Kernel analysis on Grassmann manifolds for action recognition

Mehrtash T. Harandi [a,*], Conrad Sanderson [b], Sareh Shirazi [c], Brian C. Lovell [d]

[a] NICTA, Locked Bag 8001, Canberra ACT 2601, Australia and The University of Queensland, School of ITEE, QLD 4072, Australia
[b] NICTA, PO Box 6020, St Lucia QLD 4067, Australia and Queensland University of Technology (QUT), Brisbane QLD 4000, Australia
[c] NICTA, PO Box 6020, St Lucia QLD 4067, Australia and The University of Queensland, School of ITEE, QLD 4072, Australia
[d] The University of Queensland, School of ITEE, QLD 4072, Australia

A B S T R A C T

Modelling video sequences by subspaces has recently shown promise for recognising human actions. Subspaces are able to accommodate the effects of various image variations and can capture the dynamic properties of actions. Subspaces form a non-Euclidean and curved Riemannian manifold known as a Grassmann manifold. Inference on manifold spaces usually is achieved by embedding the manifolds in higher dimensional Euclidean spaces. In this paper, we instead propose to embed the Grassmann manifolds into reproducing kernel Hilbert spaces and then tackle the problem of discriminant analysis on such manifolds. To achieve efficient machinery, we propose graph-based local discriminant analysis that utilises within-class and between-class similarity graphs to characterise intra-class compactness and inter-class separability, respectively. Experiments on KTH, UCF Sports, and Ballet datasets show that the proposed approach obtains marked improvements in discrimination accuracy in comparison to several state-of-the-art methods, such as the kernel version of affine hull image-set distance, tensor canonical correlation analysis, spatial-temporal words and hierarchy of discriminative space-time neighbourhood features.

1. Introduction

The goal of human action recognition is to automatically analyse and recognise what action is being undertaken in a given video, with one or more persons performing an action. Applications include content-based video analysis, security and surveillance, human–computer interaction, and animation (Turaga et al., 2008; Weinland et al., 2011). Subspace-based approaches, which are able to accommodate the effects of a wide range of image variations, have recently shown promising results for action recognition (Kim and Cipolla, 2009; Turaga and Chellappa, 2009; O'Hara et al., 2012). Moreover, subspaces can also capture the dynamic properties of videos (Turaga et al., 2011).

Subspaces form non-Euclidean and curved Riemannian manifolds known as Grassmann manifolds, allowing a video or an image-set to be conveniently represented as a point on a Grassmann manifold. Recent studies show that better performance can be achieved when the geometry of Riemannian spaces is explicitly considered (Hamm and Lee, 2008; Tuzel et al., 2008; Subbarao and Meer, 2009; Lui, 2012; Harandi et al., 2011; Turaga et al., 2011).

Inference on manifold spaces can be achieved by embedding the manifolds in higher dimensional Euclidean spaces, which can be considered as flattening the manifolds. A popular choice for embedding manifolds is through considering tangent spaces (Tuzel et al., 2008; Turaga et al., 2011). Two notable examples are the pedestrian detection system by Tuzel et al. (2008) and non-linear mean shift (Comaniciu and Meer, 2002) by Subbarao and Meer (2009). Nevertheless, flattening manifolds through tangent spaces is not without drawbacks. For example, the distance on a tangent space between two arbitrary points is generally not equal to the true geodesic distance,[1] which may lead to inaccurate modelling (Harandi et al., 2012). An alternative school of thought omits the use of tangent spaces and instead embeds Grassmann manifolds into Reproducing Kernel Hilbert Spaces (RKHS) (Shawe-Taylor and Cristianini, 2004) through dedicated Grassmann kernels (Hamm and Lee, 2008; Harandi et al., 2011; Shirazi et al., 2012). This in turn opens the door for employing many kernel-based machine learning algorithms (Shawe-Taylor and Cristianini, 2004).

Inference via discriminant analysis (DA) on Grassmann manifolds has been recently explored (Hamm and Lee, 2008; Wang and Shi, 2009). Given subspaces that are represented as points on a Grassmann manifold, Grassmann discriminant analysis (GDA) maps them to RKHS, such that a measure of discriminatory power in the induced RKHS is maximised. While GDA has shown promising results in (Hamm and Lee, 2008; Wang and Shi, 2009), the

* Corresponding author at: NICTA, Locked Bag 8001, Canberra ACT 2601, Australia and The University of Queensland, School of ITEE, QLD 4072, Australia. Tel.: +61 733008673.
E-mail address: mehrtash.harandi@nicta.com.au (M.T. Harandi).

[1] The geodesic distance takes into account the curvature of manifolds; an example is the distance between two points on a sphere.

Fig. 1. (a) Examples of a hand-waving action. (b) Basis vectors for a subspace of order three, modelling the entire action; the subspace is a point on a Grassmann manifold.

conventional formalism of DA suffers from not being able to take into account the local structure of data (Chen et al., 2007). For example, multi-modal classes and outliers can adversely affect the discrimination and/or generalisation ability of models based on conventional DA.

Contributions. In this work[2] we first present two methods of representing human actions on Grassmann manifolds. For the purposes of action recognition, we then extend our preliminary study on an enhanced form of GDA (Harandi et al., 2011), based on Grassmann kernels and a graph-embedding framework (Yan et al., 2007). We also show that conventional GDA (Hamm and Lee, 2008) can be seen as a special case of the enhanced graph-embedding based approach. Thorough experiments on the KTH (Schuldt et al., 2004), UCF Sports (Rodriguez et al., 2008) and Ballet (Wang and Mori, 2009) datasets, which include various realistic challenges such as background clutter, partial occlusion, changes in viewpoint, scale and illumination, and complexity of motion, show that the proposed Grassmann graph-embedding discriminant analysis (GGDA) approach obtains notable improvements in discrimination accuracy in comparison to several state-of-the-art methods. This includes the original GDA (Hamm and Lee, 2008), kernel version of affine hull image-set distance (Cevikalp and Triggs, 2010), tensor canonical correlation analysis (Kim and Cipolla, 2009), spatial-temporal words (Niebles et al., 2008) and hierarchy of discriminative space-time neighbourhood features (Kovashka and Grauman, 2010).

We continue the paper as follows. Section 2 presents various ways of representing action videos by linear subspaces. Section 3 reviews Grassmann geometry, which leads to Section 4 which presents the Grassmann graph-embedding discriminant analysis approach. In Section 5 we compare the performance of the proposed method with previous approaches on several datasets. The main findings and possible future directions are summarised in Section 6.

2. Modelling actions by linear subspaces

Let us define a video as an ordered collection of images with time-stamps (temporal information), and an image-set as an orderless collection of images. Actions can be modelled as linear subspaces through image-sets, or through linear dynamic systems that take into account the temporal information. We overview both methodologies in the following subsections.

2.1. Modelling of image-sets

An image-set $\mathbb{F} = \{\boldsymbol{f}_i\}_{i=1}^{N}; \boldsymbol{f}_i \in \mathbb{R}^n$, where $\boldsymbol{f}_i$ is the vectorised representation of frame i, can be represented as a subspace (and hence as a point on a Grassmann manifold) through any orthogonalisation procedure like Singular Value Decomposition (SVD). More specifically, let $\mathbb{F} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$ be the SVD of $\mathbb{F}$. The first p columns of $\boldsymbol{U}$ represent an optimised subspace of order p (in the mean square sense) for $\mathbb{F}$ and can be seen as a point on manifold $\mathcal{G}n,p$. Intuitively, modelling an action by a subspace as described here can be understood as a low dimensional and compact representation by a set of basis vectors, in which the appearance of action could be effectively reconstructed by linearly combining the basis vectors. See Fig. 1 for an example of modelling a hand-waving action sequence by a subspace of order three.

Modelling image-sets by linear subspaces has been shown to deliver improved performance in the presence of practical issues such as misalignment as well as variations in pose and illumination (Wolf and Shashua, 2003; Hamm and Lee, 2008; Harandi et al., 2011). Modelling of actions by image-sets can be sufficient provided that the order in which the action is performed is not very relevant to decision making. While this assumption sounds restrictive, in many practical situations this might indeed be the case. As an example, it is possible to differentiate riding a horse from jogging without having temporal information. Nevertheless, a recent study (Li et al., 2011) shows that an extended type of image-set, obtained through a block Hankel matrix formalism, can capture the temporal information.

2.2. Modelling of linear dynamic systems

A video can be represented by an Auto Regressive and Moving Average (ARMA) model to explicitly take into account dynamics and temporal information. More specifically, a set of ordered images $\{\boldsymbol{f}(t)\}_{t=1}^{\tau}; \boldsymbol{f}(t) \in \mathbb{R}^n$ can be seen as the output of an ARMA process as:

$$\boldsymbol{f}(t) = \boldsymbol{C}\boldsymbol{z}(t) + \boldsymbol{w}(t), \quad \boldsymbol{w}(t) \sim \mathcal{N}(0, \boldsymbol{R}) \tag{1}$$

$$\boldsymbol{z}(t+1) = \boldsymbol{A}\boldsymbol{z}(t) + \boldsymbol{v}(t), \quad \boldsymbol{v}(t) \sim \mathcal{N}(0, \boldsymbol{Q}) \tag{2}$$

In (1) and (2), $\boldsymbol{z}(t) \in \mathbb{R}^p$ is the latent state vector at time t, $\boldsymbol{A} \in \mathbb{R}^{p\times p}$ and $\boldsymbol{C} \in \mathbb{R}^{n\times p}$ are the transition and measurement matrices, respectively, while $\boldsymbol{w}$ and $\boldsymbol{v}$ are noise components modelled as normal distributions with zero mean and covariance matrices $\boldsymbol{R} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{Q} \in \mathbb{R}^{p\times p}$, respectively. The order of the system is given by p, while n is the number of features in a frame of the sequence. Loosely speaking, one advantage of the ARMA model is that it decouples the appearance of the spatio-temporal data (modelled by $\boldsymbol{C}$) from the dynamics (represented by $\boldsymbol{A}$).

To estimate the transition and measurement matrices (Turaga et al., 2011), we define $\boldsymbol{F}_\tau = [\boldsymbol{f}(1)|\boldsymbol{f}(2)|\cdots|\boldsymbol{f}(\tau)]$, where the symbol $|$ denotes horizontal concatenation of vectors, as the feature matrix for time indices $1, 2, \ldots, \tau$. The estimated transition ($\widehat{\boldsymbol{A}}$) and measurement ($\widehat{\boldsymbol{C}}$) matrices can then be obtained via the SVD of $\boldsymbol{F}_\tau = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$, as follows:

$$\widehat{\boldsymbol{A}} = \boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{D}_1\boldsymbol{V}(\boldsymbol{V}^T\boldsymbol{D}_2\boldsymbol{V})^{-1}\boldsymbol{\Sigma}^{-1} \tag{3}$$

$$\widehat{\boldsymbol{C}} = \boldsymbol{U} \tag{4}$$

where

$$\boldsymbol{D}_1 = \begin{bmatrix} \boldsymbol{0}_{\tau-1}^T & 0 \\ \boldsymbol{I}_{(\tau-1)\times(\tau-1)} & \boldsymbol{0}_{\tau-1} \end{bmatrix} \quad \text{and} \quad \boldsymbol{D}_2 = \begin{bmatrix} \boldsymbol{I}_{(\tau-1)\times(\tau-1)} & \boldsymbol{0}_{\tau-1} \\ \boldsymbol{0}_{\tau-1}^T, & 0 \end{bmatrix}$$

ARMA models can be compared based on the subspace angles between the column-spaces of their observability matrices (Cock and Moor, 2002). The extended observability matrix of an ARMA model is given by $\boldsymbol{O}_\infty = [\boldsymbol{C}^T|(\boldsymbol{C}\boldsymbol{A})^T|(\boldsymbol{C}\boldsymbol{A}^2)^T|\cdots|(\boldsymbol{C}\boldsymbol{A}^n)^T|\cdots]^T$ and is

[2] This work is somewhat related to (Shirazi et al., 2012), where the problem of clustering on Grassmann manifolds is explored. In the method presented here, clustering is not performed.

usually approximated by the finite observability matrix $\boldsymbol{O}_p = [\boldsymbol{C}^T|(\boldsymbol{CA})^T|(\boldsymbol{CA}^2)^T|\cdots|(\boldsymbol{CA}^{(p-1)})^T]^T$ (Turaga et al., 2011).

To represent a video on a Grassmann manifold, the finite observability matrix of the ARMA model is estimated as described above. The subspace spanned by the columns of $\boldsymbol{O}_p$ (obtained by SVD or any other orthogonalisation procedure) corresponds to a point on the Grassmann manifold $\mathcal{G}n,p$.

3. Grassmann geometry

Without delving too deeply into differential geometry and related topics, a Riemannian manifold $\mathcal{M}$ is a differentiable and smooth manifold, endowed with a Riemannian metric that allows us to extend the notion of lengths and angles from familiar Euclidean spaces to the curved and non-flat space of $\mathcal{M}$. The geodesic distance between two points $\boldsymbol{X}, \boldsymbol{Y} \in \mathcal{M}$, denoted by $d_g(\boldsymbol{X}, \boldsymbol{Y})$, is defined as the minimum length over all possible smooth curves between $\boldsymbol{X}$ and $\boldsymbol{Y}$. A geodesic curve is a curve that locally minimises the distance between points.

Subspaces form a special class of Riemannian manifolds known as Grassmann manifolds. Formally, Grassmann manifolds are defined as quotient spaces of orthogonal group[3] $\mathbb{O}(n)$. A quotient space[4] of a manifold, intuitively speaking, is the result of "gluing together" certain points of the manifold.

Definition 1. Grassmann manifold $\mathcal{G}n,p$ is a quotient space of the orthogonal group $\mathbb{O}(n)$ and is defined as the set of p-dimensional linear subspaces of $\mathbb{R}^n$. Points on the Grassmann manifold are equivalence classes of $n \times p,\ p < n$ orthogonal matrices, where two matrices are equivalent if their columns span the same p-dimensional subspace.

In practice an element $\boldsymbol{X}$ of $\mathcal{G}n,p$ is represented by an orthonormal basis as a $n \times p$ matrix, i.e., $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}_p$. The geodesic distance between two points on the Grassmann manifold can be computed as:

$$d_G(\boldsymbol{X}, \boldsymbol{Y}) = \|\Theta\|_2 \tag{5}$$

where $\Theta = [\theta_1, \theta_2, \ldots, \theta_p]$ is the principal angle vector, i.e.:

$$\cos(\theta_k) = \max_{\boldsymbol{x}\in\boldsymbol{X}, \boldsymbol{y}\in\boldsymbol{Y}} \boldsymbol{x}^T\boldsymbol{y} = \boldsymbol{x}_k^T\boldsymbol{y}_k \tag{6}$$

$$\text{s.t.}: \quad \|\boldsymbol{x}\|_2 = \|\boldsymbol{y}\|_2 = 1$$

$$\boldsymbol{x}^T\boldsymbol{x}_i = 0; \quad i = 1, 2, \ldots, k-1$$

$$\boldsymbol{y}^T\boldsymbol{y}_i = 0; \quad i = 1, 2, \ldots, k-1$$

The principal angles have the property of $\theta_i \in [0, \pi/2]$ and can be computed through the SVD of $\boldsymbol{X}^T\boldsymbol{Y}$ (Edelman et al., 1999).

The prominent theme of analysing Grassmann manifolds is to embed them in higher dimensional Euclidean spaces. In this work, we are interested in embedding Grassmann manifolds in RKHS, which can be implicitly achieved through Grassmann kernels. A function $k : \mathcal{G}_{n,p} \times \mathcal{G}_{n,p} \to \mathbb{R}^+$ is a Grassmann kernel provided that it is positive definite and well defined for all $\boldsymbol{X} \in \mathcal{G}_{n,p}$. The well-defined criterion means that the kernel is invariant to various representations of the subspaces, i.e., $k(\boldsymbol{X}, \boldsymbol{Y}) = k(\boldsymbol{XQ}_1, \boldsymbol{YQ}_2)$, $\forall \boldsymbol{Q}_1, \boldsymbol{Q}_2 \in \mathbb{O}(p)$, where $\mathbb{O}(p)$ indicates orthonormal matrices of order p (Hamm and Lee, 2008). The repertoire of Grassmann kernels includes Binet–Cauchy (Wolf and Shashua, 2003) and projection kernels (Hamm and Lee, 2008). Furthermore, the first canonical correlation of two subspaces forms a pseudo kernel[5] on Grassmann manifolds (Harandi et al., 2011). The three kernels are shown below:

$$k_{\mathtt{BC}}(\mathtt{X}, \mathtt{Y}) = \det\left(\mathtt{X}^T\mathtt{Y}\mathtt{Y}^T\mathtt{X}\right) \tag{7}$$

$$k_{\mathtt{proj}}(\mathtt{X}, \mathtt{Y}) = \mathrm{tr}\left(\mathtt{X}^T\mathtt{Y}\mathtt{Y}^T\mathtt{X}\right) \tag{8}$$

$$k_{\mathtt{CC}}(\mathtt{X}, \mathtt{Y}) = \max_{\mathtt{x}\in\mathtt{X}, \mathtt{y}\in\mathtt{Y}} \mathtt{x}^T\mathtt{y} \tag{9}$$

4. Kernel analysis on Grassmann manifolds

Given a set of input/output data $\{(\mathtt{X}_1, l_1), (\mathtt{X}_2, l_2), \ldots, (\mathtt{X}_N, l_N)\}$, where $\mathtt{X}_i \in \mathcal{G}_{n,p}$ is a Grassmann point and l_i is the corresponding class label from $\mathcal{L} = \{1, 2, \ldots, C\}$, we are interested in optimisation problems in the form of Tikhonov regularisation (Tikhonov et al., 1977):

$$\max\{\mathbf{J}(\langle \mathbb{W}, \Phi(\mathtt{X}_1)\rangle, \ldots, \langle \mathbb{W}, \Phi(\mathtt{X}_N)\rangle, l_1, \ldots, l_N) + \lambda\Omega(\mathbb{W}) : \mathbb{W} \in \mathcal{H}\} \tag{10}$$

Here, $\mathcal{H}$ is a prescribed Hilbert space of dimension h (h could be infinity) equipped with an inner product $\langle\cdot,\cdot\rangle$, $\Omega : \mathcal{H} \to \mathbb{R}$ is a regulariser, and $\mathbf{J} : \left(\mathbb{R}^h\right)^N \times \mathcal{L}^N \to \mathbb{R}$ is a cost function. For certain choices of the regulariser, solving (10) reduces to identifying N parameters and not the dimension of $\mathcal{H}$. This is more formally explained by the representer theorem (Shawe-Taylor and Cristianini, 2004) which states that the solution $\widehat{\mathbb{W}}$ of (10) is a linear combination of the inputs when the regulariser is the square of the Hilbert space norm. For vector Hilbert spaces, this result is straightforward to prove and dates back to 1970s (Kimeldorf and Wahba, 1970). Argyriou et al. (2009) showed that the representer theorem holds for matrix Hilbert spaces as well.

In the following subsections, we first elucidate the details of the proposed Grassmann graph embedding DA (GGDA) algorithm, followed by how to use the mapping obtained by GGDA to accomplish classification. We then show that the conventional Grassmann DA (Hamm and Lee, 2008) is a special case of GGDA.

4.1. Grassmann graph embedding discriminant analysis (GGDA)

A graph $(\mathtt{V}, \mathtt{G})$ in our context refers to a collection of vertices or nodes, $\mathtt{V}$, and a collection of edges that connect pairs of vertices. We note that $\mathtt{G}$ is a symmetric matrix with elements describing the similarity between pairs of vertices. Moreover, the diagonal matrix $\mathtt{D}$ and the Laplacian matrix $\mathtt{L}$ of a graph are defined as $\mathtt{L} = \mathtt{D} - \mathtt{G}$, with the diagonal elements of $\mathtt{D}$ obtained as $\mathtt{D}(i,i) = \sum_j \mathtt{G}(i,j)$.

Given N labelled points $\mathbb{X} = \{(\mathtt{X}_i, l_i)\}_{i=1}^N$ from the underlying Grassmann manifold $\mathcal{G}n,p$, where $\mathtt{X}_i \in \mathbb{R}^{n\times p}$ and $l_i \in \{1, 2, \ldots, C\}$, with C denoting the number of classes, the local geometrical structure of $\mathcal{G}n,p$ can be modelled by building a within-class similarity graph $\mathtt{G}_w$ and a between-class similarity graph $\mathtt{G}_b$. The simplest forms of $\mathtt{G}_w$ and $\mathtt{G}_b$ are based on the nearest neighbour graphs defined below:

$$\mathtt{G}_w(i,j) = \begin{cases} 1, & \text{if } \mathtt{X}_i \in N_w(\mathtt{X}_j) \quad \text{or} \quad \mathtt{X}_j \in N_w(\mathtt{X}_i) \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

$$\mathtt{G}_b(i,j) = \begin{cases} 1, & \text{if } \mathtt{X}_i \in N_b(\mathtt{X}_j) \quad \text{or} \quad \mathtt{X}_j \in N_b(\mathtt{X}_i) \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

In (11), $N_w(\mathtt{X}_i)$ is the set of ν_w neighbours $\{\mathtt{X}_i^1, \mathtt{X}_i^2, \ldots, \mathtt{X}_i^{\nu}\}$, sharing the same label as l_i. Similarly in (12), $N_b(\mathtt{X}_i)$ contains ν_b

[3] Orthogonal group $\mathbb{O}(n)$ is the space of all $n \times n$ orthogonal matrices. It is not a vector space but a differentiable manifold with two connected components.

[4] A relation $\sim$ on manifold $\mathcal{M}$ is an equivalence relation iff it is reflexive ($\boldsymbol{X} \sim \boldsymbol{X}, \forall \boldsymbol{X} \in \mathcal{M}$), symmetric ($\boldsymbol{X} \sim \boldsymbol{Y}$, iff $\boldsymbol{Y} \sim \boldsymbol{X}, \forall \boldsymbol{X}, \boldsymbol{Y} \in \mathcal{M}$) and transitive (if $\boldsymbol{X} \sim \boldsymbol{Y}$ and $\boldsymbol{Y} \sim \boldsymbol{Z}$ then $\boldsymbol{X} \sim \boldsymbol{Z}, \forall \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z} \in \mathcal{M}$). The set of all elements that are equivalent to a point $\boldsymbol{X}$ is called the equivalence class of $\boldsymbol{X}$, i.e., $[\boldsymbol{X}] = \{[\boldsymbol{Y} \in \mathcal{M} : \boldsymbol{Y} \sim \boldsymbol{X}]$. The set $\Upsilon = \mathcal{M}/\sim$ of all equivalence classes of $\sim$ in $\mathcal{M}$ i.e., $\Upsilon = \{[\boldsymbol{X}] : \boldsymbol{X} \in \mathcal{M}\} = \{[\boldsymbol{Y} \in \mathcal{M} : \boldsymbol{Y} \sim \boldsymbol{X}] : \boldsymbol{X} \in \mathcal{M}\}$, is called the quotient of $\mathcal{M}$ by $\sim$.

[5] A pseudo kernel is a function where the positive definiteness is not guaranteed to be satisfied for whole range of the function's parameters. Nevertheless, it is possible to convert a pseudo kernel into a true kernel, as discussed for example in (Chen et al., 2009).

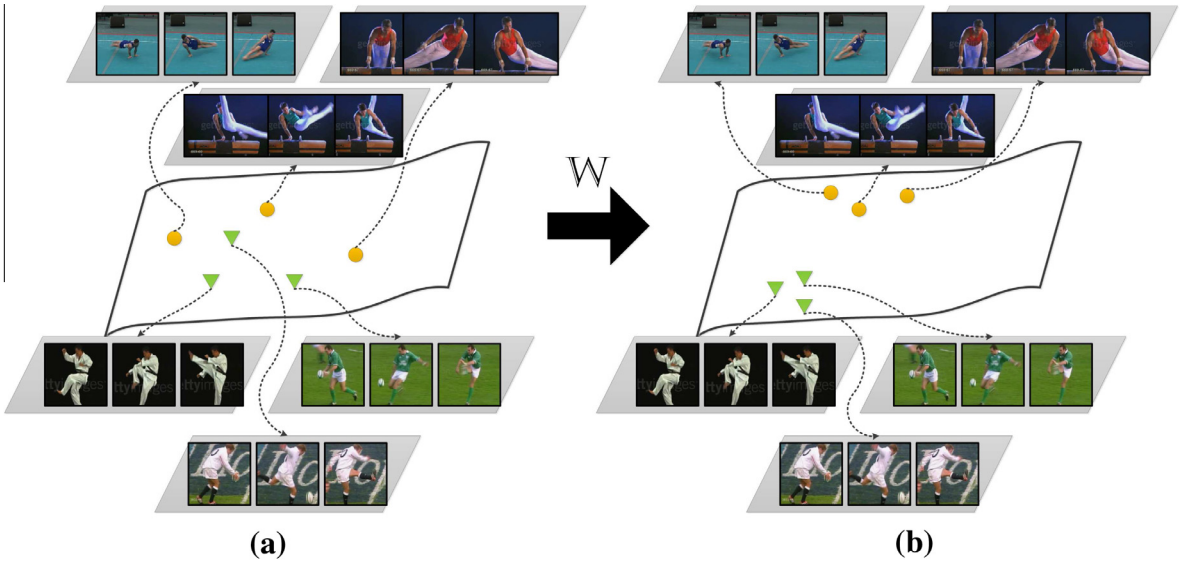

Fig. 2. A conceptual illustration of the proposed approach. (a) Actions can be modelled by linear subspaces, which in turn can be interpreted as points on a Grassmann manifold. In this figure, two types of actions (*kicking* and *swinging*) are shown. Having a proper geodesic distance between the points on the manifold, it is possible to convert the action recognition problem into a point to point classification problem. (b) Through the use of a Grassmann kernel, points on the Grassmann manifold can be mapped into an optimised RKHS, where not only certain local properties have been retained but also the discriminatory power between classes has been increased. Unlike conventional formalism of discriminant analysis, the proposed method preserves the geometrical structure and local information of a manifold by exploiting within-class and between-class similarity graphs.

neighbours having different labels. We note that more complex similarity graphs, like heat kernel graphs, can also be used to encode distances between points on Grassmann manifolds (Rosenberg, 1997).

Our aim is to simultaneously maximise a measure of discriminatory power and preserve the geometry of points (see Fig. 2 for a conceptual demonstration). This can be formalised by finding $\mathbb{W}:\Phi(\mathrm{X}_i)\rightarrow\mathrm{Y}_i$ such that the connected points of $\mathbb{G}_w$ are placed as close as possible, while the connected points of $\mathbb{G}_b$ are moved as far as possible. As such, a mapping must be sought by optimising the following two objective functions:

$$f_1=\min\frac{1}{2}\sum_{i,j}\|\mathrm{Y}_i-\mathrm{Y}_j\|^2G_w(i,j) \tag{13}$$

$$f_2=\max\frac{1}{2}\sum_{i,j}\|\mathrm{Y}_i-\mathrm{Y}_j\|^2G_b(i,j) \tag{14}$$

Eq. (13) punishes neighbours in the same class if they are mapped far away, while (14) punishes points of different classes if they are mapped close together. According to the representer theorem (Shawe-Taylor and Cristianini, 2004), the solution $\mathbb{W}=[\mathbf{\Gamma}_1|\mathbf{\Gamma}_2|\cdots|\mathbf{\Gamma}_r]$, can be expressed as a linear combination of data points, i.e., $\mathbf{\Gamma}_i=\sum_{j=1}^{N}w_{ij}\phi(\mathrm{X}_j)$. More specifically:

$$\mathrm{Y}_i=\left(\langle\mathbf{\Gamma}_1,\phi(\mathrm{X}_i)\rangle,\langle\mathbf{\Gamma}_2,\phi(\mathrm{X}_i)\rangle,\ldots,\langle\mathbf{\Gamma}_r,\phi(\mathrm{X}_i)\rangle\right)^T \tag{15}$$

We note that $\langle\mathbf{\Gamma}_l,\phi(\mathrm{X}_i)\rangle=\sum_{j=1}^{N}w_{l,j}\mathrm{tr}\left(\phi(\mathrm{X}_j)^T\phi(\mathrm{X}_i)\right)=\sum_{j=1}^{N}w_{l,j}k(\mathrm{X}_j,\mathrm{X}_i)$, $\mathrm{Y}_i=\mathbb{W}^T\mathrm{K}_i$, with $\mathrm{K}_i=\left(k(\mathrm{X}_i,\mathrm{X}_1),k(\mathrm{X}_i,\mathrm{X}_2),\ldots,k(\mathrm{X}_i,\mathrm{X}_N)\right)^T$ and

$$\mathbb{W}=\begin{pmatrix}w_{1,1}&w_{1,2}&\cdots&w_{1,r}\\w_{2,1}&w_{2,2}&\cdots&w_{2,r}\\\vdots&\vdots&\vdots&\vdots\\w_{N,1}&w_{N,2}&\cdots&w_{N,r}\end{pmatrix}$$

Plugging this into (13) results in:

$$\begin{aligned}\frac{1}{2}\sum_{i,j}\|\mathrm{Y}_i-\mathrm{Y}_j\|^2G_w(i,j)&=\frac{1}{2}\sum_{i,j}\|\mathbb{W}^T\mathrm{K}_i-\mathbb{W}^T\mathrm{K}_j\|^2G_w(i,j)\\&=\sum_i\mathrm{tr}\left(\mathbb{W}^T\mathrm{K}_i\mathrm{K}_i^T\mathbb{W}\right)G_w(i,i)\\&\quad-\sum_{i,j}\mathrm{tr}\left(\mathbb{W}^T\mathrm{K}_j\mathrm{K}_i^T\mathbb{W}\right)G_w(i,j)\\&=\mathrm{tr}\left(\mathbb{W}^T\mathbb{K}\mathbb{D}_w\mathbb{K}^T\mathbb{W}\right)-\mathrm{tr}\left(\mathbb{W}^T\mathbb{K}\mathbb{G}_w\mathbb{K}^T\mathbb{W}\right)\end{aligned} \tag{16}$$

where $\mathbb{K}=[\mathrm{K}_1|\mathrm{K}_2|\cdots|\mathrm{K}_N]$. Considering that $\mathbb{L}_b=\mathbb{D}_b-\mathbb{W}_b$, in a similar manner it can be shown that (14) can be simplified to:

$$\begin{aligned}\frac{1}{2}\sum_{i,j}\|\mathrm{Y}_i-\mathrm{Y}_j\|^2G_b(i,j)&=\mathrm{tr}\left(\mathbb{W}^T\mathbb{K}\mathbb{D}_b\mathbb{K}^T\mathbb{W}\right)-\mathrm{tr}\left(\mathbb{W}^T\mathbb{K}\mathbb{G}_b\mathbb{K}^T\mathbb{W}\right)\\&=\mathrm{tr}\left(\mathbb{W}^T\mathbb{K}\mathbb{L}_b\mathbb{K}^T\mathbb{W}\right)\end{aligned} \tag{17}$$

To solve (13) and (14) simultaneously, we need to add the following normalising constraint to the problem:

$$\mathrm{tr}\left(\mathbb{W}^T\mathbb{K}\mathbb{D}_w\mathbb{K}^T\mathbb{W}\right)=1 \tag{18}$$

This constraint enables us to convert the minimisation problem (13) into a maximisation one. Consequently, both equations can be combined into one maximisation problem. Moreover, as shown later, the imposed constraint acts as a norm regulariser in the original Tikhonov problem (10), thus satisfying the necessary condition of the representer theorem. Plugging (18) into (13) results in:

$$\begin{aligned}&\min\left\{\mathrm{tr}\left(\mathbb{W}^T\mathbb{K}\mathbb{D}_w\mathbb{K}^T\mathbb{W}\right)-\mathrm{tr}\left(\mathbb{W}^T\mathbb{K}\mathbb{G}_w\mathbb{K}^T\mathbb{W}\right)\right\}\\&\quad=\min\left\{1-\mathrm{tr}\left(\mathbb{W}^T\mathbb{K}\mathbb{G}_w\mathbb{K}^T\mathbb{W}\right)\right\}=\max\left\{\mathrm{tr}\left(\mathbb{W}^T\mathbb{K}\mathbb{G}_w\mathbb{K}^T\mathbb{W}\right)\right\}\end{aligned} \tag{19}$$

subject to the constraint shown in (18). As a result, the max versions of (13) and (14) can be merged by the Lagrangian method (C.M. Bishop, 2006) as follows:

$$\begin{aligned}&\max\left\{\mathrm{tr}\left(\mathbb{W}^T\mathbb{K}(\mathbb{L}_b+\beta\mathbb{G}_w)\mathbb{K}^T\mathbb{W}\right)\right\}\\&\text{subject to }\mathrm{tr}\left(\mathbb{W}^T\mathbb{K}\mathbb{D}_w\mathbb{K}^T\mathbb{W}\right)=1\end{aligned} \tag{20}$$

where β is a Lagrangian multiplier. The solution to the optimisation in (20) can be sought as the r largest eigenvectors of the following generalised eigenvalue problem:

$$\mathbb{K}\{\mathbb{L}_b + \beta\mathbb{G}_w\}\mathbb{K}^T\mathbb{W} = \lambda\mathbb{K}\mathbb{D}_w\mathbb{K}^T\mathbb{W} \tag{21}$$

We note that in (21), the imposed constraint (18) serves as a norm regulariser and satisfies the representer theorem condition. Algorithm 1 assembles all the above details into pseudo-code for training the Grassmann graph embedding discriminant analysis (GGDA).

Algorithm 1. Pseudocode for training Grassmann graph-embedding discriminant analysis (GGDA)

Input:
- A training set $\mathbb{X} = \{(\mathbb{X}_i, l_i)\}_{i=1}^N$ from the underlying Grassmann manifold $\mathcal{G}n,p$, where $\mathbb{X}_i \in \mathbb{R}^{n\times p}$ is a subspace and $l_i \in \{1,2,\ldots,C\}$, with C denoting the number of classes
- A kernel function k_{ij}, for measuring the similarity between two points on a Grassmann manifold

Output: The projection matrix $\mathbb{W} = [\Upsilon_1|\Upsilon_2|\cdots|\Upsilon_r]$,

Compute the Gram matrix $[\mathbb{K}]_{ij}$ for all $\mathbb{X}_i$, $\mathbb{X}_j$
for $i = 1 \rightarrow N-1$ **do**
  for $j = i+1 \rightarrow N$ **do**
    Compute the geodesic distances $d_g(i,j)$ between $\mathbb{X}_i$ and $\mathbb{X}_j$.
    $d_g(j,i) = d_g(i,j)$
  end for
end for
$\mathbb{G}_w \leftarrow 0_{N\times N}$
$\mathbb{G}_b \leftarrow 0_{N\times N}$
% Use the obtained $d_g(i,j)$ to determine neighbourhoods in the following loop.
for $i = 1 \rightarrow N$ **do**
  if ($\mathbb{X}_j$ is in the first k_w nearest neighbours of $\mathbb{X}_i$) **and** ($l_j == l_i$) **then**
    $G_w(i,j) \leftarrow 1$
    $G_w(j,i) \leftarrow 1$
  end if
  if ($\mathbb{X}_j$ is in the first k_b nearest neighbours of $\mathbb{X}_i$) **and** ($l_j \neq l_i$) **then**
    $G_b(i,j) \leftarrow 1$
    $G_b(j,i) \leftarrow 1$
  end if
end for
$\mathbb{D}_w \leftarrow 0_{N\times N}$
$\mathbb{D}_b \leftarrow 0_{N\times N}$
$D_w(i,i) \leftarrow \sum_j G_w(i,j)$
$D_b(i,i) \leftarrow \sum_j G_b(i,j)$
$\mathbb{L}_b \leftarrow \mathbb{D}_b - \mathbb{G}_b$
$\{\Upsilon_i, \tilde{\lambda}_i\}_{i=1}^r \leftarrow$ generalised eigenvectors and eigenvalues of $\mathbb{K}\{\mathbb{L}_b + \beta\mathbb{G}_w\}\mathbb{K}^T\mathbb{W} = \lambda\mathbb{K}\mathbb{D}_w\mathbb{K}^T\mathbb{W}$, with the eigenvectors ordered according to descending eigenvalues.

4.2. Classification

Upon acquiring the mapping $\mathbb{W}$, classification tasks on Grassmann manifolds are reduced to classification tasks in vector spaces. More precisely, for any query image set $\mathbb{X}_q$, a vector representation using the kernel function and the mapping $\mathbb{W}$ is acquired, i.e., $\mathbb{V}_q = \mathbb{W}^T\mathbb{K}_q$, where $\mathbb{K}_q = \big(\langle\phi(\mathbb{X}_1),\phi(\mathbb{X}_q)\rangle, \langle\phi(\mathbb{X}_2),\phi(\mathbb{X}_q)\rangle, \ldots, \langle\phi(\mathbb{X}_N),\phi(\mathbb{X}_q)\rangle\big)^T$. Similarly, gallery points $\mathbb{X}_i$ are represented by r dimensional vectors $\mathbb{V}_i = \mathbb{W}^T\mathbb{K}_i$. Classification methods such as nearest-neighbours or support vector machines (C.M. Bishop, 2006) can be employed to label $\mathbb{V}_q$.

4.3. Relation to Grassmann discriminant analysis

Here we address the relation between GGDA and Grassmann discriminant analysis (GDA) (Hamm and Lee, 2008). GDA is a learning framework on Grassmann manifolds that utilises kernel analysis to project Grassmann points into a higher discriminative Hilbert space. More specifically, let $\mathbb{X} = \{(\mathbb{X}_i, l_i)\}_{i=1}^N$ be a set of N labelled points on Grassmann manifold $\mathcal{G}n,p$. GDA seeks a transformation $\mathbb{W} : \Phi(\mathbb{X}_i) \rightarrow \mathbb{Y}_i$ such that the ratio of between-class to within-class scatters is maximised. The within-class and between-class scatters are defined as:

$$S_W = \sum_{j=1}^{C}\sum_{i:l_i=j}\|\mathbb{Y}_i - \boldsymbol{\mu}_j\|^2 \tag{22}$$

$$S_B = \sum_{j=1}^{C} n_j\|\bar{\boldsymbol{\mu}} - \boldsymbol{\mu}_j\|^2 \tag{23}$$

where $\sum_{i:l_i=j}$ denotes the summation over i such that $l_i = j$, while $\boldsymbol{\mu}_j = \frac{1}{n_j}\sum_{i:l_i=j}\mathbb{Y}_i$ is the mean of the samples in class j, and $\bar{\boldsymbol{\mu}} = \frac{1}{N}\sum_{j=1}^{C} n_j\boldsymbol{\mu}_j$ is the mean of all samples.

We note that if the data pairs in the same class are moved closer, the within-class scatter S_W gets smaller. On a similar note, if the data pairs in different classes are more separated from each other, the between-class scatter S_B gets larger. GDA can be seen as a special case of GGDA when particular within-class and between-class similarity graphs are used. The following lemma formalises the relation between GGDA and GDA.

Lemma 1. *GDA is a special case of GGDA if*

$$\mathbb{G}_w(i,j) = \begin{cases} \frac{1}{n_k}, & \text{if } l_i = l_j = k \\ 0, & \text{if } l_i \neq l_j \end{cases} \tag{24}$$

$$\mathbb{G}_b(i,j) = \begin{cases} \frac{1}{N} - \frac{1}{n_k}, & \text{if } l_i = l_j = k \\ \frac{1}{N}, & \text{if } l_i \neq l_j \end{cases} \tag{25}$$

A proof of this lemma is given in A.

5. Experiments

In this section we first compare and contrast the performance of the proposed GGDA method against several state-of-the-art approaches on the UCF sport action dataset (Rodriguez et al., 2008), the KTH human motion dataset (Schuldt et al., 2004) and the Ballet dataset (Wang and Mori, 2009). We then conclude the section by assessing the sensitivity of the proposed method against occlusion and misalignment.

5.1. Empirical evaluations

Here, we appraise the performance of the proposed GGDA method[6] against the state-of-the-art Grassmann discriminant analysis (Hamm and Lee, 2008), kernel version of affine hull image-set distance (Cevikalp and Triggs, 2010), tensor canonical correlation analysis (Kim and Cipolla, 2009), spatial-temporal words (Niebles et al., 2008) and hierarchy of discriminative space-time neighbourhood features (Kovashka and Grauman, 2010) on the UCF sport action dataset (Rodriguez et al., 2008), the KTH human motion dataset (Schuldt et al., 2004) and the Ballet dataset (Wang and Mori, 2009). In all experiments, the projection kernel has been used in GGDA.

5.1.1. UCF SPORT dataset

The UCF sport action dataset (Rodriguez et al., 2008) consists of ten categories of human actions including swinging on the pommel

[6] Source code for the proposed method is available at <http://itee.uq.edu.au/uqmhara1>.

horse, driving, kicking, lifting weights, running, skateboarding, swinging at the high bar, swinging golf clubs, and walking. The number of videos for each action varies from 6 to 22 and there are 150 video sequences in total. Furthermore, the videos presented in this dataset have nonuniform backgrounds and both the camera and the subject are moving in some actions. Frames in all video sequences are resized to 32×32. We use the region of interest provided with the dataset. The standard protocol in this dataset is the leave-one-out (LOO) cross validation (Rodriguez et al., 2008; Kovashka and Grauman, 2010; Wu et al., 2011).

Two sets of experiments are considered here. Firstly, we contrasted ARMA modelling against image-set modelling for creating Grassmann manifolds. For the ARMA process, state-space dimension $p = 20$ and the observability matrix was truncated at 5. For the image-set modelling, a subspace of order 20 (similar to ARMA modelling) was extracted from each video. The confusion matrices for GGDA using image-set and ARMA modelling are shown in Tables 1 and 2, respectively. We note that the overall classification accuracy is much better when actions are modelled by ARMA. Moreover, we can see from the confusion matrix in Table 2 that the proposed GGDA method perfectly classifies the actions of diving, kicking, pommel horsing, high bar swinging and lifting.

Secondly, we compared GGDA against three state-of-the-art Euclidean approaches, HOG3D (Wang et al., 2009), hierarchy of discriminative space-time neighbourhood features (HDN) (Kovashka and Grauman, 2010) and augmented features (Wu et al., 2011) in conjunction with multiple kernel learning (Bach et al., 2004) (AFMKL). HOG3D is an extension of histogram of oriented gradient descriptor (Dalal and Triggs, 2005) to spatio-temporal spaces. HDN learns shapes of space-time feature neighbourhoods that are most discriminative for a given action category. The idea is to form new features composed of the neighbourhoods around the interest points in a video. AFMKL exploits appearance distribution features and spatio-temporal context features in a learning scheme for action recognition. We also compared GGDA against conventional discriminant analysis on Grassmann manifolds (GDA) (Hamm and Lee, 2008). As shown in Table 3, GGDA with ARMA modelling achieves the highest accuracy. Modelling actions by image-sets (instead of ARMA) results in roughly 8 percentage points drop in the classification accuracy. Nevertheless, GGDA with only image-set modelling still outperforms Euclidean-based approaches such as HOG3D (Wang et al., 2009) and HDN (Kovashka and Grauman, 2010).

Table 3
Recognition accuracy (in %) for the UCF action recognition dataset using HOG3D, HDN (Kovashka and Grauman, 2010), AFMKL (Wu et al., 2011) and the GGDA approach.

Method	Recognition accuracy
HOG3D (Wang et al., 2009)	85.6
HDN (Kovashka and Grauman, 2010)	87.3
AFMKL (Wu et al., 2011)	91.3
GDA (Hamm and Lee, 2008) with image-set modelling	80.7
GDA (Hamm and Lee, 2008) with ARMA modelling	85.3
GGDA with image-set modelling	88.4
GGDA with ARMA modelling	96.6

5.1.2. KTH dataset

The KTH human motion data set (Schuldt et al., 2004) contains six types of human actions '*walking*', '*jogging*', '*running*', '*boxing*', '*hand waving*', and '*hand clapping*' performed several times by 25 subjects in four scenarios: '*outdoors*', '*outdoors with scale variation*', '*outdoors with different clothes*', and '*indoors*'. See Fig. 3(b) for sample frames.

We first run an automatic pre-processing step to track and stabilise the video sequences so that all of the figures appear in the center of the field of view. All videos were resized to $20 \times 20 \times 32$. In order to have a standard length of 32 frames per video, the middle 32 frames were used. Actions were modelled by an ARMA process, with state-space dimension $p = 20$ and the observability matrix was truncated at 5. For the sake of comparison, we also modelled actions by image-sets of order 20. To facilitate comparison with prior work, we followed the leave-one-out

Table 1
Confusion matrix (in %) for the GGDA method on the UCF sport action dataset using LOO protocol. Actions are modelled by image-sets.

	Diving	Golf Swing	Kicking	Riding Horse	Running	Skate boarding	Pommel Horse	High-bar Swinging	Walking	Lifting
Diving	100	0	0	0	0	0	0	0	0	0
Golf-Swing	0	83.33	5.56	0	5.56	0	0	0	5.56	0
Kicking	0	0	80	0	15	5	0	0	0	0
Riding Horse	0	0	0	58.33	16.67	0	8.33	8.33	8.33	0
Running	7.69	0	0	0	84.62	0	0	7.69	0	0
Skateboarding	0	0	0	0	0	91.67	0	0	8.33	0
Pommel-Horse	5	0	0	0	0	0	95	0	0	0
High-bar Swinging	0	0	0	0	0	0	0	100	0	0
Walking	4.55	0	0	0	0	0	0	4.55	90.91	0
Lifting	0	0	0	0	0	0	0	0	0	100

Table 2
Confusion matrix (in %) for the GGDA method on the UCF sport action dataset using LOO protocol. Actions are modelled by ARMA process.

	Diving	Golf Swing	Kicking	Riding Horse	Running	Skate Boarding	Pommel Horse	High-bar Swinging	Walking	Lifting
Diving	100	0	0	0	0	0	0	0	0	0
Golf-Swing	0	94.4	5.56	0	0	0	0	0	0	0
Kicking	0	0	100	0	0	0	0	0	0	0
Riding Horse	0	0	0	91.67	8.33	0	0	0	0	0
Running	0	0	0	0	92.31	7.69	0	0	0	0
Skateboarding	0	0	0	0	8.33	91.67	0	0	0	0
Pommel-Horse	0	0	0	0	0	0	100	0	0	0
High-bar Swinging	0	0	0	0	0	0	0	100	0	0
Walking	0	0	0	0	4.55	0	0	0	95.45	0
Lifting	0	0	0	0	0	0	0	0	0	100

Fig. 3. Examples from: (a) UCF sport action dataset (Rodriguez et al., 2008), (b) KTH dataset (Schuldt et al., 2004), and (c) Ballet dataset (Wang and Mori, 2009).

Table 4
Recognition accuracy (in %) for the KTH action recognition dataset using spatio temporal words (STW) (Niebles et al., 2008), fusion of appearance and distribution method (BoW with MKL) (Bregonzio et al., 2012), tensor canonical correlation analysis (TCCA) (Kim and Cipolla, 2009), Grassmann discriminant analysis (GDA) (Hamm and Lee, 2008) and the proposed GGDA approach.

Method	Recognition accuracy (%)
STW (Niebles et al., 2008)	83
BoW-MKL (Bregonzio et al., 2012)	94
TCCA (Kim and Cipolla, 2009)	95
GDA (Hamm and Lee, 2008) with image-set modelling	83
GDA (Hamm and Lee, 2008) with ARMA modelling	86
GGDA with image-set modelling	97
GGDA with ARMA modelling	99

Table 6
Recognition accuracy (in %) along its standard deviation for the Ballet dataset using Grassmann geodesic distance (Turaga et al., 2011), Kernel Affine Hull method (KAHM) (Cevikalp and Triggs, 2010), Grassmann discriminant analysis (GDA) (Hamm and Lee, 2008) and the proposed GGDA approach.

Method	Recognition accuracy
Geodesic distance	$77.34\% \pm 1.8$
KAHM (Cevikalp and Triggs, 2010)	$79.71\% \pm 2.3$
GDA (Hamm and Lee, 2008) with image-set modelling	$78.05\% \pm 2.9$
GDA (Hamm and Lee, 2008) with ARMA modelling	$73.70\% \pm 1.8$
GGDA with image-set modelling	$83.08\% \pm 1.8$
GGDA with ARMA modelling	$77.63\% \pm 2.9$

(LOO) cross validation protocol used in (Niebles et al., 2008; Kim and Cipolla, 2009). The classification results are reported in Table 4.

We compared GGDA against two state-of-the-art Euclidean approaches: spatial-temporal words (STW) (Niebles et al., 2008) and bag of words model in conjunction with multiple kernel learning (Bach et al., 2004) (BoW-MKL) (Bregonzio et al., 2012). In STW, a video sequence is represented by a set of spatial-temporal words, extracted from space-time interest points. The algorithm then utilises latent topic models such as the probabilistic latent semantic analysis (Hofmann, 1999) to learn the probability distributions of the spatial-temporal words. BoW-MKL exploits global spatio-temporal distribution of interest points by extracting holistic features from clouds of interest points accumulated over multiple temporal scales. Then extracted features are fused using MKL. We also compared GGDA against Tensor Canonical Correlation Analysis (TCCA) (Kim and Cipolla, 2009) and conventional discriminant analysis on Grassmann manifolds (GDA) (Hamm and Lee, 2008). TCCA is an extension of canonical correlation analysis (a principled tool to inspect linear relations between two sets of vectors) to tensor spaces and measures video-to-video volume similarity.

Table 5
Confusion matrix (in %) for the GGDA method on the KTH action recognition dataset using LOO protocol.

	Boxing	Hand clapping	Hand waving	Jogging	Running	Walking
Boxing	100	0	0	0	0	0
Hand clapping	0	100	0	0	0	0
Hand waving	0	0	100	0	0	0
Jogging	0	0	0	99	0	1
Running	0	0	0	1	97	2
Walking	0	0	0	0	2	98

Table 7
Confusion matrix (in %) for the GGDA method on the Ballet dataset. Actions are modelled by image-sets.

	LR hand opening	RL hand opening	Standing hand opening	Leg swinging	Jumping	Turning	Hopping	Standing still
LR hand opening	81.50	2.50	2.50	0.00	3.00	5.00	0.00	5.50
RL hand opening	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00
Standing hand opening	0.00	0.00	87.88	4.55	0.00	6.06	0.00	1.52
Leg swinging	3.57	0.00	0.00	85.71	3.57	3.57	0.00	3.57
Jumping	4.55	4.55	9.09	13.14	50.54	0.00	13.60	4.55
Turning	0.00	0.00	5.82	0.00	0.00	94.18	0.00	0.00
Hopping	0.00	2.33	17.23	2.33	2.33	0.00	71.14	4.65
Standing still	6.00	0.00	27.50	11.00	0.00	10.00	0.00	45.50

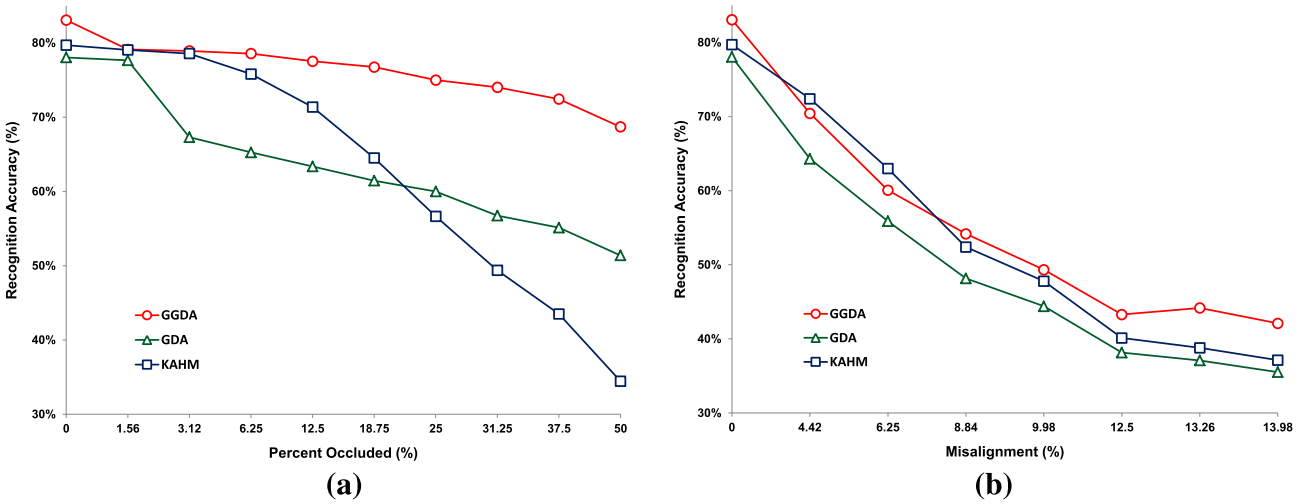

Fig. 4. Recognition rate on the Ballet dataset (Wang and Mori, 2009) for increasing amount of (a) occlusion and (b) misalignment, for KAHM (Cevikalp and Triggs, 2010), GDA (Hamm and Lee, 2008) and the proposed GGDA algorithm.

Looking at the results in Table 4, the proposed GGDA method achieves the highest classification accuracy. We note that though walking, jogging and running are sometimes confused by GGDA (see the confusion matrix in Table 5). However, in comparison to the other methods GGDA is able to reduce this ambiguity greatly.

5.1.3. Ballet dataset

The previous experiments may imply that actions should be exclusively modelled by an ARMA process. As we show in this experiment, this is not always the case, indicating that image-sets are also useful for modelling of actions. The Ballet dataset contains 44 real video sequences of 8 actions collected from an instructional ballet DVD (Wang and Mori, 2009).[7] The dataset consists of 8 complex motion patterns performed by three subjects. The actions include: '*left-to-right hand opening*', '*right-to-left hand opening*', '*standing hand opening*', '*leg swinging*', '*jumping*', '*turning*', '*hopping*' and '*standing still*'. Fig. 3(c) shows samples. This dataset is very challenging due to the significant intra-class variations in terms of speed, spatial and temporal scale, clothing and movement.

Available samples of each action were randomly split into training and testing set (the number of actions in both training and testing sets were fairly even). The process of random splitting was repeated ten times and the average classification accuracy is reported in Table 6. For comparison, the GGDA algorithm is contrasted with geodesic distance on Grassmann manifolds (Eq. (5)), and the state-of-art kernel version of affine hull set matching (KAHM) (Cevikalp and Triggs, 2010) and Grassmann discriminant analysis (GDA) (Hamm and Lee, 2008). Cevikalp and Triggs (2010) proposed to measure the similarity between image-sets using geometric distances between their convex models. In this experiment, for Grassmann-based analysis, actions were modelled by image-sets of order 10. For the sake of comparison, we also modelled actions by ARMA process with state-space dimension $p = 20$ (the observability matrix was truncated at 5).

Table 6 shows that the GGDA algorithm with image-set modelling obtains the highest accuracy and outperforms state-of-the-art methods KAHM and GDA significantly. The confusion matrix of the proposed GGDA method is shown in Table 7. We note that the lowest recognition accuracy belongs to the "standing still" action which is mainly confused with "standing hand opening". We conjecture that the low performance of GGDA method with ARMA modelling is due to lack of temporal information for some action, such as the "standing still" action.

5.2. Sensitivity analysis

The performance of a visual recognition system can be negatively affected by variations in the environment (e.g., illumination), capturing device (e.g., pose variation, occlusion) as well as preprocessing steps that prepare data for the system (e.g., registration). A detailed sensitivity analysis for the aforementioned factors is beyond the scope of this paper. However, in this paper we analyse the sensitivity of GGDA algorithm against occlusion and misalignment. To this end, we elect the Ballet dataset (Wang and Mori, 2009) for our analysis since this dataset has a uniform background and fair illumination (and therefore minimises the effect of variations in illumination and background in the analysis). Actions are modelled by image-sets and the test setup employed in Section 5.1.3 is utilised again.

[7] The study in (Wang and Mori, 2009) addresses the problem of recognising actions in still images, which is different from the work presented here.

5.2.1. Sensitivity to occlusion

In this experiment, we assess the performance at various levels of occlusion, from 1.56% up to 50%, by replacing a set of randomly located square blocks of size 4×4 in a test image with a blank block. The location of occlusion is randomly chosen for each test image and is unknown to the system. The training images do not contain occlusions. Methods that select fixed features or blocks of the image are unlikely to succeed here due to the unpredictable location of the occlusion.

Fig. 4(a) shows the recognition rates of KAHM, GDA and GGDA. The proposed GGDA method significantly outperforms the other two methods for almost all levels of occlusion. Up to 40 percent occlusion, the performance of GGDA has dropped roughly by 10 percentage points. While robustness to occlusion can be partially attributed to subspace modelling (as can also be seen for GDA), we conjecture that the proposed GGDA method has better captured the true Grassmannian geometry (through within and between graphs as the training images do not contain occlusions) and is hence more robust to the missing parts.

5.2.2. Sensitivity to misalignment

The temporal and spatial misalignment could deteriorate the performance of an action recogniser drastically (Shariat and Pavlovic, 2011). In this work we only consider spatial misalignment and assess and contrast the sensitivity of GGDA algorithm as compared to KAHM (Cevikalp and Triggs, 2010) and GDA (Hamm and Lee, 2008) methods. To this end, we have introduced random displacements to the frames of query videos and measured the accuracy for various amounts of displacements. Fig. 4(b) shows the results of misalignment analysis for KAHM (Cevikalp and Triggs, 2010), GDA (Hamm and Lee, 2008) and the proposed GGDA methods. The horizontal axis here demonstrates the degree of misalignment, i.e., the length of random displacement vector divided by the maximum possible misalignment (for frame of size $S_x \times S_y$, the maximum possible misalignment is $\frac{1}{2}\sqrt{S_x^2 + S_y^2}$). Fig. 4(b) reveals that all studied algorithms are sensitive to misalignment. The larger the displacement, the lower would be the recognition accuracy. This is mainly due to the holistic representation of images which has been shown to be highly fragile to misalignment (Wong et al., 2012).

6. Main findings and future directions

In this paper, we first demonstrated how actions can be modelled by linear subspaces. Subspaces are able to accommodate the effects of various image variations and can capture the dynamic properties of videos (Turaga et al., 2011). Since subspaces form non-Euclidean and curved Riemannian manifolds known as Grassmann manifolds, we exploited the geometry of space to design an action recogniser. As such, we proposed graph-embedding discriminant analysis on Grassmann manifolds by embedding manifolds into RKHS. The proposed method utilises within-class and between-class similarity graphs to characterise intra-class compactness and inter-class separability, respectively.

Thorough experiments on the KTH (Schuldt et al., 2004), UCF Sports (Rodriguez et al., 2008) and Ballet (Wang and Mori, 2009) datasets, which include various realistic challenges such as background clutter, partial occlusion, changes in viewpoint, scale and illumination, and complexity of motion showed that the proposed approach obtains notable improvements in discrimination accuracy in comparison to several state-of-the-art methods. This included Grassmann discriminant analysis (Hamm and Lee, 2008), kernel version of affine hull image-set distance (Cevikalp and Triggs, 2010), tensor canonical correlation analysis (Kim and Cipolla, 2009), spatial-temporal words (Niebles et al., 2008) and hierarchy of discriminative space-time neighbourhood features (Kovashka and Grauman, 2010).

The proposed GGDA algorithm (like other discriminant analysis techniques) requires several samples of each class to determine the optimum mapping. As such, GGDA cannot be applied to action recognition problems where just one sample video per class is available for training. Moreover, since the structure of manifold is encoded via between and within similarity graphs, inappropriate parameter selection (for example the size of neighbourhood N_w in Eqn. 11) might result in poor performance.

In this paper, each action is modelled by just one subspace. When the length of action video is small, this straightforward treatment appears to be appropriate. However, for very complex motions or when an action is described by a lengthy video, straightforward subspace modelling might not be adequate enough. As such, future avenues of research include exploring how several subspaces can be generated from complex motion videos. We are also keen to explore how discriminatory subspaces can be generated from interest-points (e.g., local spatio-temporal features) in a video.

Acknowledgements

NICTA is funded by the Australian Government as represented by the *Department of Broadband, Communications and the Digital Economy*, as well as the Australian Research Council through the *ICT Centre of Excellence* program.

Appendix A. Proof of Lemma 1

Plugging (24) into (13), we get:

$$\begin{aligned}
\frac{1}{2}\sum_{i,j}\|\mathrm{Y}_i - \mathrm{Y}_j\|^2 G_w(i,j) &= \sum_{i,j}\mathrm{Y}_i^T\mathrm{Y}_i G_w(i,j) - \sum_{i,j}\mathrm{Y}_i^T\mathrm{Y}_j G_w(i,j) \\
&= \sum_i \mathrm{Y}_i^T\mathrm{Y}_i \sum_j G_w(i,j) - \sum_{k=1}^{C}\frac{1}{n_k}\sum_{i,j:l_i=l_j=k}\mathrm{Y}_i^T\mathrm{Y}_j \\
&= \sum_i \mathrm{Y}_i^T\mathrm{Y}_i - \sum_{k=1}^{C}\frac{1}{n_k}\sum_{i,j:l_i=l_j=k}\mathrm{Y}_i^T\mathrm{Y}_j \\
&= \sum_{k=1}^{C}\sum_{i:l_i=k}\left(\mathrm{Y}_i - \frac{1}{n_k}\sum_{j:l_j=k}\mathrm{Y}_j\right)^T\left(\mathrm{Y}_i - \frac{1}{n_k}\sum_{j:l_j=k}\mathrm{Y}_j\right) \\
&= \sum_{k=1}^{C}\sum_{i:l_i=k}\left(\mathrm{Y}_i - \boldsymbol{\mu}_k\right)^2 = S_W
\end{aligned} \tag{A.1}$$

To show the equivalency for between-class scatters, we note that:

$$S_M = S_B + S_W = \sum_{i=1}^{N}\|\mathrm{Y}_i - \boldsymbol{\mu}\|^2 = \sum_{i=1}^{N}\mathrm{Y}_i^T\mathrm{Y}_i - \frac{1}{N}\sum_{i,j=1}^{N}\mathrm{Y}_i^T\mathrm{Y}_j \tag{A.2}$$

By plugging (25) into (14) and using $G_b(i,j) = \frac{1}{N} - G_w(i,j)$, we get:

$$\begin{aligned}
\frac{1}{2}\sum_{i,j}\|\mathrm{Y}_i - \mathrm{Y}_j\|^2 G_b(i,j) &= \frac{1}{2N}\sum_{i,j}\|\mathrm{Y}_i - \mathrm{Y}_j\|^2 - \frac{1}{2}\sum_{i,j}\|\mathrm{Y}_i \\
&\quad - \mathrm{Y}_j\|^2 G_w(i,j) \\
&= \frac{1}{N}\sum_{i,j}\mathrm{Y}_i^T\mathrm{Y}_i - \frac{1}{N}\sum_{i,j}\mathrm{Y}_i^T\mathrm{Y}_j - S_W \\
&= \sum_i \mathrm{Y}_i^T\mathrm{Y}_i - \frac{1}{N}\sum_{i,j}\mathrm{Y}_i^T\mathrm{Y}_j - S_W \\
&= S_M - S_W = S_B
\end{aligned} \tag{A.3}$$

References

Argyriou, A., Micchelli, C.A., Pontil, M., 2009. When is there a representer theorem? Vector versus matrix regularizers. J. Machine Learn. Res. 10, 2507–2529.

Bach, F.R., Lanckriet, G.R.G., Jordan, M.I., 2004. Multiple kernel learning, conic duality, and the SMO algorithm. In: Internat. Conf. on Machine Learning. ACM, New York, NY, USA.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer.

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

Cevikalp, H., Triggs, B., 2010. Face recognition based on image sets. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 2567–2573.

Chen, J., Ye, J., Li, Q., 2007. Integrating global and local structures: A least squares framework for dimensionality reduction. In: Proc. Computer Vision and Pattern Recognition (CVPR), pp. 1–8.

Chen, Y., Garcia, E.K., Gupta, M.R., Rahimi, A., Cazzanti, L., 2009. Similarity-based classification: Concepts and algorithms. J. Machine Learn. Res. 10, 747–776.

Cock, K.D., Moor, B.D., 2002. Subspace angles between ARMA models. Syst. Control Lett. 46, 265–270.

Comaniciu, D., Meer, P., 2002. Mean shift: A robust approach toward feature space analysis. IEEE Trans. Pattern Anal. Machine Intell. 24 (5), 603–619.

Dalal, N., Triggs, W., 2005. Histograms of oriented gradients for human detection. In: IEEE Conf. on Computer Vision and Pattern Recognition, pp. 886–893.

Edelman, A., Arias, T.A., Smith, S.T., 1999. The geometry of algorithms with orthogonality constraints. SIAM J. Matrix Anal. Appl. 20 (2), 303–353.

Hamm, J., Lee, D.D., 2008. Grassmann discriminant analysis: A unifying view on subspace-based learning. In: Internat. Conf. on Machine Learning (ICML), pp. 376–383.

Harandi, M.T., Sanderson, C., Shirazi, S., Lovell, B.C., 2011. Graph embedding discriminant analysis on Grassmannian manifolds for improved image set matching. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 2705–2712.

Harandi, M.T., Sanderson, C., Wiliem, A., Lovell, B.C., 2012. Kernel analysis over Riemannian manifolds for visual recognition of actions, pedestrians and textures. In: IEEE Workshop on Applications of Computer Vision (WACV), pp. 433–439.

Hofmann, T., 1999. Probabilistic latent semantic indexing. In: Internat. ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 50–57.

Kim, T.-K., Cipolla, R., 2009. Canonical correlation analysis of video volume tensors for action categorization and detection. IEEE Trans. Pattern Anal. Machine Intell. (PAMI) 31, 1415–1428.

Kimeldorf, G.S., Wahba, G., 1970. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. Ann. Math. Stat. 41, 495–502.

Kovashka, A., Grauman, K., 2010. Learning a hierarchy of discriminative space–time neighborhood features for human action recognition. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 2046–2053.

Li, B., Ayazoglu, M., Mao, T., Camps, O., Sznaier, M., 2011. Activity recognition using dynamic subspace angles. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 3193–3200.

Lui, Y.M., 2012. Advances in matrix manifolds for computer vision. Image Vision Comput. 30 (6–7), 380–388.

Niebles, J., Wang, H., Fei-Fei, L., 2008. Unsupervised learning of human action categories using spatial–temporal words. Internat. J. Comput. Vision 79, 299–318.

O'Hara, S., Lui, Y.M., Draper, B.A., 2012. Using a product manifold distance for unsupervised action recognition. Image Vision Comput. 30 (3), 206–216.

Rodriguez, M., Ahmed, J., Shah, M., 2008. Action MACH a spatio-temporal maximum average correlation height filter for action recognition. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1–8.

Rosenberg, S., 1997. The Laplacian on a Riemannian Manifold: An Introduction to Analysis on Manifolds. Cambridge University Press.

Schuldt, C., Laptev, I., Caputo, B., 2004. Recognizing human actions: A local SVM approach. In: Internat. Conf. on Pattern Recognition (ICPR), pp. 32–36.

Shariat, S., Pavlovic, V., 2011. Isotonic CCA for sequence alignment and activity recognition. In: IEEE Internat. Conf. on Computer Vision (ICCV). IEEE, pp. 2572–2578.

Shawe-Taylor, J., Cristianini, N., 2004. Kernel Methods for Pattern Analysis. Cambridge University Press.

Shirazi, S., Harandi, M.T., Sanderson, C., Lovell, B.C., 2012. Clustering on Grassmann manifolds via kernel embedding with application to action analysis. In: IEEE Internat. Conf. on Image Processing.

Subbarao, R., Meer, P., 2009. Nonlinear mean shift over Riemannian manifolds. Int. J. Comput. Vision 84 (1), 1–20.

Tikhonov, A.N., Arsenin, V.Y., 1977. Solutions of Ill-Posed Problems. V.H. Winston & Sons/John Wiley & Sons, Washington, D.C./New York.

Turaga, P., Chellappa, R., 2009. Locally time-invariant models of human activities using trajectories on the Grassmannian. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 2435–2441.

Turaga, P., Chellappa, R., Subrahmanian, V., Udrea, O., 2008. Machine recognition of human activities: A survey. IEEE Trans. Circ. Systems Video Technol. 18 (11), 1473–1488.

Turaga, P., Veeraraghavan, A., Srivastava, A., Chellappa, R., 2011. Statistical computations on Grassmann and Stiefel manifolds for image and video-based recognition. IEEE Trans. Pattern Anal. Machine Intell. 33 (11), 2273–2286.

Tuzel, O., Porikli, F., Meer, P., 2008. Pedestrian detection via classification on Riemannian manifolds. IEEE Trans. Pattern Anal. Machine Intell. (PAMI) 30, 1713–1727.

Wang, H., Ullah, M.M., Kläser, A., Laptev, I., Schmid, C., 2009. Evaluation of local spatio-temporal features for action recognition. In: British Machine Vision Conference (BMVC).

Wang, T., Shi, P., 2009. Kernel Grassmannian distances and discriminant analysis for face recognition from image sets. Pattern Recognition Lett. 30 (13), 1161–1165.

Wang, Y., Mori, G., 2009. Human action recognition by semilatent topic models. IEEE Trans. Pattern Anal. Machine Intell. 31 (10), 1762–1774.

Weinland, D., Ronfard, R., Boyer, E., 2011. A survey of vision-based methods for action representation, segmentation and recognition. Computer Vision and Image Understanding 115 (2), 224–241.

Wolf, L., Shashua, A., 2003. Learning over sets using kernel principal angles. J. Machine Learn. Res. 4, 913–931.

Wong, Y., Harandi, M., Sanderson, C., Lovell, B.C., 2012. On robust biometric identity verification via sparse encoding of faces: Holistic vs local approaches. In: Internat. Joint Conf. on Neural Networks (IJCNN), pp. 1–8.

Wu, X., Xu, D. Duan, L., Luo, J., 2011. Action recognition using context and appearance distribution features. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 489–496.

Yan, S., Xu, D., Zhang, B., Zhang, H.-J., Yang, Q., Lin, S., 2007. Graph embedding and extensions: A general framework for dimensionality reduction. IEEE Trans. Pattern Anal. Machine Intell. 29 (1), 40–51.