REPORT DOCUMENTATION PAGE

1

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
	New Reprint	-

4. TITLE AND SUBTITLE

Semi-Supervised Multiple Feature Analysis for Action Recognition

5a. CONTRACT NUMBER
W911NF-13-1-0277

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS

Zhigang Ma, Yi Yang, Xue Li, Chaoyi Pang, Alexander G. Hauptmann, Sen Wang

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES

Carnegie Mellon University
5000 Forbes Avenue

Pittsburgh, PA 15213 -3815

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)

U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
63695-CS.1

12. DISTRIBUTION AVAILIBILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

This paper presents a semi-supervised method for categorizing human actions using multiple visual features. The proposed algorithm simultaneously learns multiple features from a small number of labeled videos, and automatically utilizes data distributions between labeled and unlabeled data to boost the recognition performance. Shared structural analysis is applied in our approach to discover a common subspace shared by each type of feature. In the subspace, the proposed algorithm is able to characterize more discriminative information of each feature type. Additionally, data distribution information of each type of feature has been preserved. The aforementioned

15. SUBJECT TERMS

feature extraction, image motion analysis, machine learning, video signal processing

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Alexander Hauptmann
UU	UU	UU	UU		19b. TELEPHONE NUMBER 412-268-1448

Standard Form 298 (Rev 8/98)
Prescribed by ANSI Std. Z39.18

Report Title

Semi-Supervised Multiple Feature Analysis for Action Recognition

ABSTRACT

This paper presents a semi-supervised method for categorizing human actions using multiple visual features. The proposed algorithm simultaneously learns multiple features from a small number of labeled videos, and automatically utilizes data distributions between labeled and unlabeled data to boost the recognition performance. Shared structural analysis is applied in our approach to discover a common subspace shared by each type of feature. In the subspace, the proposed algorithm is able to characterize more discriminative information of each feature type. Additionally, data distribution information of each type of feature has been preserved. The aforementioned attributes make our algorithm robust for action recognition, especially when only limited labeled training samples are provided. Extensive experiments have been conducted on both the choreographed and the realistic video datasets, including KTH, Youtube action and UCF50. Experimental results show that our method outperforms several state-of-the-art algorithms. Most notably, much better performances have been achieved when there are only a few labeled training samples.

REPORT DOCUMENTATION PAGE (SF298)
(Continuation Sheet)

Continuation for Block 13

ARO Report Number 63695.1-CS
Semi-Supervised Multiple Feature Analysis for A...

Block 13: Supplementary Note

Approved for public release; distribution is unlimited.

Semi-Supervised Multiple Feature Analysis for Action Recognition

Sen Wang, Zhigang Ma, Yi Yang, Xue Li, Chaoyi Pang, and Alexander G. Hauptmann

***Abstract*—This paper presents a semi-supervised method for categorizing human actions using multiple visual features. The proposed algorithm simultaneously learns multiple features from a small number of labeled videos, and automatically utilizes data distributions between labeled and unlabeled data to boost the recognition performance. Shared structural analysis is applied in our approach to discover a common subspace shared by each type of feature. In the subspace, the proposed algorithm is able to characterize more discriminative information of each feature type. Additionally, data distribution information of each type of feature has been preserved. The aforementioned attributes make our algorithm robust for action recognition, especially when only limited labeled training samples are provided. Extensive experiments have been conducted on both the choreographed and the realistic video datasets, including KTH, Youtube action and UCF50. Experimental results show that our method outperforms several state-of-the-art algorithms. Most notably, much better performances have been achieved when there are only a few labeled training samples.**

***Index Terms*—Human action recognition, multiple feature learning, semi-supervised learning, shared structural analysis.**

I. Introduction

PEOPLE are more easily creating and sharing their personal videos that contain actions due to phenomenal developments in cloud computing and storage technologies. As a result, there is a heavy demand for an efficient and effective mechanism of automatic action video annotation that is able to facilitate retrieval, indexing and classification. Supervised classifiers, that only use labeled training samples, have been extensively used to address the problems. Unfortunately, labeled data are notoriously hard to obtain in the real world. By contrast, collecting unlabeled data is often effortless. When confronted with huge amounts of unlabeled data, manual annotation or labeling which is absolutely tedious and time-consuming, should always be the last choice. The goal of this work is to use multiple feature fusion to study human action recognition in video data when label information is extremely insufficient.

Human action recognition has been widely studied in computer vision [4]. The common approach is to perform feature extractions from video data and to train a classifier from the features with class information. Generally, features for action video can be divided into two groups: global features [5], [6] and local features [7], [8]. Since correlations between low-level features may provide distinctive information, more research attention [9], [10] has been put into local feature correlation mining to improve recognition results. In [11], shared structural analysis is applied to exploit multi-label correlations. Similar ideas of the shared structure learning have also been applied to many domain adaptation applications [12], [13] in which a transformation is learnt from the original feature space of both source and target domains to a subspace. This subspace is shared by all domains, which means features in every domain can be transformed into this shared subspace and then jointly learnt within it. Armed with this technique, cross-view action recognition problems have been well investigated in [14].

As mentioned above, the scarcity of labeled training samples may lead a supervised learning model to be overfitting. This work mainly focuses on recognizing actions represented by multiple features when the label information is limited. Even though semi-supervised learning and its variants are proposed to tackle the problem of insufficient labeled data for training, the ways to learn multiple features in a semi-supervised framework for action recognition with a small impact from noises and outliers have been largely ignored so far. Besides, exploiting the shared structural information has proven beneficial to action recognition in [15]. Thus, attention should also be paid to analyzing the structural information shared by action features.

To address the aforementioned challenges in action video annotation, this paper proposes a novel semi-supervised approach that does not only exploit the feature correlations within each feature type, but also automatically leverages the multiple feature fusion. First of all, semi-supervised methods, which are able to make use of both labeled and unlabeled data for training, are more suitable than supervised learning approaches for real-

Manuscript received January 16, 2013; revised June 21, 2013, August 25, 2013, and September 23, 2013; accepted October 06, 2013. Date of publication November 26, 2013; date of current version January 15, 2014. The work was supported in part by the U. S. Army Research Office (W911NF-13-1-0277), in part by the Australian Research Council the Discovery Early Career Researcher Award No. DE130101311, and in part by the Discover Project No. DP 130104614. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of ARO and Australian Research Council. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Xiao-Ping Zhang.

S. Wang and Y. Yang are with the School of Information Technology and Electrical Engineering, University of Queensland, Brisbane, Australia (e-mail: sen.wang@uq.edu.au; yi.yang@uq.edu.au).

Z. Ma is with the Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: kevinma@cs.cmu.edu).

X. Li is with the School of Information Technology and Electrical Engineering, University of Queensland, Brisbane, Australia. He is also with the Key Laboratory of Dependable Service Computing in Cyber Physical Society, Chongqing University, Chongqing, China (e-mail: xueli@itee.uq.edu.au).

C. Pang is with the Australian e-Health Research Center, CSIRO, Brisbane, Australia (e-mail: chaoyi.pang@csiro.au).

A. G. Hauptmann is with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: alex@cs.cmu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2013.2293060

world data. The recognition accuracy can be improved with a conjunction of a small amount of labeled data and a large amount of unlabeled data. Secondly, it is assumed that similar actions that are represented by the occurrence frequency of visual words, should share some common components in representation. For example, similar actions of an arm exist in both the *Tennis-Swing* and the *Golf-Swing* may have locally common components. We propose to characterize such a high-level semantic pattern through the low-level action features by applying the shared structural analysis to the Bag-of-Words (BoW) representation. By means of directly exploiting the correlations between low-level features, resemblant high-level semantic patterns are discovered between similar types of action videos. Thirdly, motivated by the latest research on video analysis that utilizes multiple features, the framework is further extended into a multiple feature based manner to achieve better classification performances. Generally speaking, the semi-supervised action video annotation is separately modeled by each type of feature with the correlations between different features simultaneously unveiled as well. In the proposed framework, training videos comprise both labeled and unlabeled. Multiple features are extracted from both the training and testing videos. For the i-th feature type, a graph model is first constructed using distributions of the i-th type of feature. Building upon this graph, virtual labels of the unlabeled data can be generated by label propagation, during which the shared structural analysis of the features is applied to uncover the feature correlations. This makes results more reliable. For each feature type, the consistency of nearby points is separately preserved, and the label prediction of the unlabeled data in the training set is made by joint consideration of the global consistency of the multiple features. In this way, a multiple feature classifier is trained for action recognition. The contributions of this paper can be summarized as follows:

- We apply a semi-supervised learning framework which analyzes structures shared by BoW features by uncovering a low-dimensional subspace based on each feature type.
- The proposed framework considers the global and local structural consistency to train a discriminating classifier for annotation.
- To maximize the holistic performance, the framework runs in a multiple feature based manner with noise handling.
- Compared with other methods, the proposed method demonstrates better performances, especially when label information is quite scarce.

The rest of this paper is organized as follows: Related works will be reviewed in Section II. The proposed framework is elaborated in Section III followed by experiments in Section IV. Lastly, Section V concludes this paper.

II. Related Work

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

A. Multiple Feature Learning

An object can be described by different features that provide different discriminating information. In light of this, research attention on feature fusion for video analysis has arisen over recent years. Feature fusion methods can be categorized into three strategies: early fusion, late fusion and multi-stage fusion. In early fusion strategy, a multimodal representation, which is constructed by heterogeneous features, is used to achieve classification tasks. A normal way to gain benefits from multiple features is to directly concatenate different types of features to form a larger feature vector. For instance, to classify human actions, Sun *et al.* [16] apply concatenation of local descriptors and holistic features as inputs to the Support Vector Machine (SVM). Though this simple fusion scheme gains a good performance, the approach usually leads to the computational burden of processing larger feature vectors. Meanwhile, it does not guarantee improved performance. It is possible that the independence among heterogeneous features may degrade the holistic performance. In contrast with early fusion methods, late fusion separately learns multiple features and builds a multimodal representation by combining learned models. In other words, late fusion occurs after independent learning for each type of feature. Farquhar *et al.* [17] propose an SVM-based late fusion algorithm, namely SVM-2K, to learn two types of features in a task of object classification. An extension from a supervised learning algorithm to a semi-supervised setting, has been proposed by Li *et al.* [18]. However, one drawback of late fusion methods is the expensive cost in learning. This is because separate learnings are carried out with respect to each feature type, and extra learning is eventually conducted for the fusion. Moreover, for most late fusion approaches, correlations of each type of feature have not been taken into account because the fusion occurs afterwards. In addition to early and late fusion strategies, the multi-stage fusion scheme has also been recently investigated. For example, Natarajan *et al.* [19] firstly combine a large set of visual and acoustic features using multiple kernel learning as the early fusion scheme. In the next stage, two different late fusion strategies are applied to MKL-based subsystems. The published results show that there exist additional performance improvements when multi-stage fusion is used.

B. Graph-Based Semi-Supervised Learning

The motivation of semi-supervised learning stems from the prohibitive cost of manually annotating a large amount of data. As one of the important branches, graph-based semi-supervised learning has attracted many research interests [20]. The main paradigm of graph-based semi-supervised learning is to utilize relations between labeled and unlabeled data by exploring the manifold structure. Since graph-based semi-supervised methods are discriminative, they have been successfully applied in a number of applications. Zhou *et al.* [21] propose a graph-based semi-supervised method that learns local and global consistency, namely LGC. Specifically, a regularization framework that iteratively predicts label information of unlabeled samples has been developed. In [22], a graph, which is constructed with a spatial Markov kernel, integrates intra-image context. Afterwards, graph-based semi-supervised learning propagates the labels of unlabeled images on the graph. Active learning is consequently combined to achieve interactive classification. Ma and *et al.* [23] use the graph-based semi-supervised framework incorporating feature selection to

learn classification information from real-world image data. In addition, graph-based semi-supervised learning has also been applied in a number of video content-based applications, including video retrieval [24], video annotation [25], action recognition [15] and multiple person tracking [26].

C. Shared Structure Analysis

Recently, shared structure analysis has been applied to multi-label learning [11], [27] and multi-task learning [28]. Taking the correlations between different labels into account, Ando *et al.* [28] have proposed an approach to minimize the total loss of a subset of predicting functions $\{f_l(x_i)\}_{l=1}^{c}$. Such a classifier is a linear combination of one classifier in the original feature space and another classifier in a low-dimensional subspace projected by a transformation matrix Θ. The classification problem is then converted into an optimization problem in the following objective function:

$$\min_{v_l,w_l,\Theta}\left(\sum_{l=1}^{c}=\left(\frac{1}{n_l}\sum_{i=1}^{n_l} loss\left(f_l\left(x_i^l\right),y_i^l\right)+r(v_l,w_l)\right)+\mu\Omega(f)\right)$$
$$\text{s.t.}\quad \Theta^T\Theta=I, \tag{1}$$

where n_l is the sample number of the l-th class. $f_l(x_i^l) = v^T x_i^l + p^T\Theta^T x_i^l$. $loss(\cdot)$ is the least squared loss function. $r(\cdot)$ and $\Omega(\cdot)$ are regularization functions using the Frobenius norm. Note that y_i^l is the ground truth label of the datum x_i which indicates whether x_i belongs to the l-th category. Ando *et al.* claim that if multiple tasks are correlated in a multi-task learning problem, benefit can be significantly obtained from a common structure shared by multiple predictors. Also, experimental results demonstrate that the shared structure learning in their linearly combined predictor is very helpful to extract the underlying correlations between tasks. In a follow-up work [11], Ji *et al.* point out it is essential to exploit correlation information contained in different labels, and propose a combined predictive function for multi-label classification. This function consists of representations in the original feature space as well as representations in a shared low-dimensional subspace. As a result, the correlation information is added to a conventional multi-label classification framework by using this joint predictive function.

III. The Proposed Approach

This section begins with an elaboration of the formulation of the proposed algorithm for action video annotation. Our method incorporates several techniques including multiple feature learning, graph-based semi-supervised learning, shared subspace analysis, the $\ell_{2,1}$-norm loss function, and manifold learning. It is named Multiple Feature Correlation Uncovering (MFCU). Following this, we present a detailed solution of how to obtain the classifier.

A. Formulation

In this work, we borrow the idea of structural learning in [11], [27] and exploit the correlations among different visual words by discovering the structural information shared by low-level features. If we properly exploit such a shared structure, a more discriminative classifier for action recognition can be obtained. Specifically, we jointly take into account the original feature space and the shared structural subspace through the following function:

$$f(X)=X^TV+X^TQP=X^TW, \tag{2}$$

where $W=V+QP$, $W\in\mathbb{R}^{d\times c}$. Q is a transformation matrix which reflects the low-dimensional subspace shared by different features. V and P are two weight matrices in the original feature space and the low-dimensional subspace, respectively. Building upon (2), Ji *et al.* [11] have proposed to learn the shared subspace by incorporating a least squared loss function. Their approach explores the shared subspace between different tasks and is easy to implement. One drawback is that the least squared loss function is sensitive to outliers. Therefore, the $\ell_{2,1}$-norm is proposed to apply to the loss function which is more sophisticated and robust [29], and obtain the following objective function:

$$\min_{W,P,Q}\|X^TW-Y\|_{2,1}+\alpha\|W\|_F^2+\beta\|W-QP\|_F^2$$
$$\text{s.t.}\quad Q^TQ=I, \tag{3}$$

where α and β are regularization parameters. $\|W\|_F^2$ controls the complexity of the model to avoid overfitting. Similar to $\|W\|_F^2$, $\|W-QP\|_F^2$ should be small as a penalty term when β is close to a large number. According to $V=W-QP$ derived from (2), the weight of representations in the original space, V, drops while the weight of representations in the transformed subspace (the shared subspace), P, increases, and vice versa. Thus, $\|W-QP\|_F^2$ regularizes the shared structure information. Through (3), we aim to construct a robust classifier that is both discriminative in the original feature space and is capable of discovering the correlations between visual words in the transformed low-dimensional subspace. Through such a joint classifier, classification performance can be further improved [15].

One limitation of the framework in [11] is that only a single type of feature is applied. Performance can be improved by applying multiple features. Another limitation is that this method relies on fully labeled training data. Our method is extended to a semi-supervised approach due to its advantage in saving labeling costs while simultaneously achieving good performance. Most semi-supervised learning methods assume that nearby points are likely to have the same labels. 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, information about density and manifold is inadequate in the real world because of the scarcity of labeled data. To deal with this problem, a graph is utilized to approximate the density and manifold information for semi-supervised learning in the framework. To begin with, the multiple feature training data set is redefined as $X_v=[X_v^l,X_v^u]$, $1\le v\le m$. m is the number of feature types. For each feature type, X_v^l and X_v^u are two subsets of data with and without labels respectively.

Inspired by [11], [15], [27], [29], we propose a joint multiple feature learning framework as follows:

$$\begin{aligned}\min_{\substack{F,F_v,W_v\\Q_v,P_v,\lambda_v}} & tr\left(F^T\sum_{v=1}^{m}\lambda_v^{\gamma}L_vF\right)+tr(F-Y)^TU(F-Y)\\ &+\mu_1\sum_{v=1}^{m}\left(\alpha\|W_v\|_F^2+\beta\|W_v-Q_vP_v\|_F^2+\|X_v^TW_v-F_v\|_{2,1}\right)\\ &+\mu_2\sum_{v=1}^{m}\|F-F_v\|_F^2\\ \text{s.t.}\quad & Q_v^TQ_v=I,\quad \sum_{v=1}^{m}\lambda_v=1,\quad \lambda_v\in[0,1],\end{aligned}\tag{4}$$

where μ_1, μ_2, α and β are regularization parameters. $\|W_v\|_F^2$ and $\|W_v-Q_vP_v\|_F^2$ undertake the same jobs as their counterparts in (3) w.r.t. each feature type. F is the global label prediction, while F_v is the label prediction of the v-th feature type. The term $\min_{F,F_v}\sum_{v=1}^{m}\|F-F_v\|_F^2$ reflects the philosophy that predictions based on each feature type should be consistent with the global type. Definition of the Laplacian matrix of the v-th feature type, L_v, can be found in [15]. Note that we add the term, $\sum_{v=1}^{m}\lambda_v^{\gamma}L_v$ to balance contributions from structural information with respect to each feature type [30]. U is a selection matrix that is defined as:

$$U_{ii}=\begin{cases}\infty & \text{if } x_i \text{ is labeled;}\\ 0 & \text{otherwise.}\end{cases}\tag{5}$$

The shared structure learning was initially proposed for multi-label learning in [11]. In our multiple feature learning framework, the idea of uncovering the shared structure is applied to exploiting shared information among different features. Moreover, this framework preserves independent structural information from each feature which contributes to a better understanding of action videos.

B. Optimization

We use an alternating approach to optimize our objective function. First, we fix $\lambda_v=1/m$ and F to optimize the other variants. Since the initialized F is the one optimized through the following objective:

$$\min_F tr(F-Y)^TU(F-Y)+tr\left(F^T\left(\sum_{v=1}^{m}\lambda_v^{\gamma}L_v\right)F\right)\tag{6}$$

The initial value of F is obtained by setting the derivative of (6) w.r.t. F to 0 as follows:

$$F=\left(U+\sum_{v=1}^{m}\lambda_v^{\gamma}L_v\right)^{-1}UY$$

After fixing F and λ_v, the optimization problem becomes:

$$\begin{aligned}\min_{\substack{F_v,W_v\\Q_v,P_v}} & \mu_1\sum_{v=1}^{m}\left(\alpha\|W_v\|_F^2+\beta\|W_v-Q_vP_v\|_F^2+\left\|X_v^TW_v-F_v\right\|_{2,1}\right)\\ &+\mu_2\sum_{v=1}^{m}\|F-F_v\|_F^2\\ \text{s.t.}\quad & Q_v^TQ_v=I\end{aligned}\tag{7}$$

By setting the derivative of the above objective w.r.t. P_v to 0, we have:

$$P_v=Q_v^TW_v\tag{8}$$

According to [31], a general $\ell_{2,1}$-norm minimization problem represented as:

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

can be solved by the following problem iteratively:

$$\begin{aligned}\min_U & f(U)+\sum_k tr((A_kU+B_k)^TD_k(A_kU+B_k),\\ \text{s.t.}\quad & U\in\mathcal{C}\end{aligned}$$

Therefore, after substituting P_v in (8), the objective problem in (7) can be solved by iteratively solving the following problem:

$$\begin{aligned}\min_{F_v,W_v,Q_v} & \mu_1\sum_{v=1}^{m}\Big(tr\left(X_v^TW_v-F_v\right)^TD_v\left(X_v^TW_v-F_v\right)\\ &+trW_v^T\left((\alpha+\beta)I-\beta Q_vQ_v^T\right)W_v\Big)\\ &+\mu_2\sum_{v=1}^{m}\|F-F_v\|_F^2\\ \text{s.t.}\quad & Q_v^TQ_v=I\end{aligned}\tag{9}$$

where D_v is a diagonal matrix with $D_{v_{ii}}=1/2\|z_v^i\|_2$, $Z_v=X_v^TW_v-F_v$ and $Z_v=[z_v^1,\ldots,z_v^n]^T\in\mathbb{R}^{n\times c}$. Note that in practice, $\|z_v^i\|_2$ could be very close to zero. In this case, we can follow the traditional regularization way and define the diagonal elements of D_v as $D_{v_{ii}}=1/(2\|z_v^i\|_2+\varsigma)$, where ς is a small constant. When $\varsigma\to 0$, it is easy to see that $1/(2\|z_v^i\|_2+\varsigma)$ approximates $1/(2\|z_v^i\|_2$. By setting the derivative of the above function w.r.t. W_v to 0, we get:

$$W_v=\left(M_v-\beta Q_vQ_v^T\right)^{-1}X_vD_vF_v,\tag{10}$$

where $M_v=X_vD_vX_v^T+(\alpha+\beta)I$. Substituting W_v into (9), the objective function becomes:

$$\begin{aligned}\min_{F_v,Q_v} & \mu_1\sum_{v=1}^{m}\left(trF_v^TD_vF_v-trF_v^TD_vX_v^TN_v^{-1}X_vD_vF_v\right)\\ &+\mu_2\sum_{v=1}^{m}\|F-F_v\|_F^2\\ \text{s.t.}\quad & Q_v^TQ_v=I,\end{aligned}\tag{11}$$

where $N_v=M_v-\beta Q_vQ_v^T$. By setting the derivative of the above objective function w.r.t. F_v to 0, we have:

$$F_v=\mu_2G_vF,\tag{12}$$

where $G_v=\left(A_v-\mu_1D_vX_v^TN_v^{-1}X_vD_v\right)^{-1}$ and $A_v=\mu_1D_v+\mu_2I$. According to the Woodbury matrix identity [32], G_v can be written as:

$$G_v=A_v^{-1}+\mu_1E_v^T\left(N_v-\mu_1X_vD_vA_v^{-1}D_vX_v^T\right)^{-1}E_v,\tag{13}$$

where $E_v = X_v D_v A_v^{-1}$. By substituting F_v in (12) into (9), we have:

$$\min_{Q_v} -\sum_{v=1}^{m} tr(F^T G_v F) \tag{14}$$

Substituting G_v in (13) into (14), the optimization problem is equivalent to the following one:

$$\max_{Q_v} \sum_{v=1}^{m} tr F^T E_v^T \left(N_v - \mu_1 X_v D_v A_v^{-1} D_v X_v^T\right)^{-1} E_v F \tag{15}$$

In (15), the term $\left(N_v - \mu_1 X_v D_v A_v^{-1} D_v X_v^T\right)^{-1}$, according to the Woodbury matrix identity [32], is rewritten as:

$$J_v^{-1} + \beta J_v^{-1} Q_v \left(Q_v^T \left(I - \beta J_v^{-1} Q_v\right)^{-1}\right) Q_v^T J_v^{-1}, \tag{16}$$

where $J_v = M_v - \mu_1 X_v D_v A_v^{-1} D_v X_v^T$. As J_v is independent on Q_v, the optimization problem therefore comes to the following objective function:

$$\begin{aligned} &\max_{Q_v} tr F^T E_v^T J_v^{-1} Q_v \left(Q_v^T \left(I - \beta J_v^{-1}\right) Q_v\right)^{-1} Q_v^T J_v^{-1} E_v F \\ &\text{s.t.} \quad Q_v^T Q_v = I \end{aligned} \tag{17}$$

For two arbitrary matrices A and B, $tr(AB) = tr(BA)$. We therefore rewrite (17) as follows:

$$\begin{aligned} &\max_{Q_v} tr\left(Q_v^T \left(I - \beta J_v^{-1}\right) Q_v\right)^{-1} Q_v^T J_v^{-1} E_v F F^T E_v^T J_v^{-1} Q_v \\ &\text{s.t.} \quad Q_v^T Q_v = I \end{aligned} \tag{18}$$

Let K_v and C_v be:

$$K_v = I - \beta J_v^{-1} \tag{19}$$

$$C_v = J_v^{-1} E_v F F^T E_v^T J_v^{-1} \tag{20}$$

After substituting K_v and C_v into (18), the objective function is reformulated as:

$$\begin{aligned} &\max_{Q_v} tr\left(Q_v^T K_v Q_v\right)^{-1} Q_v^T C_v Q_v \\ &\text{s.t.} \quad Q_v^T Q_v = I \end{aligned} \tag{21}$$

Thus, the above objective function can be solved by the eigen-decomposition of $K_v^{-1} C_v$. Next, we fix P_v, W_v and F_v to optimize F and λ_v through the following objective function:

$$\begin{aligned} \min_{F,\lambda_v} &tr(F-Y)^T U (F-Y) + tr\left(F^T \left(\sum_{v=1}^{m} \lambda_v^\gamma L_v\right) F\right) \\ &+ \mu_2 \sum_{v=1}^{m} \|F - F_v\|_F^2 \\ \text{s.t.} \quad &\sum_{v=1}^{m} \lambda_v = 1, \quad \lambda_v \in [0,1] \end{aligned} \tag{22}$$

After fixing $\lambda_v = 1/m$ and setting the derivative of (22) w.r.t. F to 0, it becomes:

$$F = \left(\sum_{v=1}^{m} \lambda_v^\gamma L_v + U + 2\mu_2 I\right)^{-1} \left(UY + \mu_2 \sum_{v=1}^{m} F_v\right) \tag{23}$$

Algorithm 1: The MFCU algorithm.

Input:
The training data are presented by m types of features $X_1, \ldots, X_m \in \mathbb{R}^{d \times n}$;
The training data labels $Y \in \mathbb{R}^{n \times c}$;
Parameters α, β and μ_1

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

Compute the graph Laplacian matrices $L_v \in \mathbb{R}^{n \times n}$;
Compute the selection matrix $U \in \mathbb{R}^{n \times n}$;
Initialize $W_v \in \mathbb{R}^{d \times c}$ and $F_v \in \mathbb{R}^{n \times c}$ randomly;
Initialize $\lambda_v = \frac{1}{m}$ and $F = (U + (\sum_{v=1}^{m} \lambda_v^\gamma L_v))^{-1} UY$;
repeat
  Compute $Z_v \in \mathbb{R}^{n \times c}$ as: $Z_v = X_v^T W_v - F_v$
  Compute the diagonal matrices D_v as:
  $$D_v = \begin{bmatrix} \frac{1}{2\|z_v^1\|_2} & & \\ & \ddots & \\ & & \frac{1}{2\|z_v^n\|_2} \end{bmatrix}$$
  Compute Q_v, F_v, W_v and P_v according to (21), (12), (10) and (8), respectively.
  Update F according to (23)
  Update λ_v according to (27)
  until *Convergence*;
Return W_v.

Now λ_v is the only variant to be solved. From the objective function, we notice that λ_v is only related to:

$$\min_{\lambda_v} tr\left(F^T \sum_{v=1}^{m} \lambda_v^\gamma L_v F\right), \quad \sum_{v=1}^{m} \lambda_v = 1, \quad \lambda_v \in [0,1] \tag{24}$$

By using a Lagrange multiplier ξ, we convert the problem to a Lagrange function as:

$$Lag(\lambda_v, \xi) = tr\left(F^T \sum_{v=1}^{m} \lambda_v^\gamma L_v F\right) - \xi\left(\sum_{v=1}^{m} \lambda_v - 1\right) \tag{25}$$

Setting the derivative w.r.t. λ_v and ξ to 0 respectively, we have:

$$\begin{cases} \gamma \lambda_v^{\gamma-1} tr(F^T L_v F) - \xi = 0 \\ \sum_{v=1}^{m} \lambda_v - 1 = 0 \end{cases} \tag{26}$$

We thus obtain λ_v by solving the following equation:

$$\lambda_v = \frac{\left(\frac{1}{tr(F^T L_v F)}\right)^{\frac{1}{(\gamma-1)}}}{\sum_{v=1}^{m} \left(\frac{1}{tr(F^T L_v F)}\right)^{\frac{1}{(\gamma-1)}}} \tag{27}$$

Consequently, an iterative algorithm is proposed to solve the objective function in Algorithm 1. The proposed iterative method in Algorithm 1 can be verified to converge by the following theorem.

Theorem 1: The objective function value shown in (4) monotonically decreases in each iteration until convergence using the iterative approach in Algorithm 1.[1]

IV. Experiments

In this section, we first introduce action video datasets, followed by a presentation of used features and compared methods. Lastly, extensive experiments are conducted to evaluate this approach and experimental results are reported and discussed.

[1]Proof can be found at https://sites.google.com/site/homepageofsenwang/.

A. Datasets and Features

In the experiments three action video datasets are used, including the KTH dataset [1], the YouTube action dataset [2] and the UCF50 dataset [3]. The **KTH actions** [1] dataset records six categories of actions. Each action is performed by 25 subjects under four different scenarios. In total, KTH contains 599 video clips (2391 sequences). The **Youtube action** [2] dataset collects 1600 action video clips of 11 categories from Youtube.com. This dataset is much more challenging than KTH due to large variations in camera motion, viewpoint, background, etc. The **UCF50 action** [3] dataset is an extension of the YouTube action dataset from 11 to 50 categories. In total, it has 6681 video clips showing identical resolution with the Youtube action dataset.

According to [33], Harris3D interest point detector [7] and HOG/HOF descriptors [34] have shown promising performance for action recognition. Besides, the MoSIFT feature [8] that treats video spatial information and temporal information separately, offers more robustness on real-world data, e.g. surveillance videos. These two features are extracted from all video data. The Bag-of-Words (BoW) model is used to represent the videos due to its popularity in the field of human action recognition. Technically, we follow the same setting utilized in [33] and randomly select two groups of 100,000 training features from HoG/HoF and MoSIFT, respectively. The unsupervised clustering algorithm, i.e. k-means, is applied to build two codebooks for these two features. To increase the precision, we choose the centers with the lowest errors as the codebook by randomly initializing k-means 10 times. The size of the two codebooks are empirically and uniformly set to 1000. For video data, the BoW is utilized to build two histograms to represent a video using two different features.

B. Compared Methods and Experimental Setup

To evaluate the performance of our framework, the proposed algorithm is compared to six state-of-the-art methods which include SVM with the χ^2 kernel [34], TaylorBoost (TBoost) [35], Semi-supervised Feature Correlation Mining (SFCM) [15], Semi-supervised Discriminative Trace Ratio analysis (SDTR) [36], simpleMKL [37] and SVM-2K [17]. SVM, TBoost and simpleMKL are three supervised state-of-the-art classification algorithms. Particularly, SVM-χ^2 has been widely applied in human action recognition due to its prominent performance for the BoW model. SFCM and SDTR are two semi-supervised algorithms. SVM-2K is a classic two-type feature learning algorithm which only deals with two types of features. Explicit feature map [38] that approximates χ^2 kernel is performed on the data for SDTR, SFCM, TBoost, SVM2K, as well as our approach. SVM-χ^2 and simpleMKL use their default kernels on the original data.

For the KTH action dataset, we use the standard data partition provided by the author: a training set (eight persons), a validation set (eight persons) and a test set (nine persons). For the YouTube action dataset and the UCF50 action dataset, we randomly split each dataset into training and testing sets. The detailed setting for comparison is followed by the convention of 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 of each dataset ($c = 6$, 11 and 50 for 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, with the remaining training videos unlabeled. The experiments are conducted on ten groups of randomly generated training and testing sets for all the methods, and average results are reported.

In the proposed algorithm, the parameter k specifying the k nearest neighbors for computing the Laplacian matrix is set as 5. P, which is the dimensionality of the shared structural subspace, and γ are set to $c - 1$ and 10 empirically as they are not sensitive. Additionally, we tune the parameters α, β and μ_1[2] from $\{10^{-4}, 10^{-2}, 1, 10^2, 10^4\}$. For SVM-$\chi^2$, SFCM, SDTR, TBoost, simpleMKL and SVM-2K, we also tune their parameters from the same range using a validation set for KTH, and 5-fold cross validation for the other two datasets. For SFCM, SDTR, TBoost, simpleMKL and SVM-χ^2, multiple features are concatenated to form a larger feature vector. For MFCU and SVM-2K, multiple features have been learnt separately. Besides accuracy, mean average precision (MAP) is used as another metric for evaluations in the experiments.

C. Experimental Results

Extensive experiments have been conducted upon three datasets in three rounds. The proposed method has been evaluated and compared with others by two measurements. Specifically, we firstly compare the proposed method to those other approaches that only apply a single type of feature. Except for our multiple feature learning approach, each compared method has been performed with both the SIFT and MoSIFT features separately, and their results have been compared in Fig. 1. Note that SVM-2K is not compared here because it leverages two different features simultaneously. Next, comparisons are made among all approaches that apply multiple features and the results, in terms of average accuracy and mean average precision, are given in Tables I and II. Lastly, the impact of shared structure analysis in the framework and the convergence demonstration are shown in Figs. 4 and 2, respectively.

From Fig. 1, it is observed that MFCU outperforms other approaches that only use one type of feature. This demonstrates that using multiple features is beneficial. In terms of average accuracy and mean average precision, our method is consistently the best on both the choreographed data (KTH) and the real-world data (Youtube and UCF50). MFCU achieves much better results especially when only a few labeled training data are available. In the case of $1 \times c$ (one labeled data per class) for the KTH dataset, for example, the accuracy and MAP of our approach score at 58.24% and 49.93% respectively, which are about two times higher than those of TBoost, SVM-χ^2 and SDTR. Compared to the second best competitor, SFCM, our multiple feature learning algorithm still has significant advantages in both accuracy and MAP.

[2]When $\mu_1(D_v - D_v X_v^T N_v^{-1} X_v D_v) + \mu_2 I$ in (11), μ_2 should be no less than the absolute value of the smallest eigenvalue of $\mu_1(D_v - D_v X_v^T N_v^{-1} X_v D_v) + \mu_2 I$ to guarantee the quadratic form of F_v is positive semi-definite.

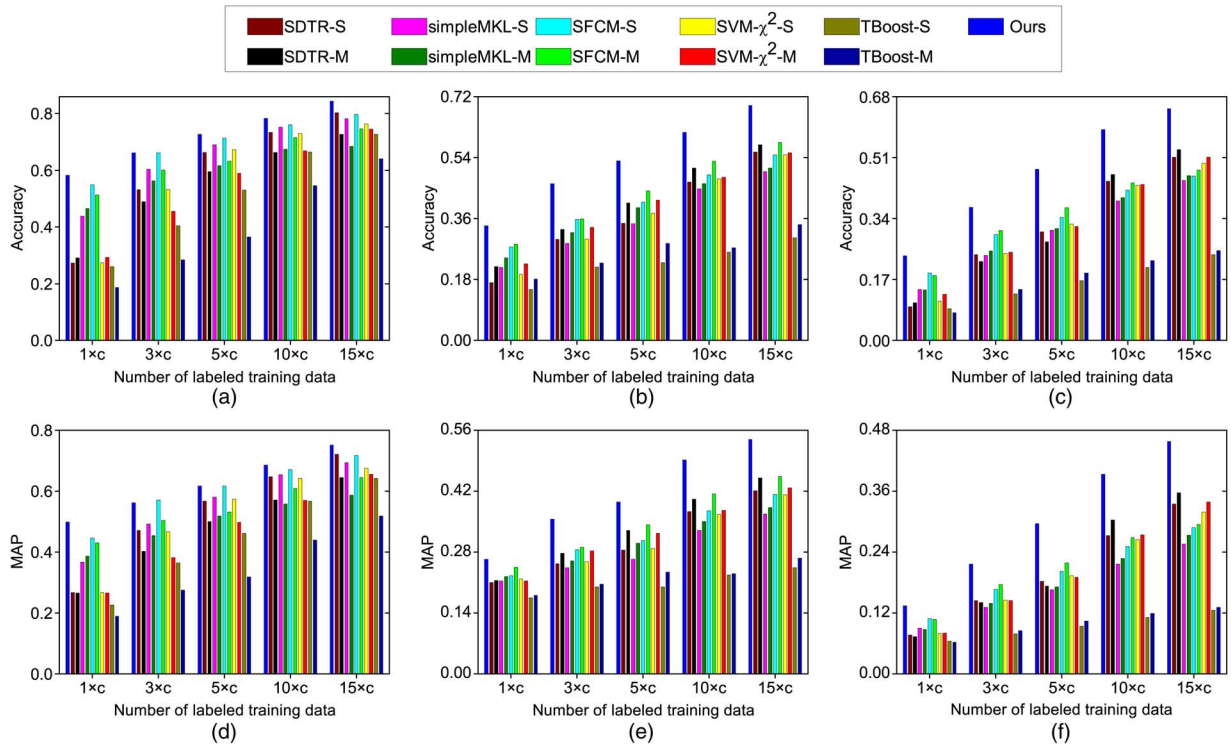


Fig. 1. Performance comparisons on the three datasets $w.r.t.$ different numbers of labeled training data. Note that each compared method has been conducted using both STIP and MoSIFT features. For example, SVM-χ^2-S and SVM-χ^2-M denote SVM with the χ^2 kernel applies the STIP and MoSIFT features respectively.

TABLE I
PERFORMANCE COMPARISON (ACCURACY)

		SDTR	simpleMKL	SFCM(Early Fusion)	SVM2K	SVM-χ^2	TBoost	Ours
KTH	$1 \times c$	0.2978 ± 0.0468	0.4737 ± 0.0833	0.5639 ± 0.0349	0.5451 ± 0.0491	0.3122 ± 0.0526	0.2572 ± 0.0248	**0.5824 ± 0.0373**
	$3 \times c$	0.5518 ± 0.0565	0.6204 ± 0.0662	0.6563 ± 0.0304	0.6447 ± 0.0457	0.5228 ± 0.0399	0.4014 ± 0.0601	**0.6614 ± 0.0379**
	$5 \times c$	0.6607 ± 0.0175	0.7092 ± 0.0072	0.7198 ± 0.0223	0.7087 ± 0.0167	0.6656 ± 0.0285	0.5745 ± 0.0569	**0.7271 ± 0.0150**
	$10 \times c$	0.7557 ± 0.0423	0.7590 ± 0.0366	0.7768 ± 0.0263	0.7731 ± 0.0229	0.7578 ± 0.0383	0.6889 ± 0.0606	**0.7838 ± 0.0211**
	$15 \times c$	0.7766 ± 0.0554	0.7914 ± 0.0242	0.8324 ± 0.0260	0.8065 ± 0.0349	0.8269 ± 0.0207	0.7620 ± 0.0162	**0.8440 ± 0.0115**
Youtube	$1 \times c$	0.2262 ± 0.0296	0.2400 ± 0.0125	0.2924 ± 0.0355	0.2789 ± 0.0318	0.2405 ± 0.0366	0.1498 ± 0.0279	**0.3383 ± 0.0278**
	$3 \times c$	0.3442 ± 0.0266	0.3261 ± 0.0257	0.3898 ± 0.0220	0.4097 ± 0.0210	0.3348 ± 0.0188	0.2392 ± 0.0332	**0.4634 ± 0.0185**
	$5 \times c$	0.4227 ± 0.0226	0.3928 ± 0.0288	0.4772 ± 0.0184	0.5011 ± 0.0319	0.4292 ± 0.0161	0.2627 ± 0.0148	**0.5308 ± 0.0346**
	$10 \times c$	0.5447 ± 0.0140	0.5036 ± 0.0116	0.5801 ± 0.0135	0.6023 ± 0.0335	0.5333 ± 0.0259	0.3019 ± 0.0175	**0.6142 ± 0.0113**
	$15 \times c$	0.6254 ± 0.0137	0.5638 ± 0.0165	0.6431 ± 0.0153	0.6722 ± 0.0122	0.5860 ± 0.0124	0.3468 ± 0.0210	**0.6941 ± 0.0162**
UCF50	$1 \times c$	0.1156 ± 0.0133	0.1638 ± 0.0116	0.2104 ± 0.0195	0.2205 ± 0.0297	0.1406 ± 0.0179	0.0890 ± 0.0193	**0.2364 ± 0.0165**
	$3 \times c$	0.2582 ± 0.0042	0.2889 ± 0.0109	0.3347 ± 0.0166	0.3649 ± 0.0104	0.2889 ± 0.0070	0.1517 ± 0.0051	**0.3708 ± 0.0160**
	$5 \times c$	0.3346 ± 0.0116	0.3749 ± 0.0105	0.4503 ± 0.0064	0.4617 ± 0.0063	0.3851 ± 0.0106	0.1976 ± 0.0087	**0.4780 ± 0.0057**
	$10 \times c$	0.5307 ± 0.0166	0.4700 ± 0.0071	0.5407 ± 0.0120	0.5629 ± 0.0036	0.5106 ± 0.0173	0.2362 ± 0.0112	**0.5878 ± 0.0100**
	$15 \times c$	0.6318 ± 0.0059	0.5307 ± 0.0048	0.5965 ± 0.0035	0.6155 ± 0.0139	0.5901 ± 0.0067	0.2736 ± 0.0093	**0.6464 ± 0.0055**

In Tables I and II, even though all approaches have used two features, MFCU still performs better than all the compared methods. Specifically, MFCU outperforms all fully supervised methods (SVM2K, SVM-2, Tboost and simpleMKL). This is because insufficient label information is unable to train a decent classifier with supervised learning algorithms. By contrast, our approach benefits from semi-supervised learning which can utilize both labeled and unlabeled data. Compared with two semi-supervised methods (SDTR and SFCM), improvements are from different sources: 1. MFCU has a more sophisticated fusion strategy than SFCM. In our late fusion strategy, local and global consistency are considered together. In this way, gains from feature fusion are augmented; 2. Compared with SDTR, the shared structural learning and the $\ell_{2,1}$-norm take advantage in terms of feature correlation mining and noise handling. From Tables I and II, it is also found that with the increase of labeled training samples, the performance of all algorithms rises. Meanwhile, the performance differences between our method and the others decrease on KTH. The differences, by contrast, are noticeable on Youtube and UCF50. We thus conclude that our method is robust for different kinds of data when the training data number varies.

To validate our claim that the proposed iterative algorithm monotonically decreases the objective function value in (4) until convergence, experiments have been conducted on all the datasets. The number of labeled training samples is set to $15 \times c$ for each dataset and the parameters are set to the median value of the tuned range. The results in Fig. 2 demonstrate by using Algorithm 1 that the objective function value monotonically decreases and converges after only a few iterations.

TABLE II
PERFORMANCE COMPARISON (MAP)

		SDTR	simpleMKL	SFCM	SVM2K	SVM-χ^2	TBoost	Ours
KTH	$1 \times c$	0.2764 ± 0.0326	0.3965 ± 0.0659	0.4735 ± 0.0489	0.4544 ± 0.0606	0.2845 ± 0.0377	0.2307 ± 0.0135	**0.4993 ± 0.0547**
	$3 \times c$	0.4648 ± 0.0480	0.5114 ± 0.0646	0.5447 ± 0.0318	0.5523 ± 0.0491	0.4435 ± 0.0408	0.3734 ± 0.0510	**0.5623 ± 0.0366**
	$5 \times c$	0.5657 ± 0.0129	0.6022 ± 0.0117	0.5998 ± 0.0320	0.6160 ± 0.0343	0.5690 ± 0.0275	0.4915 ± 0.0640	**0.6167 ± 0.0217**
	$10 \times c$	0.6634 ± 0.0436	0.6625 ± 0.0426	0.6635 ± 0.0302	0.6010 ± 0.0223	0.6633 ± 0.0434	0.5774 ± 0.0681	**0.6860 ± 0.0251**
	$15 \times c$	0.6908 ± 0.0650	0.7104 ± 0.0265	0.7335 ± 0.0119	0.6695 ± 0.0652	0.7492 ± 0.0189	0.6617 ± 0.0143	**0.7512 ± 0.0068**
Youtube	$1 \times c$	0.2145 ± 0.0167	0.2229 ± 0.0181	0.2423 ± 0.0148	0.2458 ± 0.0289	0.2191 ± 0.0174	0.1747 ± 0.0072	**0.2631 ± 0.0211**
	$3 \times c$	0.2855 ± 0.0180	0.2680 ± 0.0169	0.3171 ± 0.0216	0.3226 ± 0.0284	0.2844 ± 0.0104	0.2093 ± 0.0251	**0.3557 ± 0.0223**
	$5 \times c$	0.3443 ± 0.0128	0.2974 ± 0.0365	0.3643 ± 0.0214	0.3788 ± 0.0245	0.3305 ± 0.0144	0.2134 ± 0.0042	**0.3950 ± 0.0341**
	$10 \times c$	0.4289 ± 0.0190	0.3748 ± 0.0129	0.4492 ± 0.0233	0.4693 ± 0.0339	0.4066 ± 0.0230	0.2416 ± 0.0173	**0.4916 ± 0.0189**
	$15 \times c$	0.4863 ± 0.0113	0.4251 ± 0.0227	0.5034 ± 0.0133	0.5264 ± 0.0136	0.4452 ± 0.0112	0.2706 ± 0.0162	**0.5387 ± 0.0149**
UCF50	$1 \times c$	0.0844 ± 0.0087	0.0996 ± 0.0083	0.1117 ± 0.0126	0.1202 ± 0.0184	0.0913 ± 0.0144	0.0649 ± 0.0079	**0.1341 ± 0.0100**
	$3 \times c$	0.1704 ± 0.0016	0.1617 ± 0.0133	0.1892 ± 0.0112	0.2008 ± 0.0096	0.1749 ± 0.0043	0.0911 ± 0.0025	**0.2166 ± 0.0145**
	$5 \times c$	0.2228 ± 0.0105	0.2126 ± 0.0096	0.2671 ± 0.0040	0.2758 ± 0.0093	0.2439 ± 0.0065	0.1139 ± 0.0051	**0.2961 ± 0.0091**
	$10 \times c$	0.3789 ± 0.0172	0.2789 ± 0.0086	0.3595 ± 0.0102	0.3755 ± 0.0138	0.3357 ± 0.0145	0.1321 ± 0.0042	**0.3935 ± 0.0127**
	$15 \times c$	0.4393 ± 0.0039	0.3288 ± 0.0034	0.4204 ± 0.0085	0.4410 ± 0.0138	0.4100 ± 0.0070	0.1451 ± 0.0039	**0.4582 ± 0.0083**

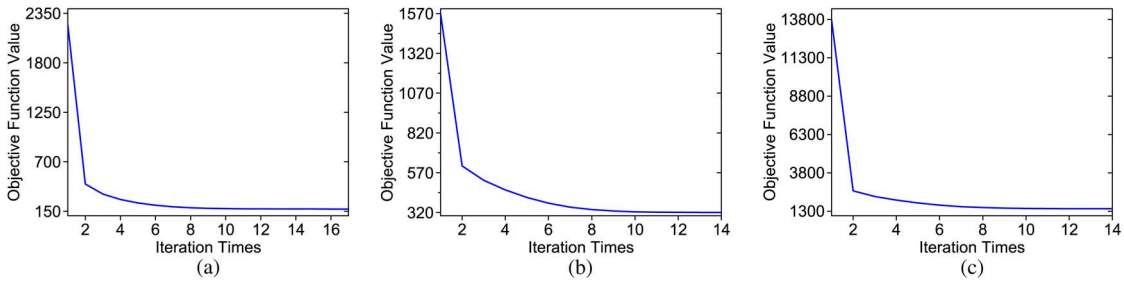


Fig. 2. The convergence curves of the objective function values in (4) by using algorithm 1 on the three datasets.

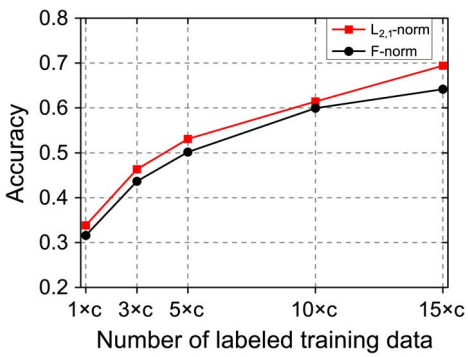


Fig. 3. Performance comparison between $\ell_{2,1}$-norm and F-norm.

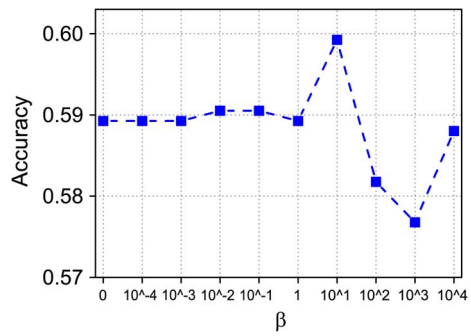


Fig. 4. The variation of accuracy $w.r.t.$ the parameter β with fixed α and μ_1.

To investigate the impact of the $\ell_{2,1}$-norm in the framework, performance comparisons between using the $\ell_{2,1}$-norm and removing it (substituted by F-norm) on the Youtube action dataset have been made. The results in Fig. 3 show that improvements are gained when using the $\ell_{2,1}$-norm for all different numbers of labeled training data. The results in Fig. 4 verify that our algorithm benefits from shared structural analysis. The real-world video dataset, Youtube, is taken as an example to demonstrate the impact of shared structure learning. We fix α and μ_1 at their optimal values, i.e. 10^0 and 10^4 respectively for $10 \times c$ labeled training data. It can be seen that as β varies from 10^{-2} to 10, the accuracy increases accordingly and reaches to the peak value when $\beta = 10$. Note that, a larger β means a larger proportion of shared structural consideration in the holistic framework, and vice versa. When $\beta = 0$, no shared structure is utilized in the framework. The results demonstrate that appropriately exploiting subspace shared by low-level features can further improve the performance. Specifically, when the number of labeled training data is $10 \times c$ (the Youtube action dataset), the extra improvement from the shared structural learning is 1.0%, while the difference between using the $\ell_{2,1}$-norm and removing it, is 1.5%. Overall, the combination of graph-based semi-supervised learning, the $\ell_{2,1}$-norm and shared structural analysis has integrally contributed to the performance boosting of our method.

D. Discussion

From the experimental results, this proposed approach, in which multi-feature learning is integrated in a graph-based semi-supervised framework, performs action recognition better than all the compared methods particularly when labeled training samples are insufficient. However, it is still

worth noting the following facts: 1. Though the ℓ_{21}-norm loss function improves performances by handling noises, its optimization requires an iterative algorithm which is more expensive than the F-norm loss function. When efficiency is a concern, the F-norm can be a substitution of the ℓ_{21}-norm in our proposed framework; 2. As indicated in [39], improved performance is not guaranteed through exploiting unlabeled data when a manifold assumption does not hold. Additionally, complementary relationships between different features may result in performance fluctuations.

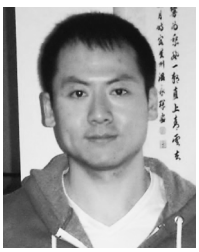
V. Conclusion

In this paper, we have proposed an approach that exploits multiple features to categorize human action videos by exploring the correlations between different visual words. Firstly, the proposed method simultaneously discovers the intrinsic relations between visual words in a low-dimensional subspace to improve the performance of the holistic classification based on each feature type. Secondly, the $\ell_{2,1}$-norm is applied to make the framework robust for noises and outliers. Thirdly, two assumptions have been utilized in the framework: 1) the label prediction should be consistent with the ground truth for each feature type; 2) the label prediction for each feature type should also be consistent with the global prediction using multiple features. Finally, the framework has been extended to semi-supervised exploiting both labeled and unlabeled videos. The framework for action video annotation has been evaluated on three datasets including both the choreographed and the realistic data. The experimental results show that our approach outperforms all the compared algorithms. The advantage is especially visible when the amount of labeled training data is quite small.

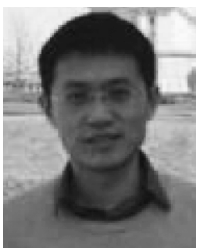
References

[1] C. Schüldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *Proc. ICPR*, 2004, vol. 3, pp. 32–36.
[2] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos "in the wild"," in *Proc. CVPR*, 2009, pp. 1996–2003.
[3] UCF50 Action Dataset [Online]. Available: http://server.cs.ucf.edu/vision/data.html
[4] R. Poppe, "A survey on vision-based human action recognition," *Image Vision Comput.*, vol. 28, no. 6, pp. 976–990, 2010.
[5] L. Wang and D. Suter, "Informative shape representations for human action recognition," in *Proc. ICPR*, 2006, pp. 1266–1269.
[6] D. Weinland, E. Boyer, and L. J. K. I. Rhône-alpes, "Action recognition using exemplar-based embedding," in *Proc. CVPR*, 2008, pp. 1–7.
[7] I. Laptev and T. Lindeberg, "Space-time interest points," in *Proc. ICCV*, 2003, pp. 432–439.
[8] M. Chen and A. Hauptmann, Mosift: Recognizing Human Actions in Surveillance Videos, Carnegie Mellon Univ., 2009, Tech. Rep..
[9] J. Sun, X. Wu, S. Yan, L.-F. Cheong, T.-S. Chua, and J. Li, "Hierarchical spatio-temporal context modeling for action recognition," in *Proc. CVPR*, Jun. 2009, pp. 2004–2011.
[10] Y. Wang and G. Mori, "Max-margin hidden conditional random fields for human action recognition," in *Proc. CVPR*, Jun. , pp. 872–879.
[11] S. Ji, L. Tang, S. Yu, and J. Ye, "A shared-subspace learning framework for multi-label classification," *ACM Trans. KDD*, vol. 4, no. 2, pp. 1–29, 2010.
[12] R. Gopalan, R. Li, and R. Chellappa, "Domain adaptation for object recognition: An unsupervised approach," in *Proc. ICCV*, Nov. 2011, pp. 999–1006.
[13] B. Gong, Y. Shi, and F. Sha, "Geodesic flow kernel for unsupervised domain adaptation," in *Proc. CVPR*, Jun. 2012, pp. 2066–2073.
[14] J. Zheng, Z. Jiang, J. Phillips, and R. Chellappa, "Cross-view action recognition via a transferable dictionary pair," in *Proc. BMVC*, 2012, pp. 125.1–125.11.
[15] S. Wang, Y. Yang, Z. Ma, X. Li, C. Pang, and A. G. Hauptmann, "Action recognition by exploring data distribution and feature correlation," in *Proc. CVPR*, 2012, pp. 1370–1377.
[16] X. Sun, M. Chen, and A. Hauptmann, "Action recognition via local descriptors and holistic features," in *Proc. CVPR Workshop*, 2009, pp. 58–65.
[17] J. Farquhar, D. Hardoon, H. Meng, J. Shawe-Taylor, and S. Szedmak, "Two view learning: Svm-2k, theory and practice," in *Proc. NIPS*, 2005.
[18] G. Li, S. Hoi, and K. Chang, "Two-view transductive support vector machines," in *Proc. SDM*, 2010, pp. 235–244.
[19] P. Natarajan, S. Wu, S. Vitaladevuni, X. Zhuang, S. Tsakalidis, U. Park, R. Prasad, and P. Natarajan, "Multimodal feature fusion for robust event detection in web videos," in *Proc. CVPR*, 2012, pp. 1298–1305.
[20] X. Zhu, Semi-Supervised Learning Literature Survey, 2005, Tech. Rep. .
[21] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Scholkopf, "Learning with local and global consistency," in *Proc. NIPS*, 2004.
[22] Z. Lu and H. Ip, "Combining context, consistency, and diversity cues for interactive image categorization," *IEEE Trans. Multimedia*, vol. 12, no. 3, pp. 194–203, 2010.
[23] Z. Ma, Y. Yang, F. Nie, J. Uijlings, and N. Sebe, "Exploiting the entire feature space with sparsity for automatic image annotation," in *Proc. ACM MM*, 2011, pp. 283–292.
[24] S. C. Hoi and M. R. Lyu, "A multimodal and multilevel ranking scheme for large-scale video retrieval," *IEEE Trans. Multimedia*, vol. 10, no. 4, pp. 607–619, 2008.
[25] T. Zhang, C. Xu, G. Zhu, S. Liu, and H. Lu, "A generic framework for video annotation via semi-supervised learning," *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 1206–1219, 2012.
[26] S. I. Yu, Y. Yang, and A. Hauptmann, "Harry Potter's Marauder's Map: Localizing and tracking multiple persons-of-interest by nonnegative discretization," in *Proc. CVPR*, Jun. 2013, pp. 3714–3720.
[27] 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 Trans. Image Process.*, vol. 21, no. 3, pp. 1339–1351, 2012.
[28] R. K. Ando and T. Zhang, "A framework for learning predictive structures from multiple tasks and unlabeled data," *J. Mach. Learn. Res.*, vol. 6, pp. 1817–1853, 2005.
[29] Z. Ma, F. Nie, Y. Yang, J. Uijlings, and N. Sebe, "Web image annotation via subspace-sparsity collaborated feature selection," *IEEE Trans. Multimedia*, vol. 14, pp. 1021–1030, 2012.
[30] Y. Feng, J. Xiao, Y. Zhuang, and X. Liu, "Adaptive unsupervised multi-view feature selection for visual concept recognition," in *Proc. ACCV*, 2012, pp. 343–357.
[31] F. Nie, H. Huang, X. Cai, and C. Ding, "Efficient and robust feature selection via joint l2, 1-norms minimization," in *Proc. NIPS*, 2010, pp. 1813–1821.
[32] G. H. Golub and C. Loan, *Matrix Computations*, 3rd ed. Baltimore, MD, USA: Johns Hopkins Univ. Press, 1996, vol. 4.
[33] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *Proc. BMVC*, 2009, pp. 124.1–124.11.
[34] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. CVPR*, 2008, pp. 1–8.
[35] M. Saberian, H. Masnadi-Shirazi, and N. Vasconcelos, "Taylorboost: First and second-order boosting algorithms with explicit margin control," in *Proc. CVPR*, 2011, pp. 2929–2934.
[36] 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 Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 723–742, Apr. 2012.
[37] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet *et al.*, "Simplemkl," *J. Mach. Learn. Res.*, vol. 9, pp. 2491–2521, 2008.
[38] A. Vedaldi and A. Zisserman, "Efficient additive kernels via explicit feature maps," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 3, pp. 480–492, Mar. 2012.
[39] Y. Yang, J. Song, Z. Huang, Z. Ma, N. Sebe, and A. Hauptmann, "Multi-feature fusion via hierarchical regression for multimedia analysis," *IEEE Trans. Multimedia*, vol. 15, no. 3, pp. 572–581, 2013.

Sen Wang received M.S. degree from Jilin University, Changchun, China, in 2009. He is pursuing his Ph.D. degree from the University of Queensland, Brisbane, Australia. He mainly focuses his research on machine learning and relevant applications in computer vision and data mining, e.g., human action recognition, social network event detection, etc.



Zhigang Ma received the B.S. and M.S. degrees from Zhejiang University, Hangzhou, China, in 2004 and 2006, respectively, and is currently working toward the Ph.D. degree from the University of Trento, Trento, Italy. His research interests include machine learning and its application to computer vision and multimedia analysis.



Yi Yang received the Ph.D. degree in computer science from Zhejiang University, in 2010. He was a postdoctoral fellow in Carnegie Mellon University. He is now a DECRA fellow at the School of Information Technology and Electrical Engineering, University of Queensland, Brisbane, Queensland, Australia. His research interests include machine learning and its applications to multimedia content analysis and computer vision, e.g., multimedia indexing and retrieval, image annotation, video semantics understanding, etc.

Xue Li received his M.Sc. and Ph.D. degrees from University of Queensland and Queensland University of Technology in 1990 and 1997. Currently, he is an Associate Professor in the School of Information Technology and Electrical Engineering at University of Queensland in Brisbane, Queensland, Australia. Xue Li's major areas of research interests and expertise include: Data Mining, Multimedia Data Security, Database Systems, and Intelligent Web Information Systems. He is a member of ACM, IEEE, and SIGKDD.

Chaoyi Pang received Ph.D. degree from the University of Melbourne. He joined the Australian e-Health Research Centre (CSIRO) in 2004 as Senior Scientist. He is a senior member of ACM. His research interests are in algorithm, data security/privacy, access control, data warehousing, data integration, database theory, and graph theory. His research performance has been evidenced by his leading authorship of a number of patents and research papers in prestigious international journals such as ACM Transactions on Database Systems (TODS) and Algorithmica.

Alexander G. Hauptmann received the B.A. and M.A. degrees in psychology from Johns Hopkins University, Baltimore, MD, the Diplom in Germany degree in computer science from the Technische Universit at Berlin, Berlin, Germany, in 1984, and the Ph.D. degree in computer science from Carnegie Mellon University (CMU), Pittsburgh, PA, in 1991. He is currently with the Faculty of the Department of Computer Science and the Language Technologies Institute, CMU. From 1984 to 1994, he was with the Informedia project for digital video analysis and retrieval, and led the development and evaluation of news-on-demand applications, where he was involved in research on speech and machine translation. His current research interests include manmachine communication, natural language processing, speech understanding and syntheses, video analyses, and machine learning.