Riemannian manifold-based SVM for human activity classification in images

RIEMANNIAN MANIFOLD-BASED SUPPORT VECTOR MACHINE FOR HUMAN ACTIVITY CLASSIFICATION IN IMAGES

Yixiao Yun *Irene Yu-Hua Gu*

Department of Signals and Systems
Chalmers University of Technology, Sweden
{yixiao, irenegu}@chalmers.se

Hamid Aghajan

Department of Electrical Engineering
Stanford University, USA
aghajan@stanford.edu

ABSTRACT

This paper addresses the issue of classification of human activities in still images. We propose a novel method where part-based features focusing on human and object interaction are utilized for activity representation, and classification is designed on manifolds by exploiting underlying Riemannian geometry. The main contributions of the paper include: (a) represent human activity by appearance features from image patches containing hands, and by structural features formed from the distances between the torso and patch centers; (b) formulate SVM kernel function based on the geodesics on Riemannian manifolds under the log-Euclidean metric; (c) apply multi-class SVM classifier on the manifold under the one-against-all strategy. Experiments were conducted on a dataset containing 2750 images in 7 classes of activities from 10 subjects. Results have shown good performance (average classification rate of 95.83%, false positive rate of 0.71%). Comparisons with three other related classifiers provide further support to the proposed method.

Index Terms— Human activity classification, Riemannian manifold, covariance descriptor, symmetric positive definite matrices, support vector machine (SVM).

1. INTRODUCTION

Recognizing human activities from visual data is one of the most important research topics in computer vision. There are many potential applications, for example, visual surveillance, image database retrieval and indexing, ambient intelligence, and computer-assisted elderly care [1].

Many existing approaches deal with recognition of human activities in video. [2] represents each video segment by a covariance matrix of optical flow, and conducts classification through sparse representation. [3] characterizes each action as a trajectory on the shape manifold, and employs a graphical model for classification. While the temporal information and motion cues in video provide discriminative features for activity classification, many human activities can be identified from key frames or individual images by purely exploiting static cues [4] [5]. Several methods on activity classification from still images have recently been proposed. [6] recognizes sport actions based on body-pose features described by circular histograms of oriented rectangles (CHORs). The method becomes less feasible when recognizing activities of daily living in indoor environments, since body poses are often similar. The presence of object interaction can be taken into account for activity classification. [7] learns spatial groups of SIFT features for classifying persons playing musical instruments. However, interacting objects can not be precisely detected and efficiently described by low-level SIFT descriptors.

Motivated by the above issues, we propose a novel method for classification of human activities in still images. The method adopts a part-based approach by focusing on image patches containing left and right hands of a person, where the interacting objects are likely to be attached. Features extracted from image patches as well as structural features are then combined and formed as points on the manifold, where a SVM classifier with a special kernel is applied. The main novelties of this paper include: representing human activity by appearance features in image patches and structural features by torso and patch distances; defining SVM with a kernel on Riemannian manifolds under the log-Euclidean metric; formulating multi-class SVM on the manifold under a one-against-all strategy.

The remainder of this paper is organized as follows: Section 2 gives a big picture of the proposed work. Section 3 briefly reviews the related work. Section 4 and 5 describe the proposed feature representation and classification methods, respectively. Section 6 shows some experimental results on an image dataset containing 7 classes of human activities. Finally, Section 7 concludes the paper.

2. THE BIG PICTURE

As shown in Fig.1, the proposed framework consists of three major parts: (a) detection of image patches centered at each hand of a target person, where hand positions are obtained from skeleton detection; (b) feature extraction from detected image patches and distance to torso, and representation of activity by the covariance matrix of features; (c) multi-class activity classification using Riemannian manifold-based SVM. The idea of using image patches centered at hands is that an

978-1-4799-2341-0/13/$31.00 ©2013 IEEE

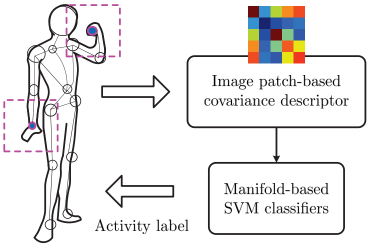
Image patch-based covariance descriptor
Manifold-based SVM classifiers
Activity label

Fig. 1. Overview of the proposed scheme for activity classification. Areas in dashed line are image patches centered at detected hand points.

interacting object is useful cue for activity classification, and it is likely to be in touch with a human hand (e.g. drinking tea, reading book). Based on hand locations obtained from the skeleton detection, image patches are formed. The basic idea of using the covariance descriptor is to combine the appearance features from image patches, and the structural features from torso and patch distances for each activity. The main motivation of applying manifold-based SVM is that each covariance activity descriptor is a symmetric positive definite matrix, descriptors for different activities do not reside in a vector space but lie on a Riemannian manifold. Hence, underlying Riemannian geometries can be exploited for training the classifier to achieve improved results.

3. RELATED WORK: REVIEW

This section briefly reviews some Riemannian geometry for the spaces associated with symmetric positive definite matrices [8], feature representation using covariance descriptors [9], and theory of support vector machines [10], for the sake of mathematical convenience in subsequent sections.

3.1. Manifold of Symmetric Positive Definite Matrices

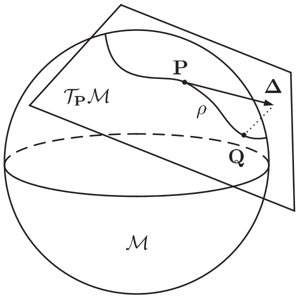

Fig. 2. Example of a 2-D manifold $\mathcal{M}$ embedded in a 3-D space $\mathbb{R}^3$. $\mathbf{P}$ and $\mathbf{Q}$ are manifold points, $\mathcal{T}_{\mathbf{P}}\mathcal{M}$ is the tangent plane at $\mathbf{P}$, $\boldsymbol{\Delta}$ is the tangent vector whose projected point on the manifold is $\mathbf{Q}$. The geodesic ρ is the shortest curve between $\mathbf{P}$ and $\mathbf{Q}$ on the manifold.

A manifold is a topological space as low dimensional subspaces embedded in a high dimensional space, that is locally Euclidean. Fig.2 depicts a 2-D manifold embedded in $\mathbb{R}^3$. The space of $d \times d$ symmetric positive definite (SPD) matrices (Sym_d^+) is an open convex cone lying on a Riemannian manifold where the tangent space at each point is endowed with a smooth inner product $\langle \cdot, \cdot \rangle_{\mathbf{P} \in \mathcal{M}}$ [8]. To compute the statistics on Sym_d^+, the *affine-invariant* metric [11] and the *log-Euclidean* metric [12] are commonly used. These two metrics are mathematically equivalent, however, numerical results can differ. This paper uses log-Euclidean metric as it is computationally more efficient [12].

As shown in Fig.2, *exponential map ($\mathcal{T}_{\mathbf{P}}\mathcal{M} \mapsto \mathcal{M}$)* and *logarithmic map ($\mathcal{M} \mapsto \mathcal{T}_{\mathbf{P}}\mathcal{M}$)* are a pair of operators mapping between the manifold $\mathcal{M}$ and the tangent space at $\mathbf{P}$:

$$\exp_{\mathbf{P}}(\boldsymbol{\Delta}) = \exp(\log(\mathbf{P}) + \boldsymbol{\Delta}) = \mathbf{Q} \tag{1}$$

$$\log_{\mathbf{P}}(\mathbf{Q}) = \log(\mathbf{Q}) - \log(\mathbf{P}) = \boldsymbol{\Delta} \tag{2}$$

where $\boldsymbol{\Delta} \in \mathcal{T}_{\mathbf{P}}\mathcal{M}$ is the tangent vector whose projected point on the manifold is $\mathbf{Q}$. *Geodesic* is the shortest curve between two points $\mathbf{P}$, $\mathbf{Q}$ on $\mathcal{M}$. The geodesic distance is computed from

$$\rho(\mathbf{P}, \mathbf{Q}) = \| \log_{\mathbf{P}}(\mathbf{Q}) \| = \| \log(\mathbf{Q}) - \log(\mathbf{P}) \| \tag{3}$$

3.2. Region Covariance

Region covariance [9] enables an effective description of object appearance features, and is shown to be robust and versatile for variations in illuminations, views and poses at modest computational cost by using integral images. Given a rectangular image region $\mathcal{R}$, let $\mathbf{f}$ be the d-dimensional feature vector for each pixel inside $\mathcal{R}$. The features can be, e.g., intensity, color, gradients, magnitudes and phase angles. The region $\mathcal{R}$ is represented by a $d \times d$ covariance matrix $\mathbf{C}_{\mathcal{R}} = \mathbb{E}[\tilde{\mathbf{f}}\tilde{\mathbf{f}}^T]$, where $\tilde{\mathbf{f}}$ is the mean-subtracted feature vector. Since covariance matrices $\mathbf{C}_{\mathcal{R}} \in Sym_d^+$, they may be viewed as connected points on a (smooth) Riemannian manifold [13].

3.3. Support Vector Machines

Support Vector Machines (SVMs) are classification method, developed under the statistical learning theory, for supervised training. A most commonly discussed form is SVMs for binary classes [10]. Given a set of labeled feature vectors $\{\mathbf{x}_i, y_i\}_{i=1}^N$ where $\mathbf{x}_i \in \mathbb{R}^l$ and $y_i \in \{-1, +1\}$, a SVM aims to find a classifier that has the minimum generalization error on the test set. This is related to finding maximum margin hyperplane, formulated by

$$\min \left(\tfrac{1}{2} \|\mathbf{w}\|^2 + \gamma \textstyle\sum_{i=1}^N \xi_i \right) \tag{4}$$
$$\text{s.t.} \quad y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1 - \xi_i \quad \forall i$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product, $\mathbf{w}$ is a weight vector, b is a bias, $\gamma > 0$ is a regularization coefficient, $\xi_i \geq 0$ is a slack variable. This optimization problem can be formed by Lagrange multiplier, and solved by applying quadratic programming to its dual form.

For nonlinear separable classes, a mapping ($\phi : \mathbb{R}^l \mapsto \mathcal{H}$) is usually applied to map the feature vectors $\mathbf{x}_i \in \mathbb{R}^l$ to a higher dimensional space where classes may be more close to linearly separable. This produces a kernel Hilbert space $\mathcal{H}$ with an inner product (kernel function) $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle_{\mathcal{H}}$. For extension of a binary SVM to a multiclass SVM, *one-against-all* or *one-against-one* strategies are often adopted for simplicity [14].

4. IMAGE PATCH-BASED COVARIANCES AS ACTIVITY DESCRIPTORS

This section describes the proposed feature representation method for human activities using a part-based approach from still images. By detecting the skeleton of human body in images, e.g. using a *Kinect* sensor, positions of hands and head, and the torso axis of human can be detected. The basic idea of detecting hand points is that interacting objects, a useful cue for activity classification, are likely to appear in the vicinity of human hands. It is also beneficial to detect the head point and the torso axis, as it may provide structural information on the hands and the body.

Given a detected hand points $\mathbf{p}_i = (x_i, y_i)^T$ $(i = 1, 2)$ of a person in an image, a normalized image patch $\mathcal{I}_i$ of size $w \times h$ centered at $\mathbf{p}_i$ is obtained. For each pixel in $\mathcal{I}_i$, a feature vector $\mathbf{f}_{i,j}$ can be formed by using the following two component vectors:

a) Appearance feature vector

$$\mathbf{g}_{i,j} = [r, g, b, |I_x|, |I_y|, |I_{xx}|, |I_{yy}|, \sqrt{I_x^2 + I_y^2}, \tan^{-1}(\frac{I_y}{I_x})]^T$$

where r, g, b are RGB values of pixel, $|I_x|$, $|I_y|$, $|I_{xx}|$, $|I_{yy}|$ are magnitudes of the first and second derivatives along x, y directions, $\sqrt{I_x^2 + I_y^2}$ and $\tan^{-1}(\frac{I_y}{I_x})$ are the gradient magnitude and orientation, respectively.

b) Structural feature vector

$$\mathbf{h}_{i,j} = [x, y, d_1, d_2, d_3]^T$$

where $(x, y)^T$ is the pixel coordinate, d_1, d_2 and d_3 are the distances between the pixel and the head point $\mathbf{p}_0 = (x_0, y_0)^T$, the other hand point $\mathbf{p}_k$ $(k \neq i, k = 1, 2)$ and the torso axis, normalized by the length of the torso axis L, respectively.

Finally, a r-dimensional feature vector $\mathbf{f}_{i,j}$ is defined for the patch related to the left (or, right) hand, as

$$\mathbf{f}_{i,j} = [(\mathbf{\Omega}\, \mathbf{h}_{i,j})^T \quad (\mathbf{g}_{i,j})^T]^T \tag{5}$$

where $\mathbf{\Omega} > 0$ is a diagonal matrix adjusting the weight of features. The image patch $\mathcal{I}_i$ is represented by an $r \times r$ covariance matrix as

$$\mathbf{C}_i = \frac{1}{|\mathcal{I}_i| - 1} \sum_{j=1}^{|\mathcal{I}_i|} \tilde{\mathbf{f}}_{i,j} \tilde{\mathbf{f}}_{i,j}^T \quad \in Sym_r^+, \quad i = 1, 2 \tag{6}$$

where $|\mathcal{I}_i|$ is the total number of pixels in $\mathcal{I}_i$, and $\tilde{\mathbf{f}}_{i,j}$ is the mean-subtracted feature vector. Finally, assuming two patches are statistically independent, a covariance matrix of $d \times d$ $(d = 2r)$ is formed for each activity by using the two image patch-related descriptors as follows:

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{i^*} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_k \end{bmatrix} \quad \in Sym_d^+ \tag{7}$$

where $i^* = \arg\min_{i=1,2} \|\mathbf{p}_0 - \mathbf{p}_i\|$ and $k \neq i^*$, $k = 1, 2$.

5. RIEMANNIAN MANIFOLD-BASED MULTI-CLASS SVM

This section presents the proposed multi-class classification scheme that learns soft-margin SVM classifiers under the one-against-all strategy by exploiting the Riemannian geometry. The basic idea is to apply a two-layer mapping for the manifold points, by first using the logarithmic mapping under the log-Euclidean metric, and then by using a Radial Basis Function (RBF). The associated kernel function is shown to be based on geodesic distances on the manifold, disregarding the choice of base point at which the tangent space is embedded.

For the k-th classifier that distinguishes the k-th class activities from all remaining ones, a base point $\boldsymbol{\mu}^k$ on the Riemannian manifold $\mathcal{M}$ is selected (e.g., by the Karcher mean [15]), where an inner product can be defined in the associated tangent space (see Section 3). The logarithmic mapping function that maps manifold points onto tangent spaces is given in (2), based on which the first-layer mapping function $(\phi_1 : Sym_d^+ \mapsto \mathcal{H}_1)$ is defined:

$$\phi_1(\mathbf{C}_n) = \log_{\boldsymbol{\mu}^k}(\mathbf{C}_n) = \log(\mathbf{C}_n) - \log(\boldsymbol{\mu}^k) = \mathbf{\Delta}_n^k \tag{8}$$

where $\mathbf{\Delta}_n^k$ is the tangent vector of $\mathbf{C}_n$ in the tangent space $\mathcal{T}_{\boldsymbol{\mu}^k}\mathcal{M}$ originated at $\boldsymbol{\mu}^k$. The second-layer mapping $(\phi_2 : \mathcal{H}_1 \mapsto \mathcal{H}_2)$ is the RBF kernel. Consider the concatenation of two mapping functions as one, the kernel function associated with this mapping function $(\phi : Sym_d^+ \mapsto \mathcal{H}_2)$ for the k-th SVM model is given by

$$\begin{aligned} \mathcal{K}^k(\mathbf{C}_n, \mathbf{C}_m) &= \langle \phi(\mathbf{C}_n), \phi(\mathbf{C}_m) \rangle_{\mathcal{H}_2} \\ &= \exp(-\lambda_k \| \log_{\boldsymbol{\mu}^k}(\mathbf{C}_n) - \log_{\boldsymbol{\mu}^k}(\mathbf{C}_m) \|^2) \\ &= \exp(-\lambda_k \| \log(\mathbf{C}_n) - \log(\mathbf{C}_m) \|^2) \\ &= \exp(-\lambda_k \rho^2(\mathbf{C}_n, \mathbf{C}_m)) \end{aligned} \tag{9}$$

where $\lambda_k > 0$ is the kernel parameter, and $\rho(\mathbf{C}_n, \mathbf{C}_m)$ is the geodesic distance between manifold points $\mathbf{C}_n$ and $\mathbf{C}_m$ defined in (3) under the log-Euclidean metric.

One of the K binary SVM classifiers that has the largest margin (or, the highest confidence) is chosen for the unknown pattern. The decision rule is given by

$$c = \arg \max_{k=1}^{K} (\alpha^k) \tag{10}$$

where $c \in \{1, \cdots, K\}$ is the label of selected activity class, and α^k is the normalized output margin of the k-th SVM model.

6. EXPERIMENTAL RESULTS

Dataset: Images used for the training and test sets are collected from the dataset [16] (140 videos selected, approx. 20 key frames per video) with provided skeleton annotations. Seven activity classes are chosen for the classifier: *drinking*, *eating*, *reading*, *phone calling*, *using laptop*, *vacuum cleaning*, and *playing guitar*. The total number of activity images is 2750. Detail about the dataset images in each class are given in Table 1. Fig.3 shows some example images from the 7 classes.

Setup: All image patches are normalized to 32 × 32 pixels. The range of regularization coefficient in (4) is $\gamma \in [76.11, 181.02]$, and the kernel parameter in (9) is $\lambda_k = 4.76 \times 10^{-3}$, and the weighting matrix in (5) is $\mathbf{\Omega} = \text{diag}(1, 1, 2, 2, 4)$. Images in the dataset are partitioned into

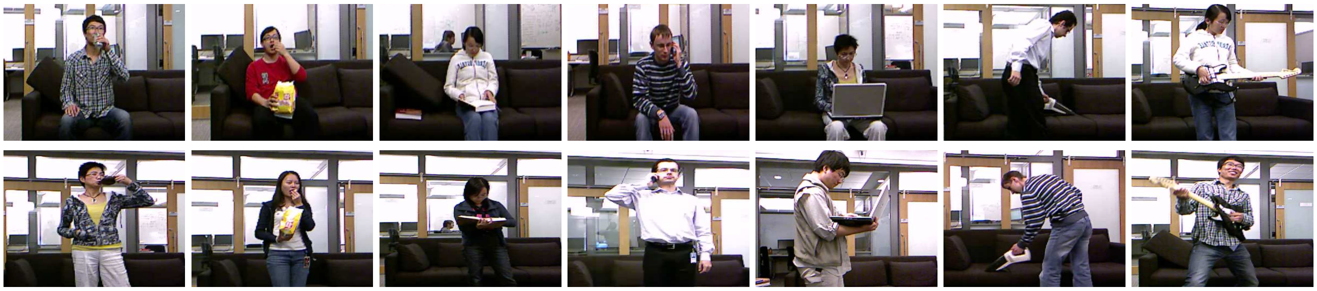

Fig. 3. Example images of human activities in each class (zoomed in for better inspection of subjects). From left to right columns: drinking, eating, reading, phone calling, using laptop, vacuum cleaning, and playing guitar.

Class #	Class name	# images	# subjects
1	drinking	400	10
2	eating	392	10
3	reading	378	10
4	phone calling	400	10
5	using laptop	400	10
6	vacuum cleaning	380	10
7	playing guitar	400	10

Table 1. Information on image dataset containing 7 classes of activities.

2 sets, where images from 7 subjects (approx. 70% in each class) are used for training, and images from the remaining 3 subjects (approx. 30%) are used for testing.

Results, Performance and Comparisons: The performance of the proposed classifier is evaluated according to classification rate, false positive rate [17]. Further, comparison are made with three closely related classification methods.

- *Method-1 (M1)* (non-manifold SVM): directly applies RBF kernel SVM to $\mathbf{C}$ in (7);
- *Method-2 (M2)*: uses the identity matrix $\mathbf{I} \in Sym_d^+$ as the base point, and applies SVM to the tangent space of $\mathbf{C}$;
- *Method-3 (M3)*: uses a global mean $\boldsymbol{\mu} \in Sym_d^+$ as the base point, and applies SVM to the tangent space of $\mathbf{C}$;

Classification rate (%)								
Method	1	2	3	4	5	6	7	Average
M1	58.33	75.00	95.00	13.33	100	82.50	95.83	74.29
M2	45.83	100	94.17	60.00	100	86.67	100	83.81
M3	36.67	83.33	92.50	54.17	99.17	85.00	100	78.69
Proposed	**87.50**	**100**	**100**	**83.33**	**100**	**100**	**100**	**95.83**

False positive rate (%)								
Method	1	2	3	4	5	6	7	Average
M1	5.69	5.71	0.88	13.03	0	2.46	0.7	4.07
M2	8.54	0	1.03	6.45	0	2.19	0	2.60
M3	10.90	2.95	1.36	7.86	0.14	2.50	0	3.67
Proposed	**2.12**	**0**	**0**	**2.82**	**0**	**0**	**0**	**0.71**

Table 2. Performance: proposed method and 3 other methods in terms of classification rate, false positive rate on the testing set.

Observing and comparing the results in Tables 2, one can see that the proposed classifier provides good classification rate while maintaining small false positive rate. Comparing with *Method-1*, *Method-2* and *Method-3*, the proposed method has significantly improved the performance.

7. CONCLUSION

The proposed activity classification scheme, using part-based features, emphasizing on interacting between hands and object and applying Riemannian manifold-based multi-class SVM, is shown to be effective in obtaining high classification rate (average 95.83%) with low false positive rate (0.71%) in our experiments. Comparisons with three closely related classification methods have provided further support to the robustness of the proposed scheme. Future study is to be conducted on tests on more images and more datasets, on extending to larger number of activity classes, and on comparisons with other state-of-the-art methods.

8. REFERENCES

[1] S. Gong, T. Xiang, "Visual Analysis of Behaviour: From Pixels to Semantics," *Springer*, 2011.

[2] K. Guo, P. Ishwar, J. Konrad, "Action recognition using sparse representation on covariance manifolds of optical flow," *AVSS*, 2010.

[3] M.F. Abdelkader, et al, "Silhouette-based gesture and action recognition via modeling trajectories on Riemannian shape manifolds," *J. CVIU*, 115(3):439–455, 2011.

[4] Y. Wang, H. Jiang, et al, "Unsupervised discovery of action classes," *CVPR*, 2006.

[5] N. Ikizler-Cinbis, et al, "Learning actions from the web," *ICCV*, 2009.

[6] N. Ikizler, R.G. Cinbis, et al, "Recognizing actions from still images," *ICPR*, 2008.

[7] B. Yao, F. Li, "Grouplet: A structured image representation for recognizing human and object interactions," *CVPR*, 2010.

[8] J.M. Lee, "Introduction to Smooth Manifolds," *Springer*, 2006.

[9] O. Tuzel, F. Porikli, P. Meer, "Region covariance: a fast descriptor for detection and classification," *ECCV*, 2006.

[10] C. Cortes, V.N. Vapnik, "Support-vector networks," *J. Machine Learning*, 20(3):273–297, 1995.

[11] X. Pennec, P. Fillard, N. Ayache, "A Riemannian framework for tensor computing," *Int'l J. Computer Vision*, 66(1):41–66, 2006.

[12] V. Arsigny, P. Fillard, et al, "Geometric means in a novel vector space structure on symmetric-positive definite matrices," *SIAM J. Matrix Analysis and Applications*, 66(1):328–347, 2008.

[13] O. Tuzel, F. Porikli, P. Meer, "Pedestrian detection via classification on Riemannian manifolds," *IEEE Trans. PAMI*, 30(10):1713–1727, 2008.

[14] C.W. Hsu, C.J. Lin, "A comparison of methods for multi-class support vector machines," *IEEE Trans. Neural Networks*, 13(2):415–425, 2002.

[15] H. Karcher, "Riemannian center of mass and mollifier smoothing," *J. Comm. Pure and Applied Math.*, 30:509–541, 1977.

[16] Microsoft Research, "MSR Daily Activity 3D dataset," `http://research.microsoft.com/en-us/um/people/zliu/actionrecorsrc/default.htm`.

[17] T.K. Moon, W.C. Stirling, "Mathematical Methods and Algorithms for Signal Processing," *Prentice Hall*, 1999.