbf kernel ($\sigma = 1$). We also compared the classification results with LDA and its kernelised variance kernel linear discriminant analysis (KLDA), such a kernel-based classifier is formally derived by optimising a safety margin measured by the distance metric defined in the kernel-induced intrinsic space [58]. The parameter $\sigma$ for the kernel-based methods is set to be 0.5 for consistent and fair comparison. The reduced dimensionality for the feature reduction techniques under comparison is $k = 4$. This is chosen based on cross-validation on the dataset and then applied to the rest of the datasets. The recognition comparison with features from RGB images, depth images, and both RGB images and depth images are shown in the following three tables. The results shown are based on the averaged error rate of one-versus-one scheme for all classes from the datasets.

As shown in Tables 1, 2 and 3, in the case of RGB information or depth information explored alone, the result is not satisfied. If the features both from RGB images and depth images are extracted, the static gesture recognition performance increased significantly. Obviously, RGB and depth information collaborate with each other in the HGR task, and with the adoption of these features, the HGR performance can be improved significantly.

Compare with the other feature space discriminant analysis methods, the proposed deflation-based orthogonal discriminant analysis achieves the best classification results. If the original data is adopted directly, it is obvious that the original space of data suffers from over fitting using SVM with rbf kernel, which results in a 59% error probability. In such scenarios, PCA/kernel PCA (KPCA) with extremely low dimensionality will do an even better job than the original space. LDA outperforms SVM on original space in all cases. Since the parameter selection of kernel SVM is

**Table 2** Recognition accuracy with depth features (HOG, LBP and structure features)

| | Original | PCA | KPCA | LDA | KLDA | SODA [55] | MMLDA [56] | Proposed |
|---|---|---|---|---|---|---|---|---|
| AC | 0.36 | 0.38 | 0.4 | 0.31 | 0.33 | 0.37 | 0.36 | 0.39 |
| Close | 0.2 | 0.22 | 0.24 | 0.19 | 0.19 | 0.23 | 0.22 | 0.25 |
| Ctrl | 0.62 | 0.64 | 0.66 | 0.56 | 0.58 | 0.58 | 0.55 | 0.59 |
| Flat | 0.36 | 0.38 | 0.39 | 0.31 | 0.34 | 0.56 | 0.55 | 0.59 |
| Open | 0.37 | 0.39 | 0.42 | 0.32 | 0.33 | 0.45 | 0.43 | 0.5 |
| Pope | 0.42 | 0.45 | 0.48 | 0.38 | 0.4 | 0.44 | 0.44 | 0.48 |
| Side | 0.32 | 0.35 | 0.37 | 0.3 | 0.33 | 0.35 | 0.33 | 0.39 |
| Nega | 0.61 | 0.63 | 0.65 | 0.55 | 0.58 | 0.64 | 0.64 | 0.69 |
| Ave. | 0.41 | 0.43 | 0.45 | 0.37 | 0.39 | 0.45 | 0.44 | 0.48 |

**Table 3** Recognition accuracy with RGB and depth features

| | Original | PCA | KPCA | LDA | KLDA | SODA [55] | MMLDA [56] | Proposed |
|---|---|---|---|---|---|---|---|---|
| AC | 0.71 | 0.73 | 0.75 | 0.67 | 0.69 | 0.72 | 0.7 | 0.74 |
| Close | 0.48 | 0.49 | 0.51 | 0.44 | 0.45 | 0.54 | 0.52 | 0.56 |
| Ctrl | 0.88 | 0.89 | 0.91 | 0.82 | 0.83 | 0.9 | 0.89 | 0.91 |
| Flat | 0.42 | 0.44 | 0.45 | 0.38 | 0.39 | 0.5 | 0.48 | 0.52 |
| Open | 0.51 | 0.52 | 0.55 | 0.45 | 0.47 | 0.55 | 0.53 | 0.57 |
| Pope | 0.52 | 0.55 | 0.56 | 0.49 | 0.51 | 0.56 | 0.53 | 0.58 |
| Side | 0.39 | 0.43 | 0.45 | 0.37 | 0.37 | 0.43 | 0.42 | 0.45 |
| Nega | 0.78 | 0.79 | 0.79 | 0.74 | 0.75 | 0.81 | 0.81 | 0.84 |
| Ave. | 0.59 | 0.61 | 0.62 | 0.55 | 0.56 | 0.63 | 0.61 | 0.65 |

**Table 4** Performance comparison on the 20 development data batches

| Methods | Extend-MHI EA [60] | Manifold LSR EA [61] | 3DHOF+GHOG EA [43] | Motion history EA [49] | Proposed EK | Proposed EA |
|---|---|---|---|---|---|---|
| TeLev | 0.26 | 0.28 | 0.43 | 0.31 | 0.3 | 0.25 |

TeLev is the sum of the Levenshtein distance divided by the true number of gestures

a key for high performance, LDA enjoys the advantage of simplicity. By obtaining an orthogonal transformation matrix which maps the features to a new low-dimensional feature space, SODA achieves better performance than LDA and MMLDA. By releasing the restriction of 1D subspace, the proposed method achieves more flexibility than LDA, and by evading pseudo-inverse procedure and updating the difference between the positive and negative centroids, the proposed method has advantages both in computation complexity and discriminative ability. Compared with these methods, the proposed method achieves the state-of-the-art performance.

### 5.2 Comparison with other dynamic HGR methods

In this section, extensive experimental results are presented to compare with other state-of-the-art methods of dynamic HGR. ChaLearn Gesture Dataset (CGD2011) is used in the experiments [59]. CGD2011 is the largest gestures dataset recorded with Kinect, which consists of 50 000 dynamic gestures (grouped in 500 batches, each batch including 47 sequences and each sequence containing of 1–5 gestures drawn from one of 30 small gesture vocabularies of 8–15 gestures), with frame size $240 \times 320$, 10 fps, recorded by 20 different users. The proposed view point independent gesture recognition is compared by average performance with other recent representative methods on the first 20 development data batches. On the basis of the proposed discriminating features extraction, two strategies are adopted.

- *Extracting features from key frame (EK):* In a dynamic hand gesture sequence, the key frame, that is, the frame that has the minimal relative motion quantity is more discriminative than the other frames. First, the features are extracted from the key frames in the reference and the test gesture sequences, and then the distance between the key frames is calculated with the proposed method in Section 4.1. The test gesture sequence is classified as the gesture whose key frame has the smallest distance with the key frame of test gesture sequence.
- *Extracting features from all frames (EA):* We extract the discriminating feature form each frame in the gesture sequence, and self-organisation feature map (SOFM)/HMM [62] is adopted as the classification model. Here, SOFM/HMM is a three-state left-to-right model allowing possible skips and the covariance matrix is a diagonal matrix with all diagonal elements being 0.2. The comparison results are reported in Table 4.

In the experiments, Levenshtein distance is used to evaluate the HGR performance, which is also used in CHALEARN gesture challenge. Levenshtein distance is the minimum number of edit operations (substitution, insertion or deletion) that have to perform from one sequence to another (or vice versa).

The proposed deflation-based orthogonal discriminant analysis is applied to overcome the limitation of LDA, aiming at finding a projection vector such that the projected values of data from both classes have maximum class separability. In LDA, only one such vector can be found due to the rank deficiency for binary classification problems. The proposed deflation-based orthogonal discriminant analysis attempts to obtain a transformation matrix instead of a vector. We compare the proposed approaches with some popular gesture matching methods as shown in Table 4, including the extended motion history image and maximum correlation coefficient (extended-MHI) [60], non-linear regression framework on manifolds (manifold least squares regression (LSR)) [61], the motion history-based silhouettes and Euclidean distance-based classifiers (motion history) [49]. Our proposed EA method achieves the best performance in the comparison. Compare with the features adopted with the other methods, relative simple features are used in our method, that is, HOG, LBP, structure and 3D-voxel features. With the proposed deflation-based orthogonal discriminant analysis, these features are projected into a subspace with more discriminative ability.

It is noted that the compared other methods extracted features from all the frames. By extracting the discriminating features only from the key frames, our proposed EK method further reduces the computation complexity and achieves satisfying results. Compared with the DHOF and GHOG [43], the proposed EK method achieves significantly better performance. The final dynamic gesture recognition performance depends not only on the discriminative features extraction, but also on the classifier adopted. Both extended-MHI [60] and manifold LSR [61] adopted more complex classifiers. In the case of motion history feature, if Euclidean distance-based classifiers is adopted [49], our prosed EK method can also get better performance. In extended-MHI, the similar motion history feature is used, whereas a correlation coefficient method with higher computation complexity is adopted, leading to a higher accuracy than the proposed EK. From these comparison results, we can see that the proposed EK method has relatively high recognition accuracy while low computational complexity. The proposed methods achieve about 15 and 11 fps for EK and EA, respectively, which is faster than the video recording speed (10 fps) of CGD 2011. Besides, our work indicates that the performance of gesture recognition can be significantly improved by the adoption of deflation-based orthogonal discriminant analysis, which will inspire other researchers in this field to develop HGR along this direction.

## 6 Conclusion

In this paper, a discriminating features extraction strategy for HGR is proposed. Both RGB images and depth images are explored; more specifically, HOG feature, LBP feature, structure feature and 3D-voxel feature are first extracted from RGB images and depth images, then a novel deflation-based orthogonal discriminant analysis is explored for further feature reduction and enhancing the discriminative ability, which allows high flexibility than LDA. The proposed method has high recognition accuracy while low computational complexity, and outperforms several state-of-the-art methods in gesture recognition.

## 7 Acknowledgments

This work was supported in part by the Major State Basic Research Development Program of China (973 Program 2015CB351804) and the National Natural Science Foundation of China under grant nos. 61272386, 61100096 and 61300111.

## 8 References

1 Mitra, S., Acharya, T.: 'Gesture recognition: a survey', *IEEE Trans. Syst. Man Cybern. C, Appl. Rev.*, 2007, **37**, (3), pp. 311–324
2 Vogler, C., Metaxas, D.: 'Parallel hidden Markov models for American sign language recognition'. Proc. 7th IEEE Int. Conf. on Computer Vision, 1999, vol. 1, pp. 116–122

3 Baklouti, M., Monacelli, E., Guitteny, V., Couvet, S.: 'Intelligent assistive exoskeleton with vision based interface'. Proc. of the Sixth Int. Conf. on Smart Homes and Health Telematics, 2008, vol. 5120, pp. 123–135
4 Nickel, K., Stiefelhagen, R.: 'Visual recognition of pointing gestures for human-robot interaction', *Image Vis. Comput.*, 2007, **25**, (12), pp. 1875–1884
5 Wachs, J., Kölsch, M., Stern, H., Edan, Y.: 'Vision-based hand-gesture applications', *Commun. ACM*, 2011, **54**, (2), pp. 60–71
6 Kadous, M.W.: 'Machine recognition of Auslan signs using Powergloves: toward large-lexicon recognition of sign language'. Proc. Workshop Integration Gesture Language Speech, 1996, pp. 165–174
7 Fels, S.S., Hinton, G.E.: 'Glove-talk: a neural network interface between a data-glove and a speech synthesizer', *IEEE Trans. Neural Netw.*, 1993, **4**, (1), pp. 2–8
8 Kim, J.S., Jang, W., Bien, Z.: 'A dynamic gesture recognition system for the Korean sign language (KSL)', *IEEE Trans. Syst. Man Cybern. B*, 1996, **26**, (4), pp. 354–359
9 Fang, G., Gao, W., Zhao, D.: 'Large-vocabulary continuous sign language recognition based on transition-movement models', *IEEE Trans. Syst. Man Cybern. A, Syst. Hum.*, 2007, **37**, (1), pp. 1–9
10 Jiang, F., Gao, W., Yao, H., Zhao, D., Chen, X.: 'Synthetic data generation technique in signer-independent sign language recognition', *Pattern Recognit. Lett.*, 2009, **30**, (5), pp. 513–524
11 Cooper, H., Holt, B., Bowden, R.: 'Sign language recognition'. Looking at People: Automatic visual analysis of humans, Part D
12 Dala, N., Triggs, B.: 'Histograms of oriented gradients for human detection'. CVPR, 2005
13 Ojala, T., Pietikäinen, M., Harwood, D.: 'A comparative study of texture measures with classification based on featured distribution', *Pattern Recognit.*, 1996, **29**, (1), pp. 51–59
14 Ong, S.C.W., Ranganath, S.: 'Automatic sign language analysis: a survey and the future beyond lexical meaning', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2005, **27**, (6), pp. 873–891
15 Stenger, B.: 'Template based hand pose recognition using multiple cues'. Proc. Seventh Asian Conf. on Computer Vision: ACCV, 2006
16 Bretzner, L., Laptev, I., Lindeberg, T.: 'Hand gesture recognition using multi scale colour features, hierarchical models and particle filtering'. Proc. 5th IEEE Int. Conf. on Automatic Face and Gesture Recognition, Washington, DC, USA, 2002, pp. 423–428
17 Argyros, A., Lourakis, M.: 'Vision-based interpretation of hand gestures for remote control of a computer mouse'. Proc. of the Workshop on Computer Human Interaction, 2006, pp. 40–51
18 Starner, T., Weaver, J., Pentland, A.: 'Real-time American sign language recognition using desk and wearable computer based video', *IEEE Trans. Pattern Anal. Mach. Intell.*, 1998, **20**, (12), pp. 1371–1375
19 Wang, C., Wang, K.: 'Hand gesture recognition using AdaBoost with SIFT for human robot interaction' (Springer, Berlin, 2008), pp. 317–329, ISSN 0170-8643
20 Barczak, A., Dadgostar, F.: 'Real-time hand tracking using a set of co-operative classifiers based on Haar-like features', *Res. Lett. Inf. Math. Sci.*, 2005, **7**, pp. 29–42
21 Chen, Q., Georganas, N., Petriu, E.: 'Real-time vision-based hand gesture recognition using Haar-like features'. IEEE Instrumentation and Measurement Technology Conf. Proc., IMTC, 2007
22 Wikipedia. Kinect – Wikipedia, the free encyclopedia. Available at http://www.en.wikipedia.org/wiki/Kinect, 2012, accessed 24 September 2012
23 Oikonomidis, I., Kyriazis, N., Argyros, A.A.: 'Efficient model-based 3D tracking of hand articulations using Kinect'. Proc. of BMVC, Dundee, September 2011
24 Ershaed, H., Al-Alali, I., Khasawneh, N., Fraiwan, M.: 'An Arabic sign language computer interface using the Xbox Kinect'. Annual Undergraduate Research Conf. on Applied Computing, Dubai, UAE, May 2011
25 Dardas, N.: 'Real-time hand gesture detection and recognition for human computer interaction'. PhD thesis, University of Ottawa, 2012
26 Awad, G., Han, J., Sutherland, A.: 'A unified system for segmentation and tracking of face and hands in sign language recognition'. Proc. of the 18th Int. Conf. on Pattern Recognition, 2006, vol. 1, pp. 239–242
27 Maraqa, M., Abu-Zaiter, R.: 'Recognition of Arabic sign language (ArSL) using recurrent neural networks'. Proc. of the First Int. Conf. on the Applications of Digital Information and Web Technologies, 2008, pp. 478–481
28 Ramamoorthy, A., Vaswani, N., Chaudhury, S., Banerjee, S.: 'Recognition of dynamic hand gestures', *Pattern Recognit.*, 2003, **36**, (9), pp. 2069–2081
29 Kadir, T., Bowden, R., Ong, E.J., Zisserman, A.: 'Minimal training, large lexicon, unconstrained sign language recognition'. Proc. of the British Machine Vision Conf., 2004, vol. 1, pp. 1–10
30 Cutler, R., Turk, M.: 'View-based interpretation of real-time optical flow for gesture recognition'. Proc. of the 10th IEEE Int. Conf. and Workshops on Automatic Face and Gesture Recognition, 1998, p. 416
31 Mahbub, U., Roy, T., Rahman, M.S., Imtiaz, H.: 'One-shot-learning gesture recognition using motion history based gesture silhouettes'. Proc. of the Int. Conf. on Industrial Application Engineering, 2013, pp. 186–193
32 Zhang, L.-G., Chen, Y., Fang, G., Chen, X., Gao, W.: 'A vision-based sign language recognition system using tied-mixture density HMM'. Proc. of the Sixth Int. Conf. on Multimodal Interfaces, 2004, pp. 198–204
33 Stergiopoulou, E., Papamarkos, N.: 'Hand gesture recognition using a neural network shape fitting technique', *Eng. Appl. Artif. Intell.*, 2009, **22**, (8), pp. 1141–1158
34 Maung, T.H.H.: 'Real-time hand tracking and gesture recognition system using neural networks', *World Acad. Sci. Eng. and Technol.*, 2009, **50**, pp. 466–470
35 Agrawal, T., Chaudhuri, S.: 'Gesture recognition using motion histogram'. Proc. of the Indian National Conf. of Communications, 2003, pp. 438–442
36 Shao, L., Ji, L.: 'Motion histogram analysis based key frame extraction for human action/activity representation'. Proc. of Canadian Conf. on Computer and Robot Vision, 2009, pp. 88–92
37 Zahedi, M., Keysers, D., Ney, H.: 'Appearance-based recognition of words in American sign language', *Pattern Recognit. Image Anal.*, 2005, pp. 511–519
38 Wong, S.-F., Kim, T.-K., Cipolla, R.: 'Learning motion categories using both semantic and structural information'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, 2007, pp. 1–6
39 Cooper, H., Holt, B., Bowden, R.: 'Sign language recognition'. In Visual Analysis of Humans, 2011, pp. 539–562
40 Rauschert, I., Agrawal, P., Sharma, R., Fuhrmann, S., Brewer, I., MacEachren, A.: 'Designing a human-centered, multimodal GIS interface to support emergency management'. Proc. of the 10th ACM Int. Symp. on Advances in Geographic Information Systems, 2002, pp. 119–124
41 Ren, Z., Yuan, J., Meng, J., Zhang, Z.: 'Robust part-based hand gesture recognition using Kinect sensor', *IEEE Trans. Multimedia*, 2013, **15**, (5), pp. 1110–1120
42 Wan, J., Ruan, Q., Li, W., Deng, S.: 'One-shot learning gesture recognition from RGB-D data using bag of features', *J. Mach. Learn. Res.*, 2013, **14**, (1), pp. 2549–2582
43 Fanello, S.R., Gori, I., Metta, G., Odone, F.: 'One-shot learning for real-time action recognition', *Pattern Recognit. Image Anal.*, 2013, **7887**, pp. 31–40
44 Fukumizu, K., Bach, F.R., Jordan, M.I.: 'Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces', *J. Mach. Learn. Res.*, 2004, **5**, (1), pp. 73–99
45 Hastie, T., Tibshirani, R., Friedman, J.: 'The elements of statistical learning: data mining, inference, and prediction' (Springer, New York, NY, 2009, 2nd edn.)
46 Guyon, I., Elisseeff, A.: 'An introduction to variable and feature selection', *J. Mach. Learn. Res.*, 2003, **3**, pp. 1157–1182
47 Pearson, K.: 'On lines and planes of closest fit to systems of points in space', *Philos. Mag. Ser.*, 1901, **6**, (2), pp. 559–572
48 Hotelling, H.: 'Analysis of a complex of statistical variables into principal components', *J. Educ. Psychol.*, 1933, **24**, pp. 498–520
49 Mahbub, U., Roy, T., Rahman, M.S., Imtiaz, H.: 'One-shot-learning gesture recognition using motion history based gesture silhouettes'. Int. Conf. on Industrial Application Engineering, 2013, pp. 186–193
50 Jolliffe, I.T.: 'Principal component analysis. Series: springer series in statistics' (Springer, New York, NY, 2002, 2nd edn.)
51 Golub, G., Van Loan, C.F.: 'Matrix computations' (Johns Hopkins University Press, Baltimore, MD, 1996, 3rd edn.)
52 Golub, G.H., Kahan, W.: 'Calculating the singular values and pseudo-inverse of a matrix', *J. Soc. Ind. Appl. Math. Ser. B, Numer. Anal.*, 1965, **2**, (2), pp. 205–224
53 Armstrong, S.A., Staunton, J.E., Silverman, L.B., *et al.*: 'MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia', *Nat. Gen.*, 2002, **30**, (1), pp. 41–47
54 Aronszajn, N.: 'Theory of reproducing kernels', *Trans. Am. Math. Soc.*, 1950, **68**, pp. 337–404
55 Yu, Y., Mckelvey, T., Kung, S.Y.: 'A classification scheme for 'high-dimensional-small-sample-size' data using SODA and ridge-SVM with microwave measurement applications'. Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, Canada, May 2013
56 Xu, J., Yang, J., Gu, Z., Zhang, N.: 'Median–mean line based discriminant analysis', *Neurocomputing*, **123**, (10), pp. 233–246
57 Mavroforakis, M., Theodoridis, S.A.: 'Geometric approach to s ne (SVM) classification', *IEEE Trans. Neural Netw.*, 2006, **17**, (3), pp. 671–683
58 Kung, S.Y.: 'Kernel method in machine learning' (Cambridge University Press, Cambridge, 2014)
59 Guyon, I., Athitsos, V., Jangyodsuk, P., Escalante, H.J.: 'The ChaLearn gesture dataset (CGD 2011)', *Mach. Vis. Appl.*, 2014, **25**, (8), pp. 1929–1951
60 Wu, D., Zhu, F., Shao, L.: 'One shot learning gesture recognition from RGBD images'. IEEE Computer Society Conf. on CVPRW, 2012, pp. 7–12
61 Lui, Y.M.: 'A least squares regression framework on manifolds and its application to gesture recognition'. IEEE Computer Society Conf. on CVPRW, 2012, pp. 13–18
62 Fang, G., Gao, W., Zhao, D.: 'Large vocabulary sign language recognition based on fuzzy decision trees', *IEEE Trans. Syst. Man Cybern. A, Syst. Hum.*, 2004, **34**, (3), pp. 539–562