Image and Vision Computing 55 (2016) 138–148

Contents lists available at ScienceDirect

Image and Vision Computing

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

Statistical adaptive metric learning in visual action feature set recognition☆

Shuanglu Dai*, Hong Man

Department of Electrical and Computer Engineering, Stevens Institute of Technology, 1 Castle Point on Hudson, Hoboken, NJ NJ07030, USA

ARTICLE INFO

Article history:
Received 22 August 2015
Received in revised form 22 February 2016
Accepted 7 April 2016
Available online 16 April 2016

Keywords:
Feature set classification
Hybrid statistic modeling
Metric learning
Manifold selection

ABSTRACT

Great variances in visual features often present significant challenges in human action recognitions. To address this common problem, this paper proposes a statistical adaptive metric learning (SAML) method by exploring various selections and combinations of multiple statistics in a unified metric learning framework. Most statistics have certain advantages in specific controlled environments, and systematic selections and combinations can adapt them to more realistic “in the wild” scenarios. In the proposed method, multiple statistics, include means, covariance matrices and Gaussian distributions, are explicitly mapped or generated in the Riemannian manifolds. Typically, d-dimensional mean vectors in R^d are mapped to a $R^{d\times d}$ space of symmetric positive definite (SPD) matrices Sym_d^+. Subsequently, by embedding the heterogeneous manifolds in their tangent Hilbert space, subspace combination with minimal deviation is selected from multiple statistics. Then Mahalanobis metrics are introduced to map them back into the Euclidean space. Unified optimizations are finally performed based on the Euclidean distances. In the proposed method, subspaces with smaller deviations are selected before metric learning. Therefore, by exploring different metric combinations, the final learning is more representative and effective than exhaustively learning from all the hybrid metrics. Experimental evaluations are conducted on human action recognitions in both static and dynamic scenarios. Promising results demonstrate that the proposed method performs effectively for human action recognitions in the wild.

1. Introduction

Set classification has been studied within computer vision communities for a long period of time. In gait recognition, for example, frame by frame static features of a certain object are considered as a feature set. Similarly in human action recognition, Spatial–Temporal features uniformly extracted from frames of an action atom are considered as a feature set [1]. In addition, image sets have been commonly used in face recognitions [2,3]. The task of feature set classification is to classify an input feature set to one of the sets in the training gallery [4]. Compared to image sets, feature sets are more diverse. They cannot be easily assumed to follow certain distribution or lie in some scale and affine invariant linear subspace. One of the effective techniques handling such problem is by using statistical representations to substitute the original feature samples. For action recognition in the wild scenarios, combinations of statistics from lower-order to higher-order have shown promising representation capabilities, while how to combine these multiple statistics in a near optimal way remains a technical challenge [4,5].

In general, three types of statistics have been commonly applied on set modeling, i.e. sample-based statistics (SAS) [6-9], subspace-based statistics (SUS) [10-16] and distribution-based statistics (DIS) [2,3]. Utilizing affine transformation and centroid of samples, sample-based statistics represent d-dimension feature sets with first-order statistics in the R^d space. A great advantage of SAS is that samples are considered as vectors so the nearest neighbor (NN) classification can be easily implemented with unified distance measures. But sample tests performed at every individual sample are often computationally expensive. Well-known sample-based methods include Minimum Mean Discrepancy (MMD) [17], Affine (Convex) Hull based Image Set Distance (AHISD, CHISD) [18], Set-to-Set Distance Metric Learning (SSDML) [9] and Information Theoretic Metric Learning (ITML) [19]. Differing from sample-based statistics, subspace-based statistics analyze sets lying on a specific Riemannian manifold. By learning the kernel functions or statistical metrics, the subspaces are projected back to the Euclidean spaces. The distance measures from the Riemannian manifold to the Euclidean space is often considered as the true geodesic distances, which lie in a Hilbert space [20]. Distance discriminant functions are

☆ This paper has been recommended for acceptance by Ling Shao.
* Corresponding author.
E-mail address: sdai1@stevens.edu (S. Dai).

http://dx.doi.org/10.1016/j.imavis.2016.04.003
0262-8856/© 2016 Elsevier B.V. All rights reserved.

then performed on the Hilbert space, and recognitions can finally be achieved by using Nearest Neighbors (NNs) method. Second-order statistic based methods have better representation of the data, but it is hard to design a discriminant function with a unified distance measure for the manifolds. Typical subspace-based methods include Mutual Subspace Method (MSM) [11], Discriminant Canonical Correlations (DCC) [10], Manifold Discriminant Analysis (MDA) [14], Grassmann Discriminant analysis (GDA) [13], Covariance Discriminative Learning (CDL) [15], Localized Multi-Kernel Metric Learning (LMKML) [16] etc. Distribution based statistic model each sample in the feature set with a distribution, which can be expressed as an expansion of the Riemannian manifold from the 2nd-order statistic space Sym_d^+ to Sym_{d+1}^+. Such methods are often with 3rd-order statistics and may lead to complex parametric distribution comparison. Typical examples include Single Gaussian Models (SGM) [2], Gaussian Mixture Models (GMM) [3] and kernel version of ITML with DIS-based set model (DIS-ITML). Although 3rd-order statistics model sets with more consolidated representations, the hypothesis tests often require significant amount of computation in distribution comparisons.

More adaptive forms of set modeling methods have been proposed by combining multiple statistical metrics in certain heuristic ways. Some of the recent hybrid statistical models include Projection Metric Learning (PML) on Grassmann manifold and hybrid Euclidean-and-Riemannian Metric Learning (HERML) [21]. The main idea of multiple statistic combination is to project measurements from multiple heterogeneous spaces into high-dimensional Hilbert spaces. The key issue then becomes the learning of unified discriminant functions from the training sets. Due to the simplicity of the subspace mapping, discriminant functions can frequently be found from single statistics. For instance, Set-to-Set Distance Metric Learning (SSDML) learns a proper metric between pairs of single vectors in Euclidean space to obtain more accurate set-to-set affine hull based distance for classification; Localized Multi-Kernel Metric Learning (LMKML) maps the 3rd-order statistics into Euclidean spaces by learning a unified metric discriminant function through reproducing kernel Hilbert spaces (RKHS). While hybrid multiple statistics perform well in realistic outdoor scenarios, their unified discriminant functions are often more difficult to design. Addressing the subspace discrimination problems, Shao et al. proposed a kernelized multiview projection (KMP) for action feature set recognition. KMP discriminatively assigns weights to multiple kernelized sets with a single feature to achieve a low-dimensional subspace. However, weighting kernels directly in the linear subspaces is not an optimal way for learning kernelized sets with multiple features in different scales [22].

Aiming at classification for sets with multiple features, this paper proposes an adaptive subspace analysis method for learning hybrid statistical metrics. The analyzed single or multiple statistics can be used to classify sets through various feature combinations in different scales. Inspired by the discriminant function design in the second-order based methods, LogDet divergence is introduced as a unified discriminant function for our metric learning. With this discriminant function, our method effectively unifies different statistics into a common measurement. Thus nearest neighbor method can be easily performed for classification. The whole process of modeling and learning consists of several steps. Firstly, heterogeneous statistics including mean, covariance matrix and Gaussian distribution are introduced to project data into high-dimensional Hilbert spaces. Typically, d-dimensional mean vectors represent samples from R^d to Sym_d^+ expanded by a Point-to-Set projection; covariance matrices lie in Riemannian manifold Sym_d^+ and multivariate Gaussian distributions expand the second order statistics into Riemannian manifold Sym_{d+1}^+ expressed by relative entropy. Secondly, by embedding the heterogeneous spaces into high-dimensional Hilbert spaces, the Mahalanobis distance is introduced as our discriminant metric. Then, the Hilbert space selection is conducted based on the minimum Hilbert subspaces. The hybrid statistics are then reduced to single or multiple statistic combination. Finally, with LogDet divergence that maps all the Hilbert space points into R^d, a constrained kernel learning is performed. Recognitions are mainly conducted on video sequences in both static and dynamic scenarios using spatial image features such as edges, SIFT, HOG, and texture features.

2. Statistical feature set modeling

2.1. Statistics and subspace embedding

2.1.1. Data statistics

Let $X = [X_1, \dots, X_N]$ denotes the training set formed by N feature sets, where $X_i = [x_1, x_2, \dots, x_M] \in R^{n_i \times d}$ indicates the i-th feature set, $1 \leq i \leq N$, and n_i is the number of samples in this set. It is known that the kernel function is always defined by firstly mapping the original features to a high dimensional Hilbert space, that is $\phi : R^d \rightarrow F$ or $Sym^+ \rightarrow F$, and then calculating the dot product of high dimensional statistics Φ_i and Φ_j in the new space. Considering ϕ as an explicit mapping with the statistical kernels, Φ_i^r denotes the high dimensional feature of r-th statistics extracted from the feature set X_i. Here, $1 < r < R$ and R is the number of statistics being used.

We uniformly map feature set $X_i, 1 \leq i \leq N$ with following three statistics: sample-based, subspace-based and distribution-based statistics.

Sample-based statistics (SAS): Supported by Bregman divergence, mean vector is considered as one of the important properties describing the probability distributions. It is often used to measure the central tendency of set of samples. Given sample $x_k \in X_i$, $1 \leq k \leq M$, the mean vector μ_i of X_i is computed as: $u_i = \frac{1}{M}\sum_{k=1}^{M} x_k$.

Subspace-based statistics (SUS): Within subspaces derived from eigen-decomposition, set variances are influenced by covariant matrix. Given sample $x_k \in X_i$, $1 \leq k \leq M$, the covariant matrix C_i of X_i is computed as: $C_i = \frac{1}{M-1}\sum_{k=1}^{M}(x_k - \mu_i)(x_k - \mu_i)^T$.

Distribution-based statistics (DIS): Gaussian distribution is a very commonly occurring probability distribution, which is a continuous distribution with the maximum entropy for a given mean and variance. Therefore, the d-dimensional distribution of set X_i is modeled as a Single Gaussian Model (SGM) with an estimated d-dimensional mean vector $\hat{m}_{k,i}, 1 \leq k \leq d$ and a covariance matrix $\hat{C}_i$: $x \sim N\left(\hat{m}_i, \hat{C}_i\right)$.

2.1.2. Subspace embedding and canonical correlation-based selection

A point μ in the Euclidean space R^d can be mapped into a symmetrical positive definite matrix $\left\{\mu\mu^T | R^{d\times d}, |\mu\mu^T| > 0\right\} \in Sym_d^+$. For DIS, the space of Gaussian distribution is able to be embedded into a Riemannian manifold Sym_{d+1}^+ [20].

Theorem 1. *Let $G = \{\gamma|dx|, x \in R^d\}$ be a space of normal distribution, where $|dx|$ is Lebesgue measure. Then its positive definite affine space Aff_d^+ has an explicit embedding $Aff_d^+ \rightarrow Sym_{d+1}^+$ lying on the Riemannian symmetric space Sl_{d+1}/SO_{d+1}.*

Proof. Denote an affine group of G in R^d: $Aff_d = \{(m, Q)|x \rightarrow Qx + m, Q \in G_d, m \in R^d\}$ acts transitively on G by $\gamma|dx| \rightarrow (m, Q) \otimes \gamma|dx|$, where $\otimes$ denotes the transitive operator.

Assume $\gamma_0|dx| = (2\pi)^{-d/2} e^{-\frac{1}{2}|x|^2}|dx|$ as an standard Gaussian distribution on R^d, where $\pi^{-1}(\gamma_0|dx|) = O_d$ is a positive measure. The transitive operation $(m, Q) \otimes \gamma|dx|$ can be explicitly written as

$$(m, Q) \otimes \gamma|dx| = (2\pi)^{-d/2}(\det Q)^{-1} e^{-\frac{1}{2}|Q^{-1}(x-m)|^2}|dx| \tag{1}$$

only when $\det Q > 0$. Therefore, we add a restriction θ to keep affine group positive definite $\theta : Aff_d^+ = \left\{(m, Q)|\det Q > 0\right\}$. Since the