Action Recognition by Exploring Data Distribution and Feature Correlation

Sen Wang[1,2], Yi Yang[3], Zhigang Ma[4], Xue Li[1], Chaoyi Pang[2], Alexander G. Hauptmann[3]
[1]School of ITEE, University of Queensland, Australia
[2]The Australian e-Health Research Centre, CSIRO, Australia
[3]School of Computer Science, Carnegie Mellon University, USA
[4]Department of Information Engineering and Computer Science, University of Trento, Italy
{sen.wang,xueli}@uq.edu.au {yiyang,alex}@cs.cmu.edu ma@disi.unitn.it chaoyi.pang@csiro.au

Abstract

Human action recognition in videos draws strong research interest in computer vision because of its promising applications for video surveillance, video annotation, interactive gaming, etc. However, the amount of video data containing human actions is increasing exponentially, which makes the management of these resources a challenging task. Given a database with huge volumes of unlabeled videos, it is prohibitive to manually assign specific action types to these videos. Considering that it is much easier to obtain a small number of labeled videos, a practical solution for organizing them is to build a mechanism which is able to conduct action annotation automatically by leveraging the limited labeled videos. Motivated by this intuition, we propose an automatic video annotation algorithm by integrating semi-supervised learning and shared structure analysis into a joint framework for human action recognition. We apply our algorithm on both synthetic and realistic video datasets, including KTH [20], CareMedia dataset [1], Youtube action [12] and its extended version, UCF50 [2]. Extensive experiments demonstrate that the proposed algorithm outperforms the compared algorithms for action recognition. Most notably, our method has a very distinct advantage over other compared algorithms when we have only a few labeled samples.

1. Introduction

Human action recognition has been widely studied in computer vision [18]. As an important tool for video concept understanding, action recognition plays an essential role in a number of applications, such as video surveillance [15], video annotation and retrieval [6, 10] and human-machine interactions. Nowadays, people are creating and sharing their personal videos due to the phenomenal development of network and storage technologies. When confronted with the huge volumes of videos related to human actions, an efficient mechanism of annotation is necessarily demanded to facilitate retrieval, indexing, classification and so on. Manual annotation, however, is tedious and considerably time-consuming. As a result, we focus on automatic video annotation to label human action types.

Many achievements on automatic video annotation of human actions have been published. D. Ramanan *et al*. [17] have constructed a system which tracks the actions in videos first and then annotates them by matching the actions with the existing labeled action library, which is dedicatedly built by synthesizing motions. Recently, a text-based classification approach has been proposed by Laptev *et al*. [10] for movie annotation by using movie scripts. In a follow-up work, Duchenne *et al*. [6] have proposed an approach to locate and annotate the actions with the weakly-labeled training data in realistic movies with dialogues. However, the availability of scripts is quite rare in the real-world. Moreover, some motions yield no scripts or even sound, *i.e*. *walking*. Liu *et al*. [11] directly associate low-level features of human actions with a set of attributes that are manually specified. Consequently, these high-level attributes are learned to improve the action classification.

Despite the promising results that have been achieved in the past, some challenges still exist in this research area. On one hand, the amount of labeled data is extremely scarce compared to the unlabeled data in the real world. This fact makes the supervised learning methods not very suitable since the training data are inadequate. In contrast, semi-supervised methods are able to make use of both labeled and unlabeled data for training. The learning accuracy can be greatly improved with the conjunction of small amount of labeled data and large amount of unlabeled data. In this sense, semi-supervised learning is more practical for the real-world video action recognition. However, semi-supervised learning based action recognition has been largely unaddressed. On the other hand, it is natural that two different actions can have locally common compo-

978-1-4673-1228-8/12/$31.00 ©2012 IEEE

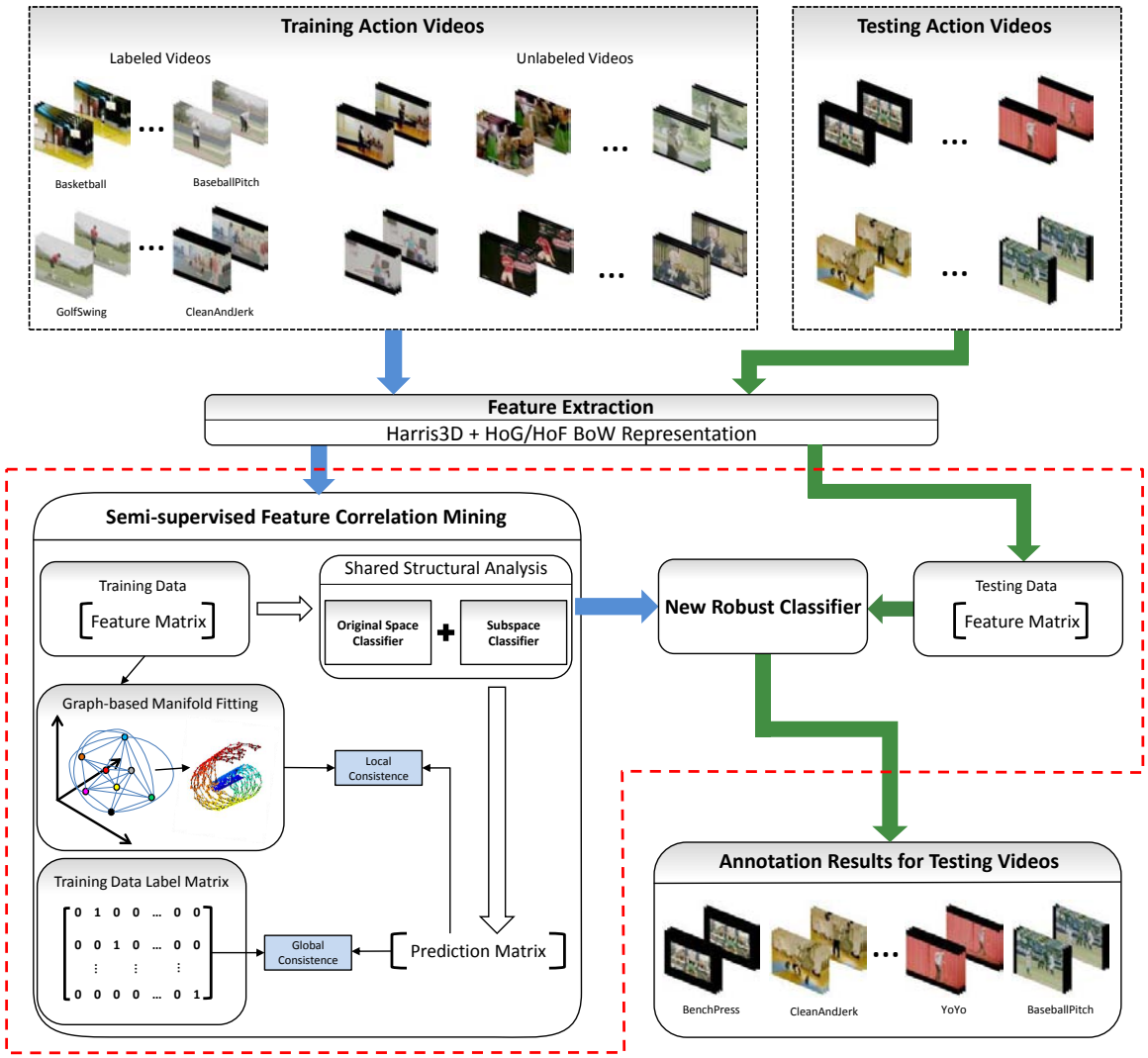


Figure 1. The illustration of the Semi-supervised Feature Correlation Mining video action recognition framework. Our semi-supervised algorithm which analyses the shared structural information is indicated with the red dashed lines.

nents, *e.g*. similar actions of arm exist in both *Tennis-Swing* and *Golf-Swing*. In [11], these common components can be viewed as the same action attribute, *e.g*. *arm pendulum-like motion*, shared by the two actions. However, it remains unclear how to manually define correlations, which are sufficient enough for different action types. In this paper, using a learning model, we directly exploit the correlations between low-level action features rather than constructing associations between low-level features and high-level attributes. Specifically, the spatio-temporal detector and descriptor applied in our work, *i.e*., Harris3D+HOG/HOF, locally characterize the actions. In that way, if we a use Bag-of-Words (BoW) model to represent the videos, similar actions should have similar occurrences of visual words. Furthermore, the aforementioned action attributes and correlations are uncovered by a shared structure learning model. We aim to apply such sort of shared structural analysis at feature level to improve the recognition performance.

In this paper, we propose a novel semi-supervised scheme which leverages shared structural analysis for action recognition. Figure 1 displays our framework for recognizing actions in videos. Features are extracted for both training and testing videos to represent them. According to the distribution of the visual features, a graph model is first constructed in training. Building upon the graph, virtual labels of the unlabeled data can be generated, during which shared structural analysis is applied to uncover the feature correlations to make the results more reliable. In this way, a classifier is trained for action recognition. The contributions of this paper can be summarised as follows:

- We apply a semi-supervised learning algorithm which considers structures shared by different words of BoW features by uncovering a low-dimensional subspace. Moreover, our framework considers the global and local structural consistency to train a discriminating and robust classifier for annotation by using $\ell_{2,1}$-norm.
- We demonstrate the advantages of combining manifold learning with feature analysis for action recognition, which is verified by extensive experiments.
- Compared with other methods, our method shows outstanding performance especially when the label information is quite scarce.

The rest of the paper is organized as follows: We briefly review the related work in Section 2. Our proposed framework is elaborated in Section 3 followed by the experiments in Section 4. Lastly, Section 5 concludes this paper.

2. Related Work

In this section, we briefly review the related research on semi-supervised learning and shared structural analysis.

The motivation of semi-supervised learning stems from the prohibitive cost of manually annotating a considerably large amount of data. The main paradigm of graph-based semi-supervised learning is achieved to utilize the relation between labeled and unlabeled data by exploring the manifold structure. A variety of applications [26, 22, 14] using graph-based semi-supervised learning have been proposed with promising performance. In these works, the vertices of the graph are the labeled and unlabeled samples while the edges are the similarities between pairs. Since the graph Laplacian is the mostly used tool to implement semi-supervised algorithms, we integrate it into our framework.

Recently, the shared structure analysis has been widely studied. Taking into account the shared information between multiple tasks, Ando *et al*. [3] have used a linear transformation matrix to characterize structural information shared by multiple tasks. The shared structure learning has been then introduced into the applications based on multi-label data [23, 8] and the classification is usually achieved through the least square regression. These research efforts have shown that shared structural learning is a powerful tool to improve the performance in data analysis. Nie *et al*. [16] have proposed a method using $\ell_{2,1}$-norm on the loss function and they apply their method to feature selection which shows prominent performance. Inspired by the work of Nie *et al*., we therefore propose to adopt the $\ell_{2,1}$-norm into the shared subspace analysis to obtain a classifier which is robust for outliers. On top of that, we extend the proposed approach to a semi-supervised way to address the shortage of labeled data. As a result, our method incorporates several advanced techniques including shared subspace analysis, the $\ell_{2,1}$-norm loss function and manifold learning to achieve better action recognition performance.

3. The Proposed Approach

In this section, we first elaborate the formulation of our method. We name it Semi-supervised Feature Correlation Mining (SFCM). Then we present the detailed solution of how to obtain the classifier.

3.1. Formulation

The visual words of video sequences are correlated as they jointly reflect the action types. Thus, it is reasonable to assume that these visual words share a common structure in a low-dimensional space. If we properly exploit such a shared structure, a more discriminative classifier for action recognition can be obtained. Motivated by [8, 23], we take the original feature space and the shared structural subspace into account jointly through the following function:

$$f(x) = v^T x + p^T Q^T x \tag{1}$$

where x is a datum, v and p are weight vectors and Q is a transformation matrix to characterize the shared information by different words in BOW feature.

Given a training set $X = [x_1, x_2, ..., x_n] \in \mathbb{R}^{d\times n}$, we can then associate it with its labels $Y = [y_1, y_2, ..., y_n]^T \in \{0, 1\}^{n\times c}$ via a statistical approach based on the above function. Note that $x_i \in \mathbb{R}^d (1 \le i \le n)$ is the i-th datum and n is the size of X. c stands for the class number and $y_i \in \mathbb{R}^c (1 \le i \le n)$ is the label vector. Denoting $V = [v_1, v_2, ..., v_c]$ and $P = [p_1, p_2, ..., p_c]$, we get the following equation:

$$f(X) = X^T V + X^T QP. \tag{2}$$

By defining $W = V + QP$ where $W \in \mathbb{R}^{d\times c}$, the above function becomes:

$$f(X) = X^T W. \tag{3}$$

Building upon (3), Ji *et al.* [8] have proposed to achieve shared subspace learning by incorporating least squares loss function:

$$\begin{aligned} &\min_{W,P,Q} \left\| X^T W - Y \right\|_F^2 + \alpha \left\| W \right\|_F^2 + \beta \left\| W - QP \right\|_F^2 \\ &s.t. \quad Q^T Q = I \end{aligned} \tag{4}$$

where I is identity matrix. Their approach explores the shared subspace between different labels. In this paper, we propose to apply the $\ell_{2,1}$-norm on the loss function, which is more robust [16, 13], and obtain the following objective:

$$\begin{aligned} &\min_{W,P,Q} \left\| X^T W - Y \right\|_{2,1} + \alpha \left\| W \right\|_F^2 + \beta \left\| W - QP \right\|_F^2 \\ &s.t. \quad Q^T Q = I \end{aligned} \tag{5}$$

To step further, we extend the above function to a semi-supervised method for its advantage in saving labeling cost while simultaneously achieving good performance as shown in [26, 23]. Most of semi-supervised learning methods assume that the nearby data points are likely to have the same label. Specifically, data points which can be connected via a path through high density regions on the data manifold are likely to have the same label. In fact, the information of density and manifold is inadequate in the real-world since the amount of labeled data is often quite small. To address this problem, a graph model is utilized to approximate the density and manifold information. Motivated by [26, 23], we further consider extending our method to semi-supervised one for its advantage in saving labeling cost while simultaneously achieving good performance. To begin with, we redefine the training data set as $X = [X_l^T, X_u^T]^T$ where $X_l = [x_1, \ldots, x_m]^T$ and $X_u = [x_{m+1}, \ldots, x_n]^T$ are the two subsets of data with labels and without labels respectively. Accordingly, its label matrix of X is $Y = [Y_l^T, Y_u^T]^T$, where $Y_l = [y_1, \cdots, y_m]^T \in \{0,1\}^{m\times c}$ and $Y_u = [y_{m+1}, \ldots, y_n]^T \in \mathbb{R}^{(n-m)\times c}$ is a matrix with all zeros. Then we introduce a predicted label matrix $F = [F_1, \ldots, F_n]^T \in \mathbb{R}^{n\times c}$, where $F_i \in \mathbb{R}^{c\times 1}$ is the predicted label vector of the i-th data x_i by the classifier. According to [25], F is consistent with the nearby points on the same manifold and is also consistent with the ground truth labels of the labeled training data. The idea of manifold and label consistency can be generalized as:

$$\begin{aligned}&\min_F \sum_{l=1}^{c}\left[\frac{1}{2}\sum_{i,j=1}^{n}(F_{il}-F_{jl})^2 A_{ij} + \sum_{i=1}^{n} U_{ii}(F_{il}-y_{il})^2\right] \\ &\quad \Rightarrow \min_F tr(F^T L F) + tr(F-Y)^T U (F-Y)\end{aligned} \tag{6}$$

where $tr(\cdot)$ denotes the trace operator. A_{ij} is an element of the local structure graph A. The graph A is defined as follows:

$$A_{ij} = \begin{cases} 1 & \text{if } x_i \in N_k(x_j) \text{ or } x_j \in N_k(x_i); \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

where $N_k(x_i)$ is the set of k-nearest neighbours of x_i. $L = S - A$ is the Laplacian matrix and S is a diagonal matrix with $S_{ii} = \sum_{j=1}^{n} A_{ij}$. U in (6) is a selection matrix and is defined as:

$$U_{ii} = \begin{cases} \infty & \text{if } x_i \text{ is labeled;} \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

Inspired by [8, 16, 23], we integrate (5) and (6) into a joint framework as follows:

$$\begin{aligned}&\min_{F,W,Q,P} tr(F^T L F) + tr(F-Y)^T U (F-Y) \\ &+ \mu\left[\alpha\|W\|_F^2 + \beta\|W - QP\|_F^2 + \left\|X^T W - F\right\|_{2,1}\right], \\ &s.t. Q^T Q = I\end{aligned} \tag{9}$$

where μ, α and β are regularization parameters. $\|W\|_F^2$ controls the complexity of the model to avoid overfitting. $\|W - QP\|_F^2$ regularizes the shared information among different features. Shared structure learning was initially proposed for multi-label learning in [8, 23]. In our work, the idea of uncovering shared structure is applied to exploit the shared information among different visual words of BoW features for a better analysis of human actions.

3.2. Solution

According to [16], a general $\ell_{2,1}$-norm minimization problem represented as:

$$\begin{aligned}&\min_U f(U) + \sum_k \|A_k U + B_k\|_{2,1} \\ &\text{s.t. } U \in \mathcal{C}\end{aligned}$$

can be solved by the following problem iteratively:

$$\begin{aligned}&\min_U f(U) + \sum_k tr((A_k U + B_k)^T D_k (A_k U + B_k). \\ &\text{s.t. } U \in \mathcal{C}\end{aligned}$$

Therefore, the objective problem in (9) can be solved by iteratively solving the following problem:

$$\begin{aligned}&\min_{F,W,Q,P} tr(F^T L F) + tr(F-Y)^T U (F-Y) \\ &+ \mu\left[\alpha\|W\|_F^2 + \beta\|W - QP\|_F^2\right] \\ &+ \mu\left(tr(X^T W - F)^T D (X^T W - F)\right) \\ &s.t. Q^T Q = I\end{aligned} \tag{10}$$

where D is a diagonal matrix with $D_{ii} = \frac{1}{2\|z^i\|_2}$, $Z = X^T W - F$ and $Z = [z^1, \ldots, z^n]^T \in \mathbb{R}^{n\times c}$. Note that in practice, $\left\|z^i\right\|_2$ could be very close to zero. In this case, we can follow the traditional regularization way and define the diagonal elements of D as $D_{ii} = \frac{1}{2\|z^i\|_2 + \varsigma}$, where ς is a small constant. When $\varsigma \to 0$, it is easy to see that $\frac{1}{2\|z^i\|_2+\varsigma}$ approximates $\frac{1}{2\|z^i\|_2}$. The same as [23], we then set the derivative of (9) w.r.t. P to zero and have:

$$P = Q^T W \tag{11}$$

Substituting (11) into (10), it becomes:

$$\begin{aligned}&\min_{F,W,Q} tr(F^T LF) + tr(F-Y)^T U(F-Y)\\&+\mu\left(tr(X^TW-F)^T D(X^TW-F)\right.\\&\left.+trW^T\left[(\alpha+\beta)I-\beta QQ^T\right]W\right)\\&s.t. Q^TQ=I,\end{aligned} \tag{12}$$

By setting the derivative of (12) w.r.t. W to zero, we have:

$$W=(M-\beta QQ^T)^{-1}XDF, \tag{13}$$

where

$$M=XDX^T+(\alpha+\beta)I \tag{14}$$

Let N be:

$$N=M-\beta QQ^T \tag{15}$$

and substitute $W = N^{-1}XDF$ into (12), the objective function becomes:

$$\begin{aligned}&\min_{F,Q} tr(F^T LF) + tr(F-Y)^T U(F-Y)\\&+\mu\left(trF^T DF - trF^T DX^T N^{-1}XDF\right)\\&s.t. Q^TQ=I,\end{aligned} \tag{16}$$

By setting the derivative of the above function w.r.t. F to zero, we then have:

$$F=(B-\mu DX^TN^{-1}XD)^{-1}UY, \tag{17}$$

where $B = L + U + \mu D$. According to the Woodbury matrix identity [7], (16) is equivalent to solving the following optimization problem after substituting F into it [23]:

$$\begin{aligned}&\max_Q trY^TUB^{-1}DX^TJ^{-1}Q(Q^TKQ)^{-1}Q^TJ^{-1}XDB^{-1}UY\\&s.t. Q^TQ=I,\end{aligned} \tag{18}$$

Let J be

$$J=M-\mu XDB^{-1}DX^T \tag{19}$$

and K be

$$K=I-\beta(M-\mu XDB^{-1}DX^T)^{-1} \tag{20}$$

For two arbitrary matrices A and B, $tr(AB) = tr(BA)$. We thus rewrite (18) as:

$$\begin{aligned}&\max_Q tr(Q^TKQ)^{-1}Q^TJ^{-1}XDB^{-1}UYY^TUB^{-1}DX^TJ^{-1}Q\\&s.t. Q^TQ=I\end{aligned} \tag{21}$$

Let C be

$$C=J^{-1}XDB^{-1}UYY^TUB^{-1}DX^TJ^{-1}, \tag{22}$$

the optimization problem becomes:

$$\begin{aligned}&\max_Q tr(Q^TKQ)^{-1}Q^TCQ\\&s.t. Q^TQ=I\end{aligned} \tag{23}$$

The above objective function can be solved by eigen-decomposition of $K^{-1}C$. Consequently, we propose an iterative algorithm to solve our objective function in Algorithm 1. It can be proved that the objective in (9) monotonically decreases until convergence by Algorithm 1.

Algorithm 1: The SFCM algorithm.

Input:
The training data $X \in \mathbb{R}^{d\times n}$;
The training data labels $Y \in \mathbb{R}^{n\times c}$;
Parameters α, β and μ.

Output:
Converged $W \in \mathbb{R}^{d\times c}$.

1: Compute the graph Laplacian matrix $L \in \mathbb{R}^{n\times n}$;
2: Compute the selection matrix $U \in \mathbb{R}^{n\times n}$;
3: Initialize $W \in \mathbb{R}^{d\times c}$ randomly;
4: Initialize $F \in \mathbb{R}^{n\times c}$ randomly.
5: **repeat**
Compute $Z \in \mathbb{R}^{n\times c}$ as: $Z = X^TW - F$
Compute the diagonal matrix D as:

$$D=\begin{bmatrix}\frac{1}{2\|z^1\|_2} & & \\ & \ddots & \\ & & \frac{1}{2\|z^n\|_2}\end{bmatrix}$$

Compute K according to (20)
Compute C according to (22)
Compute Q by eigen-decomposition using (23)
Update F according to (17)
Update W according to (13)
until *Convergence*;
6: Return W.

4. Experiments

In this section, we evaluate our method for action recognition in videos. We first present the feature used for data representation, followed by an introduction of the video datasets. Then we discuss the experimental setup and give the results lastly.

4.1. Spatial-Temporal Features

The Bag-of-Words model is popular in the field of human action recognition [10, 11, 12]. According to [21], Harris3D interest point detector [9] and HOG/HOF descriptors [10] have shown promising performance for action recognition. We therefore use this approach to extract the features for the videos. Specifically, we follow the approach recommended in [21] which randomly selects 100,000 training features and uses the k-means to build the codebook. The size of the codebook is empirically set to 1000, and then the Bag-of-Words (BoW) features for each video are represented by the histograms of the visual word occurrences. To

increase the precision, we choose the centers with the lowest error as the codebook by initializing k-means 10 times.

4.2. Datasets

Four datasets are used in our experiments which are KTH dataset [20], YouTube action dataset [12], UCF50 dataset [2] and CareMedia dataset [1].

The **KTH actions** [20] dataset records 6 categories of actions: *walking, jogging, running, boxing, hand-waving* and *hand-clapping*. Each action is performed by 25 subjects under 4 different scenarios: outdoors, outdoors with scale variation, outdoors with different clothes and indoors. In total, KTH contains 599 video clips (2391 sequences) with resolution of 160×120 pixels.

The **Youtube action** [12] dataset collects 1600 action video clips with the resolution of 320×240 pixels over 11 categories: *basketball shooting, cycling, diving, golf swinging, horse back riding, soccer juggling, swinging, tennis swinging, trampoline jumping, volleyball spiking,* and *walking with a dog*. For each category, 25 actors perform more than 4 types of actions. Youtube action dataset is much more challenging than KTH due to large variations in camera motion, viewpoint, background, *etc*.

The **UCF50 action** [2] dataset is the extension of the YouTube action dataset from 11 categories to 50 categories. Totally, it has 6681 video clips with the identical resolution with Youtube action dataset.

Figure 2. Sample frames from the CareMedia corpus.

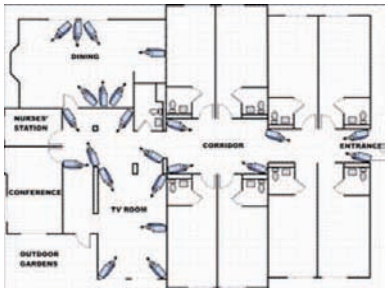


Figure 3. Camera placement in the nursing home.

The **CareMedia** [1] dataset collected by Carnegie Mellon University, consists mainly of 15 geriatric patients' activities in the public areas of a nursing home. Video and audio data are recorded by 23 cameras and microphones mounted in several fixed locations that are unobtrusive to the patients. We select a subset containing 3017 video sequences recorded by a particular camera fixed in the dining room. These video data are annotated as five categories: *Pose and/or Motor Action* (*e.g. Tremors*), *Positive* (*e.g. Smiles and Dancing*), *Physically Aggressive* (*e.g. Punching*), *Physically Non-aggressive* (*e.g. Eating*), and *Staff Activities* (*e.g. Feeding*). We show some sample frames in Figure 2 and camera placement in Figure 3.

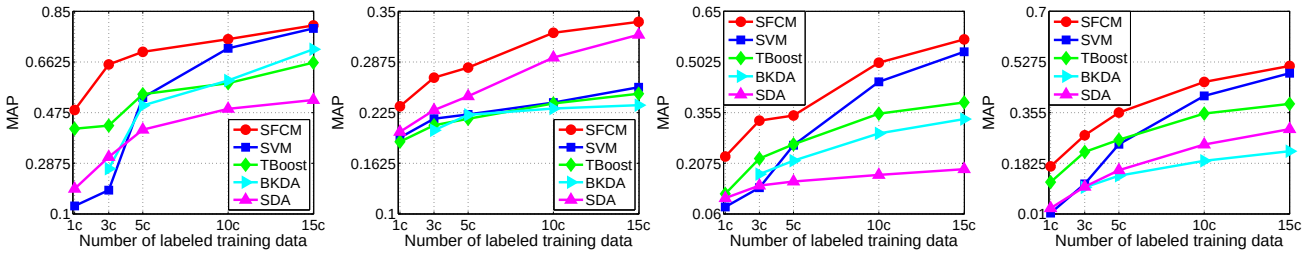
4.3. Compared Methods

To evaluate performances of our algorithm, we compare four methods which are SVM with χ^2 kernel [10], TaylorBoost (T-Boost) [19], Bayes Optimal Kernel Discriminant Analysis (BKDA) [24] and Semi-supervised Discriminant Analysis (SDA) [4]. The SVM with χ^2 kernel is widely applied in human action recognition due to its prominent performance for the Bag-of-Words model. TaylorBoost and BKDA are two supervised state of the art classification algorithms. For our algorithm, full rank Kernel Principal Analysis (with χ^2 kernel) is performed to kernelize the data. SDA is a semi-supervised discriminant analysis method, for which we use the RBF kernel SVM for classification after it processes the data.

4.4. Experiment Setup

For KTH action dataset, we use the standard data partition provided by the author in which 9 subjects (2,3,5,6,7,8,10 and 22) are utilized as testing set and the rest of data for training set. For YouTube action dataset, UCF50 action dataset and CareMedia dataset, we randomly split each dataset into training set and testing set since there are not any standard partitions available. The detailed setting for comparison is followed by the convention of the semi-supervised learning approaches. Specifically, the training set contains both labeled and unlabeled data, and the testing set is not available during the training phrase. Denote c as the class number for each dataset ($c = 5, 6, 11$ and 50 for CareMedia, KTH, Youtube and UCF50 respectively). We randomly sample m labeled videos ($m = 1, 3, 5, 10$ and 15) per category in the training set, thus resulting in $1 \times c$, $3 \times c$, $5 \times c$, $10 \times c$ and $15 \times c$ randomly labeled videos while the remaining training videos are unlabeled. We conduct experiments on 10 groups of randomly generated training and testing sets for all the methods. The average results are reported.

In our algorithm, the parameter k specifying the k-nearest neighbors for computing Laplacian matrix is empirically set to 5. r, which is the dimensionality of shared structural subspace, is set to 4 empirically as it is not sensitive. Additionally, there are three parameters, α, β and μ. We tune them from



(a) KTH (b) CareMedia (c) Youtube (d) UCF50

Figure 4. Performance comparisons on four datasets *w.r.t.* different numbers of labeled training data. When the number of labeled data is less equal than $10 \times c$, our method outperforms all other algorithms dramatically. When $15 \times c$ data are labeled, our method is ranked at top position for all datasets. Besides our method yields significantly better performances in realistic dataset, including Youtube, UCF50 and CareMedia.

$\{10^{-8}, 10^{-6}, 10^{-4}, 10^{-2}, 1, 10^{2}, 10^{4}, 10^{6}, 10^{8}\}$. For SDA [4] and SVM, we also tune the their parameters from the same range.

4.5. Experiment Results

Figure 4, Table 1 to Table 4 show the action recognition results on the four datasets w.r.t. different number of labeled training data. Note that the classification task cannot be conducted by BKDA [24] under the $1 \times c$ setting since it is unable to perform classification when there is only one sample per class. We observe that: 1) All methods achieve better results on KTH compared to those on another three datasets. This is probably due to the simplicity of KTH, which consists of simple actions collected with clean background. 2) The recognition accuracy of all methods is improved with the increase of the amount of labeled training videos. 3) Our method consistently attains the best recognition performance. 4) Our method gains much better performance when the amount of labeled data is small. For example, when only $3 \times c$ (18 out of 1524 training data for KTH) training data are labeled, our method achieves the recognition accuracy of 65.32% which is significantly better than others. Those results have indicated that our algorithm benefits from the analysis of the correlations between different visual words in a shared structure.

Next, We take the largest dataset used in this paper, UCF50, as an example to show the impact of shared structure analysis. Figure 5 shows the recognition performance

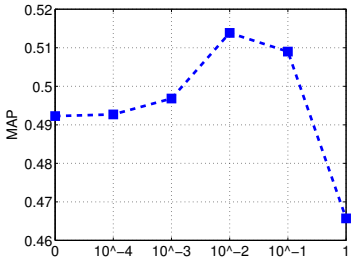


Figure 5. The performance variance of accuracy w.r.t. the parameter β with fixed α and μ.

variance w.r.t. β after fixing α and β at their optimal values, *i.e.* 10^{-2} and 10^{6}. The best performance is obtained when $\beta = 10^{-2}$. In fact, larger β indicates more shared structural information is utilized in our classifier. In contrast, when $\beta = 0$ no shared structure is utilized to contribute to the framework. The result clearly demonstrates that the performance is improved by using the shared structural information.

Table 1. Performance comparison (MAP±Standard Deviation) when $3 \times c$ training videos are labeled.

	KTH	YouTube	UCF50	CareMedia
SFCM	**0.653±0.038**	**0.332±0.025**	**0.278±0.013**	**0.268±0.062**
SVM	0.188±0.109	0.136±0.041	0.112±0.012	0.218±0.076
TBoost	0.427±0.054	0.221±0.019	0.221±0.018	0.209±0.052
BKDA	0.267±0.134	0.176±0.020	0.101±0.007	0.203 ± 0.033
SDA	0.311±0.020	0.143±0.015	0.103±0.010	0.228± 0.05

Table 2. Performance comparison (MAP±Standard Deviation) when $5 \times c$ training videos are labeled.

	KTH	YouTube	UCF50	CareMedia
SFCM	**0.700±0.034**	**0.347±0.034**	**0.355±0.011**	**0.281± 0.056**
SVM	0.531±0.111	0.260±0.064	0.247±0.020	0.223± 0.052
TBoost	0.543±0.042	0.263±0.028	0.263±0.027	0.217 ± 0.049
BKDA	0.501±0.070	0.215±0.024	0.140±0.008	0.222± 0.014
SDA	0.412±0.024	0.154±0.041	0.159±0.010	0.245± 0.025

Table 3. Performance comparison (MAP±Standard Deviation) when $10 \times c$ training videos are labeled.

	KTH	YouTube	UCF50	CareMedia
SFCM	**0.747±0.031**	**0.501±0.022**	**0.460±0.010**	**0.324 ± 0.068**
SVM	0.713±0.047	0.445±0.038	0.412±0.013	0.237± 0.007
TBoost	0.584±0.035	0.352±0.028	0.352±0.026	0.236 ± 0.040
BKDA	0.594±0.068	0.294±0.034	0.191±0.008	0.230 ±0.027
SDA	0.489±0.032	0.174±0.019	0.246±0.008	0.293 ± 0.136

Table 4. Performance comparison (MAP±Standard Deviation) when $15 \times c$ training videos are labeled.

	KTH	YouTube	UCF50	CareMedia
SFCM	**0.798±0.023**	**0.568±0.019**	**0.514±0.010**	**0.337 ± 0.027**
SVM	0.787±0.019	0.532±0.019	0.489±0.011	0.256 ± 0.052
TBoost	0.660±0.026	0.385±0.015	0.385±0.014	0.248± 0.043
BKDA	0.709±0.020	0.336±0.021	0.224±0.014	0.234± 0.013
SDA	0.522±0.027	0.190±0.024	0.299±0.012	0.321± 0.056

5. Conclusion

In this paper, we have proposed an approach to categorize human actions in videos by exploring the correlations between different visual words. First, our method simultaneously discovers the intrinsic relationship between visual words in a low-dimensional subspace to improve the performance of the holistic classification. Second, $\ell_{2,1}$-norm is applied to make the classifier robust for outliers. Finally, we extend the classifier into semi-supervised scenario to exploit on both labeled and unlabeled videos. We evaluate our framework for action video annotation on four datasets containing both synthetic and realistic ones. The experimental results show that our approach outperforms all compared algorithms, especially when the amount of labeled data is relatively small.

Acknowledgement

This work was partially supported by the National Science Foundation under Grant No. IIS-0812465, and partially supported by the National Institutes of Health (NIH) Grant No. 1RC1MH090021-01. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the National Institutes of Health.

References

[1] CareMedia Dataset. `http://www.informedia.cs.cmu.edu/caremedia/index.html`.

[2] UCF50 Action Dataset. `http://server.cs.ucf.edu/~vision/data.html`.

[3] R. K. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853, 2005.

[4] D. Cai, X. He, and J. Han. Semi-supervised discriminant analysis. In *ICCV*, 2007.

[5] J. Chen, L. Tang, J. Liu, and J. Ye. A convex formulation for learning shared structures from multiple tasks. In *ICML*, 2009.

[6] O. Duchenne, I. Laptev, J. Sivic, F. Bach, and J. Ponce. Automatic annotation of human actions in video. In *ICCV*, 2009.

[7] G. H. Golub and C. Loan. Matrix computations, 3rd edition. the johns hopkins university press. 4, 1996.

[8] S. Ji, L. Tang, S. Yu, and J. Ye. A shared-subspace learning framework for multi-label classification. *ACM Transactions on Knowledge Discovery*, 4(2):1–29, 2010.

[9] I. Laptev and T. Lindeberg. Space-time interest points. In *ICCV*, 2003.

[10] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[11] J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *CVPR*, 2011.

[12] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos "in the wild". In *CVPR*, 2009.

[13] Z. Ma, F. Nie, Y. Yang, J. Uijlings, and N. Sebe. Web image annotation via subspace-sparsity collaborated feature selection. *IEEE Transactions on Multimedia*, 2012.

[14] Z. Ma, Y. Yang, F. Nie, J. Uijlings, and N. Sebe. Exploiting the entire feature space with sparsity for automatic image annotation. In *ACM MM*, 2011.

[15] J. Nascimento, M. Figueiredo, and J. Marques. Semi-supervised learning of switched dynamical models for classification of human activities in surveillance applications. In *ICIP*, 2007.

[16] F. Nie, H. Huang, X. Cai, and C. Ding. Efficient and robust feature selection via joint l2,1-norms minimization. In *NIPS*, 2010.

[17] D. Ramanan and D. A. Forsyth. Automatic annotation of everyday movements. In *NIPS*, 2003.

[18] P. Ronald. A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6):976–990, 2010.

[19] M. Saberian, H. Masnadi-Shirazi, and N. Vasconcelos. Taylorboost: First and second-order boosting algorithms with explicit margin control. In *CVPR*, 2011.

[20] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *ICPR*, volume 3, 2004.

[21] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009.

[22] Y. Yang, F. Nie, D. Xu, J. Luo, Y. Zhuang, and Y. Pan. A multimedia retrieval framework based on semi-supervised ranking and relevance feedback. *IEEE Transactions on Pattern Analysis Machine IntelligenceI*, 34(4):723 –742, April 2012.

[23] Y. Yang, F. Wu, F. Nie, H. Shen, Y. Zhuang, and A. Hauptmann. Web and personal image annotation by mining label correlation with relaxed visual graph embedding. *IEEE Transactions on Image Processing*, 21(3):1339 –1351, 2012.

[24] D. You and A. Martinez. Bayes optimal kernel discriminant analysis. In *CVPR*, 2010.

[25] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Scholkopf. Learning with local and global consistency. In *NIPS*, 2004.

[26] X. Zhu. Semi-supervised learning literature survey. Technical report, 2005.