FROM *CONCEPTS* TO *EVENTS*: A PROGRESSIVE PROCESS FOR MULTIMEDIA CONTENT ANALYSIS

ZHIGANG MA

Advisor: Prof. Dr. Nicu Sebe

Department of Information Engineering and Computer Science
University of Trento

Italy
December 2013

CONTENTS

PUBLICATIONS

This thesis consists of the following publications:

- Chapter 2:

  -**Z. Ma**, F. Nie, Y. Yang, J. Uijlings and N. Sebe: "Web Image Annotation via Subspace-Sparsity Collaborated Feature Selection". *IEEE Transactions on Multimedia*, 14(4): 1021-1030, 2012.

- Chapter 3:

  -**Z. Ma**, F. Nie, Y. Yang, J. Uijlings, N. Sebe and A. G. Hauptmann: "Discriminating Joint Feature Analysis for Multimedia Content Understanding". *IEEE Transactions on Multimedia*, 14(6): 1662-1672, 2012.

  Idea previously appeared in:

  -**Z. Ma**, Y. Yang, F. Nie, J. Uijlings and N. Sebe: "Exploiting the Entire Feature Space with Sparsity for Automatic Image Annotation". In *Proceedings of the ACM International Conference on Multimedia*, pages 283-292, 2011.

- Chapter 4:

  -**Z. Ma**, Y. Yang, N. Sebe, K. Zheng, A. G. Hauptmann: "Multimedia Event Detection Using A Classifier-Specific Intermediate Representation". *IEEE Transactions on Multimedia*, 15(7):1628-1637, 2013.

  Idea previously appeared in:

  -**Z. Ma**, Y. Yang, A. G. Hauptmann and N. Sebe: "Classifier-specific Intermediate Representation for Multimedia Tasks". In *Proceedings of the ACM International Conference on Multimedia Retrieval*, 2012.

- Chapter 5:

  -**Z. Ma**, Y. Yang, N. Sebe and A. G. Hauptmann: "Knowledge Adaptation with Partially Shared Features for Event Detection Using Few Exemplars". Pending minor revision in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

  Idea previously appeared in:

  -**Z. Ma**, Y. Yang, Y. Cai, N. Sebe and A. G. Hauptmann: "Knowledge Adaptation for Ad Hoc Multimedia Event Detection with Few Exemplars". In *Proceedings of the ACM International Conference on Multimedia*, pages 469-478, 2012.

The following are the papers published during the course of the Ph.D but not included in this thesis:

- S. Wang, **Z. Ma**, Y. Yang, X. Li, C. Pang and A. G. Hauptmann: "Semi-supervised Multiple Feature Analysis for Action Recognition". *IEEE Transactions on Multimedia*, 2014.

- Y. Yang, **Z. Ma**, Z. Xu, S. Yan and A. G. Hauptmann: "How Related Exemplars Help Complex Event Detection in Web Videos?" In *IEEE International Conference on Computer Vision*, 2013.

- **Z. Ma**, Y. Yang, Z. Xu, N. Sebe, A. G. Hauptmann: “We Are Not Equally Negative: Fine-grained Labeling for Multimedia Event Detection". In *Proceedings of the ACM International Conference on Multimedia*, pages 293-302, 2013.
- Y. Yan, Z. Xu, G. Liu, **Z. Ma** and N. Sebe: “GLocal Structural Feature Selection with Sparsity for Multimedia Data Understanding". In *Proceedings of the ACM International Conference on Multimedia*, pages 537-540, 2013.
- **Z. Ma**, Y. Yang, F. Nie and N. Sebe: “Thinking of Images as What They Are: Compound Matrix Regression for Image Classification". In *International Joint Conferences on Artificial Intelligence*, 2013.
- **Z. Ma**, Y. Yang, Z. Xu, S. Yan, N. Sebe and A. G. Hauptmann: “Complex Event Detection via Multi-source Video Attributes". In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2627-2633, 2013.
- Y. Yang, **Z. Ma**, N. Sebe and A. G. Hauptmann: “Feature Selection for Multimedia Analysis by Sharing Information among Multiple Tasks". *IEEE Transactions on Multimedia*, 15(3):661-669, 2013.
- Y. Yang, J. Song, Z. Huang, **Z. Ma**, N. Sebe and A. G. Hauptmann: “Multi-Feature Fusion via Hierarchical Regression for Multimedia Analysis". *IEEE Transactions on Multimedia*, 15(3):572-581, 2013.
- Y. Han, Z. Xu, **Z. Ma** and Z. Huang: “Image Classification with Manifold Learning for Out-of-sample Data". *Signal Processing*, 93(8):2169-2177, 2013.
- H. Zhang, Y. Liu, **Z. Ma**: “Fusing Inherent and External Knowledge with Nonlinear Learning for Cross-media Retrieval". *Neurocomputing*, 119:10-16, 2013.
- Y. Yang, Y. Yang, Z. Huang, J. Liu and **Z. Ma**: “Robust Cross-Media Transfer for Visual Event Detection". In *Proceedings of the ACM International Conference on Multimedia*, pages 1045-1048, 2012.
- S. Wang, Y. Yang, **Z. Ma**, X. Li, C. Pang and A. G. Hauptmann: “Action Recognition by Exploring the Data Distribution and Feature Correlation". In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1370-1377, 2012.
- Y. Yang, H. T. Shen, **Z. Ma**, Z. Huang and X. Zhou: “L21-Norm Regularized Discriminative Feature Selection for Unsupervised Learning". In *International Joint Conferences on Artificial Intelligence*, pages 1589-1594, 2011.

1

INTRODUCTION

How do we memorize moments of our life? We take pictures to capture the beauty of nature, happy smiles of our beloved people or the prosperity of the cities that we build. We record videos to memorize our daily life in a more vivid way. Be it a birthday party or a wedding ceremony. Videos can perfectly capture each joyful and romantic moment. We also like to share those digitalized memories with our family and friends even when we are not close by since internet makes us always connected. Online portals like Flickr and Youtube and social network like Facebook and Instagram are flourishing all the time. These comprise just a part of the multimedia data nowadays, not to mention the tremendous number of other news, documentary and surveillance resources. The abundant images and videos serve as a huge information pool that can be utilized for our daily life. For the exploitation of them per se, effective indexing techniques are highly desired [19].

Images and videos depict semantic contents in different degrees of richness. Generally speaking, people tend to record static concepts such as objects, scenes or moments of human activities. Videos, in contrast, are used to record dynamic events that are more complicated than static concepts. For example, we can capture a flower with an image but a wedding ceremony needs a long lasting video. Therefore, images and videos consist of the main multimedia data and it is important to develop effective analyzing techniques for both of them. In this thesis we address the problem of image and video understanding and specifically, we tackle the problem with machine learning techniques. The generic framework of thesis is displayed in Figure 1 which shows that the primary techniques harnessed in our work are comprised of feature selection, semi-supervised learning, intermediate representation learning and knowledge adaptation.

We begin from image analysis as static concepts are the components of complicated video events. So what is the basis of image analysis? It is probably the feature representation of an image. In the literature, many different types of feature have been proposed to capture the semantic information of images. Impressive progress on image analysis has been witnessed based on these feature representations. However, it is inevitable that the feature representation has certain amount of noise and redundancy. Consequently, the following questions are raised up:

- Is it possible to get a more compact representation? Would the analyzing accuracy be improved as a result?

We work on these issues in Chapter 2, which was published by IEEE Transactions on Multimedia [52]. Technically, feature selection is utilized to select a compact subset from the original feature sets and a novel sparse model formulates our algorithm, which corresponds to the Module 1 in Figure 1. Another benefit of feature selection is that the dimension of the feature representation is reduced, thus leading to the improvement of computational efficiency.

To step further, we include videos that contain simple activities and human actions into our work besides images. Is there any common problem existing in the research of both of them? We pose this question as it would really beneficial if we can come up with a solution for such a common problem as it can be applied to both domains. Given a closer look, we notice that both image and video understanding face a reality that precisely labeled images and videos are difficult to obtain.

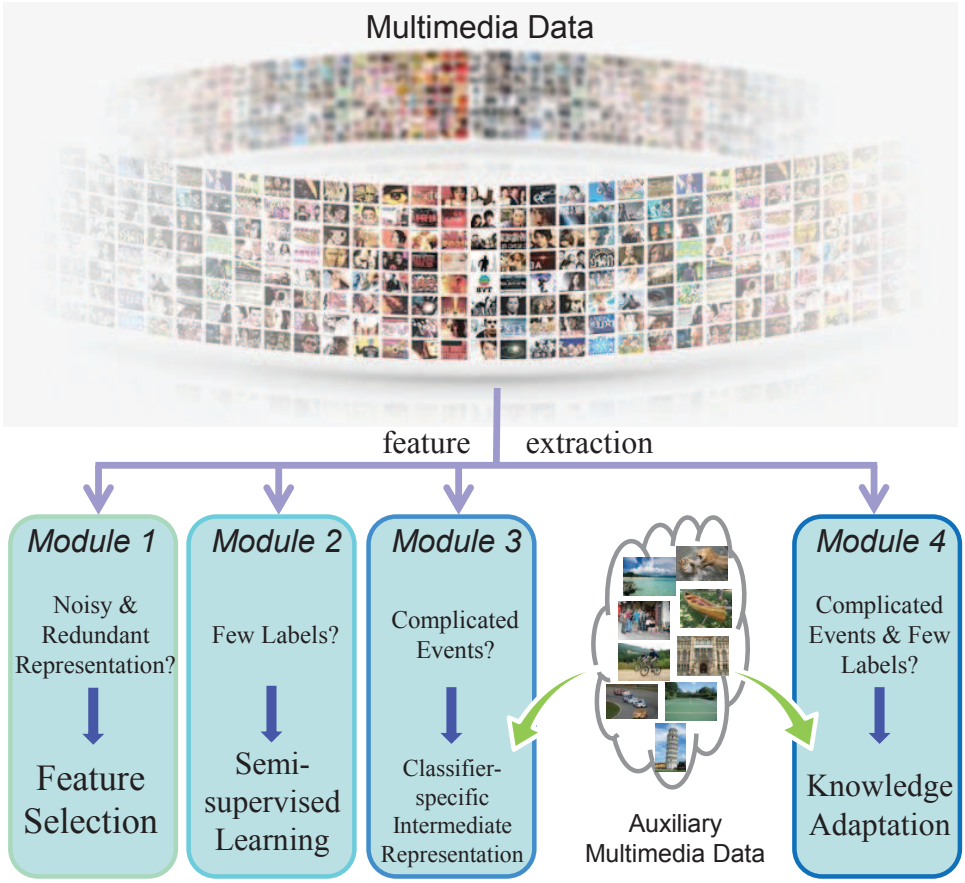


Figure 1: The illustration of our approach for multimedia content analysis.

Though images and videos on the Web are usually associated with tags (labels), they are subjective and sometimes noisy. As for image and video understanding we need to learn models with labeled training data, noisy and incorrect labels would potentially lead to incompetent analyzing models. Hence, the following question comes up:

- Is there any way to attain reasonable analyzing performance with only few labeled images and videos are available?

Searching for a possible answer, we propose a semi-supervised feature analyzing framework for image and video understanding in Chapter 3, which was published by IEEE Transactions on Multimedia [53]. This work corresponds to the Module 2 in Figure 1. Semi-supervised learning is known to be able to handling the paucity of precise labels by exploiting both labeled and unlabeled data. Our approach is based on semi-supervised learning and simultaneously considers eliminating feature noise and redundancy. Through extensive experiments on image and video classification, we validate that properly utilizing unlabeled data does contribute to the performance boost.

Following the work on videos with simple activities and human actions in Chapter 3, we move on to understanding more complicated videos that depict a multimedia event such as *landing a fish*. A multimedia event is a higher level semantic abstraction of video sequences than a concept and consists of multiple concepts. In addition, a multimedia event usually lasts much longer than a concept that can be detected in a shorter video sequence or even in a single frame. Another

challenge is that different video sequences of a particular event may have huge variations. Despite its arduousness, we propose to work on multimedia event analysis as it is more closely related to user interest. We base our research on the multimedia event detection task that has been drawing increasing attention recently. Detection task is more challenging than the widely studied annotation task. Multimedia annotation, also known as recognition, aims to associate a datum with one or multiple semantic labels (tags). Detection identifies the occurrence of a class of interest in a large pool of data. In contrast with annotation for which both the training and testing data are from a fixed number of classes, the training and testing data in detection can be from an infinite number of classes. Hence, multimedia event detection has posed a great research challenge. As multimedia event builds upon several basic elements of objects, scenes and human actions we may refer to an approach suggested by previous work that uses semantic concept representation obtained from concept detectors for event videos [28] [25]. Yet this approach requires the training of many concept detectors in advance, which is tedious and the video understanding performance heavily depends on the accuracy of those concept detectors. As a result, we think about the following question:

- Can we skip the explicit concept detection process but learn an intermediate representation using available multimedia archives related to various concepts for complicated events?

Probing for a positive answer, we propose to learning an intermediate representation coupled with the classifier learning for multimedia event detection in Chapter 4, which was published by IEEE Transactions on Multimedia [56]. Our method corresponds to the Module 3 in Figure 1. Since the intermediate representation learning is bounded to the classifier learning, both of them attain mutual benefit, thus resulting in an optimized event detector that carries more informative cues from the intermediate representation.

We have witnessed encouraging results in Chapter 4 by leveraging the idea that a multimedia event consists of several relevant concepts of objects, scenes and actions. The progress motivates us to further investigate improving multimedia event detection in this direction. Particularly, we tackle a similar problem in line with the second problem addressed in this thesis:

- How can we guarantee reasonable multimedia event detection accuracy when only few positive exemplars are provided?

Note that we expect a solution tailored for multimedia event detection and Chapter 4 has shed a light upon us that other concepts-based multimedia data can be useful. Hence, rather than semi-supervised learning, we approach the problem by a novel knowledge adaptation algorithm in Chapter 5, the extension of our ACM MM paper [54]. We propose to adapt the knowledge from concept level to assist in event detection. Specifically, we use the available video corpora with annotated concepts as our auxiliary resource and event detection is performed on the target videos. Our approach has another desirable property that it is able to adapt knowledge from the source to the target even if the features of them are partially different, but overlapping. Avoiding the requirement that the two domains are consistent in feature types is desirable as data collection platforms change or augment their capabilities and we should be able to respond to this with little or no effort.

The final result of this thesis delivers a comprehension of how we can improve multimedia analysis through a variety of machine learning techniques. From the representation perspective, feature selection is potentially helpful. From the classification perspective, semi-supervised learning and transfer learning both bring in reasonable performance by using only few labeled training data.

2

WEB IMAGE ANNOTATION VIA SUBSPACE-SPARSITY COLLABORATED FEATURE SELECTION[1]

The number of web images has been explosively growing due to the development of network and storage technology. These images make up a large amount of current multimedia data and are closely related to our daily life. To efficiently browse, retrieve and organize the web images, numerous approaches have been proposed. Since the semantic concepts of the images can be indicated by label information, automatic image annotation becomes one effective technique for image management tasks. Most existing annotation methods use image features that are often noisy and redundant. Hence, feature selection can be exploited for a more precise and compact representation of the images, thus improving the annotation performance. In this chapter, we propose a novel feature selection method and apply it to automatic image annotation. There are two appealing properties of our method. First, it can jointly select the most relevant features from all the data points by using a sparsity-based model. Second, it can uncover the shared subspace of original features, which is beneficial for multi-label learning. To solve the objective function of our method, we propose an efficient iterative algorithm. Extensive experiments are performed on large image databases that are collected from the web. The experimental results together with the theoretical analysis have validated the effectiveness of our method for feature selection, thus demonstrating its feasibility of being applied to web image annotation.

2.1 INTRODUCTION

As digital cameras become very common gadgets in our daily life, we have witnessed an explosive growth of digital images. On the other hand, the popularity of many social networks such as Facebook and Flickr helps boost the sharing of these personal images on the web. In fact, digital images now take up a very large proportion of multimedia contents in the network and are utilized intensively with different purposes. However, it is not straightforward to effectively organize and access these web images because we are facing an overwhelmingly large amount of them. Aiming to manage the images efficiently, automatic image annotation has been proposed as an important technique in multimedia analysis. The key idea for image annotation is to correlate keywords or detailed text descriptions with images to facilitate image indexing, retrieval, organization and management.

The sheer amount of web images itself provides us free and rich image repository for research. Researchers have been developing many automatic image annotation methods by leveraging the web scale databases such as Flickr which consist of a large number of user-generated images annotated with user-defined tags [80]. Appearance-based annotation, which is one popular approach, is generally realized through two processes, namely searching and mining. Similar images of the unannotated images are first found out from the web scale databases through the searching process and then the mining process extracts annotation from the textual information of these retrieved similar images. Research work using this approach has demonstrated promising performance for automatic

1 Z. MA, F. NIE, Y. YANG, J. UIJLINGS AND N. SEBE: "WEB IMAGE ANNOTATION VIA SUBSPACE-SPARSITY COLLABORATED FEATURE SELECTION". *IEEE TRANSACTIONS ON MULTIMEDIA*, 14(4): 1021-1030, 2012.

image annotation [86] [69]. Appearance-based image annotation has its effectiveness, but a major problem is that it can be negatively affected when user-generated tags do not reflect the concepts precisely. Learning-based automatic annotation is another effective approach and has gained much research interest. This approach is dependent on certain amount of available annotated images as the training data to learn classifiers for image annotation. Many algorithms have been rendered using learning-based approach these years with varying degrees of success for multimedia semantic analysis [48] [98] [111] [55] [94]. Therefore, this chapter focuses on exploiting learning based methods for image annotation.

Images are normally represented by multiple features, which can be quite different from each other [99]. As it is inevitable to bring in irrelevant and/or redundant information in the feature representation, feature selection can be used to preprocess the data to facilitate subsequent image annotation task [89]. Hence, it is of great value to propose effective feature selection methods. Existing feature selection algorithms are achieved by different means. For instance, classical feature selection algorithms such as Fisher Score [22] compute the weights of different features, rank them accordingly and then select features one by one. These classical algorithms generally evaluate the importance of each feature individually and neglect the useful information of the correlation between different features. To overcome the disadvantage of selecting features individually, researchers have proposed another approach which selects features jointly across all data points by taking into account the relationship of different features [89] [62]. These methods have shown promising performance in different applications. In this chapter we propose a feature selection technique which builds upon the latest mathematical advances in sparse, joint feature selection and apply this to automatic image annotation.

Image annotation is basically a classification problem. However, most web images are multi-labeled, that is to say, an image can reflect several semantic concepts. This intrinsic characteristic of web images makes it a complicated problem to classify them. A simple way to annotate multi-label images is to transform the problem to a couple of binary classification problems for each concept respectively. Though it is easy to implement, this approach neglects the correlation between different concept labels which is potentially useful. Therefore, many recent works [32] have proposed to exploit the shared subspace learning for multi-label tasks by incorporating the relational information of concept labels into multi-label learning. Inspired by their success, we apply shared subspace learning to the problem of feature selection.

To summarize, we combine the latest advances in joint, sparse feature selection with multi-label learning to create a novel feature selection technique which uncovers a feature subspace that is shared among classes. We name our method Sub-Feature Uncovering with Sparsity and demonstrate its effectiveness for automatic web image annotation. The main contributions of our work are:

- Our method leverages the prominent joint feature selection with sparsity, which can select the most discriminative features by exploiting the whole feature space.
- Our method considers the correlation between different concept labels to facilitate the feature selection.
- We conduct several experiments on large scale databases collected from the web. The results demonstrate the effectiveness of utilizing sparse feature selection and label correlation simultaneously.

The rest of this chapter is organized as follows. We briefly introduce the state of the art on shared feature subspace uncovering, feature selection and automatic image annotation in section II. Then we elaborate the formulation of our method followed by the proposed solution in section III. We conduct extensive experiments in section IV to verify the advantage of our method for web image annotation. The conclusion is drawn in section V.

2.2 RELATED WORK

Our work is geared towards better image annotation performance by exploiting effective feature selection. In this section, we briefly review the three related topics of our work, *i.e.*, shared feature subspace uncovering, feature selection and automatic image annotation.

2.2.1 *Shared Feature Subspace Uncovering*

Let x be a datum represented by a feature vector. The general goal of supervised learning is to predict for the input x an output y. To achieve this objective, learning algorithms usually use training data $\{(x_i, y_i)\}_{i=1}^{n}$ to learn a prediction function f that can correlate x with y. A common approach to obtain f is to minimize the following regularized empirical error:

$$\min_{f} \sum_{i=1}^{n} \mathrm{loss}(f(x_i), y_i) + \mu\Omega(f), \tag{2.1}$$

where $\mathrm{loss}(\cdot)$ is the loss function and $\mu\Omega(f)$ is the regularization with μ as its parameter.

It is reasonable to assume that multi-label images share certain common attributes. For example, a picture related to “parade", “people" and “street" share the component “people" with another picture related to “party", “people." Intuitively, we can leverage such label correlations for image annotation. In multi-label learning problems, Ando *et al.* assume that there is a shared subspace for the original feature space [7]. The concepts of an image are predicted by its vector representation in the original feature space together with the embedding in the shared subspace, which can be generalized as the following demonstration:

$$f(x) = v^T x + p^T Q^T x, \tag{2.2}$$

where v and p are the weight vectors and Q is a common subspace shared by all the features.

Suppose the images are related to c concepts in multi-label learning and there are m_t training data $\{x_i\}_{i=1}^{m_t}$ belonging to the t-th concept labeled as $\{y_i\}_{i=1}^{m_t}$. Then (2.1) can be redefined as:

$$\begin{aligned} \min_{f_t, Q} \sum_{t=1}^{c} \left(\frac{1}{m_t} \sum_{i=1}^{m_t} \mathrm{loss}\left(f_t(x_i), y_i\right) + \mu\Omega(f_t) \right) \\ \text{s.t.} \quad Q^T Q = I \end{aligned} \tag{2.3}$$

Note that the constraint $Q^T Q = I$ in (2.3) is imposed to make the problem tractable.

By incorporating the shared feature subspace uncovering of (2.2) into (2.3), we get:

$$\begin{aligned} \min_{\{v_t, p_t\}, Q} \sum_{t=1}^{c} \left(\frac{1}{m_t} \sum_{i=1}^{m_t} \mathrm{loss}\left((v_t + Q p_t)^T x_i, y_i \right) + \mu\Omega\left(\{v_t, p_t\}\right) \right) \\ \text{s.t.} \quad Q^T Q = I \end{aligned} \tag{2.4}$$

Shared feature subspace learning has received increasing attention for its effectiveness on multi-label data [32]. Its theory has also been applied in multimedia analysis and proved its advantage. For instance, Amores *et al.* have leveraged the idea of sharing feature across multiple classes for object-class recognition and achieved prominent performance [6]. As a result, we adopt shared feature subspace uncovering in our feature selection framework and build our mathematical formulation on (2.4).

2.2.2 *Feature Selection*

Feature selection is widely adopted in many multimedia analysis applications. Its principle is to select the most discriminating features from the original ones while simultaneously eliminate the noise, thus resulting in better performance in practice. Another advantage of feature selection lies in its attribute that it reduces the dimensionality of the original data, which in turn reduces the computational cost of the classification.

According to the availability of label information, feature selection algorithms can be classified into two groups: supervised and unsupervised. Unsupervised feature selection [40] [87] [12] is used when there is no label information. An effective way of unsupervised feature selection is to use the manifold structure of the whole feature set to select the most meaningful features [12].

In contrast, supervised feature selection is preferable when there is available label information that can be leveraged by using the correlation between features and labels. In the literature, plenty of supervised feature selection methods have been proposed. For example, Fisher Score [22] and ReliefF [36] are traditional supervised feature selection methods and are exploited widely in multimedia analysis. However, traditional feature selection usually neglects the correlation among different features [12]. Therefore, another approach has been developed recently, namely sparsity-based feature selection [37] [62] which can exploit the feature correlation. This approach is built upon the comprehension that many real world data can be sparsely represented, thus rendering the possibility of searching the sparse representation of the data to realize feature selection. The $l_{2,1}$-norm regularization is known to be an effective model for sparse feature selection [107] and has drawn increasing attention [62] [95].

The $l_{2,1}$-norm of an arbitrary matrix $W \in \mathbb{R}^{d\times c}$ is defined as:

$$\|W\|_{2,1} = \sum_{i=1}^{d} \sqrt{\sum_{j=1}^{c} W_{ij}^2} \tag{2.5}$$

In [62] and [95], $l_{2,1}$-norm is leveraged to conduct feature selection jointly across the entire feature space with promising performance. Their works demonstrate that the $l_{2,1}$-norm of W makes W sparse, meaning that some of its rows shrink to zero. Consequently, W can be viewed as the combination coefficients for the most discriminative features. Feature selection is then realized by W where only the features associated with the non-zero rows in W are selected. Sparsity-based feature selection is efficient as it can select discriminative features jointly across all data points. However, few works have incorporated sparsity-based feature selection and shared feature subspace uncovering into one joint framework.

2.2.3 *Automatic Image Annotation*

Image annotation can be viewed as a classification task. It aims to correlate concept labels with specific images by classifying images to different classes. The ultimate goal is that the predicted labels via annotation algorithms can precisely reflect the real semantic contents of images. Nonetheless, the web image resources are countless so it is infeasible to annotate all of them manually. Hence, automatic image annotation becomes an essential tool for handling web scale images for retrieval, index and other management tasks.

Existing automatic image annotation methods have utilized a plethora of techniques [80] [69] [48] [24] [14]. Since images are usually represented by different features, much work [24] [89] [55] has focused on optimizing the feature selection process in their annotation frameworks. By finding the discriminative subset of original features and eliminating the noise, feature selection can help improve image annotation performance. For instance, Ma *et al.* have exploited a sparse

selection model to select discriminative features that are closely related to image concepts for image annotation [55].

Thanks to the continuous effort made by researchers, we have witnessed great advance in automatic annotation for web images. However, the performance of automatic image annotation is yet to be satisfactory, thus requiring more research work in this domain. Inspired by the recent advanced techniques of feature selection and shared feature subspace uncovering, we propose a novel framework to extract the most discriminating features to boost the image annotation performance.

2.3 THE PROPOSED FRAMEWORK

In this section, we first illustrate the formulation of our Sub-Feature Uncovering with Sparsity (S-FUS) framework. Then a detailed approach is rendered to solve the objective problem.

2.3.1 *Problem Formulation*

Our method roots from the shared feature subspace uncovering as given by (2.4).

Denote the training data matrix as $X = [x_1, x_2, ..., x_n]$ where $x_i \in \mathbb{R}^d (1 \leqslant i \leqslant n)$ is the i-th datum and n is the total number of the training data. Let $Y = [y_1, y_2, ..., y_n]^T \in \{0,1\}^{n \times c}$ be the label matrix. c stands for the class number and $y_i \in \mathbb{R}^c (1 \leqslant i \leqslant n)$ is the label vector with c classes. Denote $V = [v_1, v_2, ..., v_c] \in \mathbb{R}^{d \times c}$ and $P = [p_1, p_2, ..., p_c] \in \mathbb{R}^{sd \times c}$ where sd is the dimension of the shared subspace. We can then present (2.4) in a more compact way as:

$$\begin{aligned} \min_{V,P,Q} \ & loss\left((V + QP)^T X, Y\right) + \mu\Omega(V, P) \\ & \text{s.t.} \quad Q^T Q = I \end{aligned} \tag{2.6}$$

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

$$\begin{aligned} \min_{W,V,P,Q} \ & loss\left(W^T X, Y\right) + \mu\Omega(V, P) \\ & \text{s.t.} \quad Q^T Q = I \end{aligned} \tag{2.7}$$

It can be seen from the above function that by applying a different loss function and regularization, we can realize shared feature subspace uncovering in different ways. The least square loss has been widely used in research which can be illustrated as $\left\|X^T W - Y\right\|_F^2$ where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix. By utilizing the least square loss, Ji *et al.* [32] have proposed to achieve shared subspace learning in the following way:

$$\begin{aligned} \min_{W,P,Q} \ & \left\|X^T W - Y\right\|_F^2 + \alpha \|W\|_F^2 + \beta \|W - QP\|_F^2 \\ & \text{s.t.} \quad Q^T Q = I \end{aligned} \tag{2.8}$$

In the above function, $\alpha \|W\|_F^2 + \beta \|W - QP\|_F^2$ is the regularization term. The first part regulates the information to each specific label and the second part controls the complexity of the objective function. This approach is mathematically tractable and can be easily implemented. However, there are two issues worthy of further consideration. First, the least square loss is very sensitive to outliers, thus demanding a more robust loss function. Second, as we aim to conduct effective feature selection, it is advantageous to exert the sparse feature selection models on the regularization term. In [62], Nie *et al.* have proved that $l_{2,1}$-norm based models can handle both the aforementioned issues.

We therefore propose the following objective function as our foundation to realize feature selection:

$$\begin{aligned} \arg\min_{W,P,Q} \left\|X^T W - Y\right\|_{2,1} + \alpha \|W\|_{2,1} + \beta \|W - QP\|_F^2 \\ \text{s.t.} \quad Q^T Q = I \end{aligned} \tag{2.9}$$

The loss function in our objective, that is to say, $\left\|X^T W - Y\right\|_{2,1}$ is robust to outliers as indicated in [62]. At the same time, $\|W\|_{2,1}$ in the regularization term guarantees that W is sparse to achieve feature selection across all data points [95] [62].

2.3.2 *Solution*

As can be seen in (2.9), our problem involves the $l_{2,1}$-norm which is non-smooth and cannot be solved in a closed form. As a result, we propose to solve it as follows.

By denoting $X^T W - Y = [z^1, ..., z^n]^T$ and $W = [w^1, ..., w^d]^T$, the objective in (2.9) is equivalent to:

$$\begin{aligned} \arg\min_{W,P,Q} Tr\left((X^T W - Y)^T \tilde{D}(X^T W - Y)\right) + \alpha Tr\left(W^T D W\right) + \beta \|W - QP\|_F^2 \\ \text{s.t.} \quad Q^T Q = I, \end{aligned} \tag{2.10}$$

where $\tilde{D}$ and D are two matrices with their diagonal elements $\tilde{D}_{ii} = \frac{1}{2\|z^i\|_2}$ and $D_{ii} = \frac{1}{2\|w^i\|_2}$ respectively.

Note that for any arbitrary matrix A, $\|A\|_F^2 = Tr\left(A^T A\right)$. Thus, (2.10) becomes:

$$\begin{aligned} &\arg\min_{W,P,Q} Tr\left((X^T W - Y)^T \tilde{D}(X^T W - Y)\right) + \alpha Tr\left(W^T D W\right) \\ &+\beta Tr\left((W - QP)^T (W - QP)\right) \\ &\text{s.t.} \quad Q^T Q = I, \end{aligned} \tag{2.11}$$

By setting the derivative of (2.11) *w.r.t* P to zero, we have:

$$\beta(2Q^T QP - 2Q^T W) = 0 \qquad \Rightarrow \qquad P = Q^T W \tag{2.12}$$

Substituting P in (2.11) with (2.12) we have:

$$\begin{aligned} &\arg\min_{W,Q} Tr\left((X^T W - Y)^T \tilde{D}(X^T W - Y)\right) + \alpha Tr\left(W^T D W\right) \\ &+\beta Tr\left((W - QQ^T W)^T (W - QQ^T W)\right) \\ &\Rightarrow \arg\min_{W,Q} Tr\left((X^T W - Y)^T \tilde{D}(X^T W - Y)\right) + \alpha Tr\left(W^T D W\right) \\ &+\beta Tr\left(W^T (I - QQ^T)(I - QQ^T) W\right) \\ &\text{s.t.} \quad Q^T Q = I \end{aligned} \tag{2.13}$$

Since $(I - QQ^T)(I - QQ^T) = (I - QQ^T)$, the problem becomes:

$$\begin{aligned} \arg\min_{W,Q} Tr\left((X^T W - Y)^T \tilde{D}(X^T W - Y)\right) + Tr\left(W^T(\alpha D + \beta I - \beta QQ^T) W\right) \\ \text{s.t.} \quad Q^T Q = I \end{aligned} \tag{2.14}$$

By setting the derivative of (2.14) *w.r.t* W to zero, we get:

$$\begin{aligned} & 2X\tilde{D}X^T W - 2X\tilde{D}Y + 2(\alpha D + \beta I - \beta QQ^T)W = 0 \\ \Rightarrow \quad & (X\tilde{D}X^T + \alpha D + \beta I - \beta QQ^T)W = X\tilde{D}Y \\ \Rightarrow \quad & W = (M - \beta QQ^T)^{-1} X\tilde{D}Y \\ \Rightarrow \quad & W = N^{-1} X\tilde{D}Y, \end{aligned} \tag{2.15}$$

where $M = X\tilde{D}X^T + \alpha D + \beta I$, $N = (M - \beta QQ^T)^{-1}$ and $N = N^T$.

Note that (2.14) can be rewritten as:

$$\begin{aligned} & \arg\min_{W,Q} Tr\left(W^T X\tilde{D}X^T W\right) - 2Tr\left(W^T X\tilde{D}Y\right) + Tr\left(Y^T \tilde{D}Y\right) \\ & + Tr\left(W^T(\alpha D + \beta I - \beta QQ^T)W\right) \\ & \Rightarrow \arg\min_{W,Q} Tr\left(W^T(X\tilde{D}X^T + \alpha D + \beta I - \beta QQ^T)W\right) - 2Tr\left(W^T X\tilde{D}Y\right) \\ & + Tr\left(Y^T \tilde{D}Y\right) \\ & \Rightarrow \arg\min_{W,Q} Tr\left(W^T(M - \beta QQ^T)W\right) - 2Tr\left(W^T X\tilde{D}Y\right) + Tr\left(Y^T \tilde{D}Y\right) \\ & \Rightarrow \arg\min_{W,Q} Tr\left(W^T NW\right) - 2Tr\left(W^T X\tilde{D}Y\right) + Tr\left(Y^T \tilde{D}Y\right) \\ & \qquad \text{s.t.} \quad Q^T Q = I \end{aligned} \tag{2.16}$$

By incorporating the W obtained with (2.15) into the above function, we have:

$$\begin{aligned} & \arg\min_{Q} Tr\left(Y^T \tilde{D}X^T N^{-1} NN^{-1} X\tilde{D}Y\right) - 2Tr\left(Y^T \tilde{D}X^T N^{-1} X\tilde{D}Y\right) + Tr\left(Y^T \tilde{D}Y\right) \\ & \Rightarrow \arg\min_{Q} Tr\left(Y^T \tilde{D}Y\right) - Tr\left(Y^T \tilde{D}X^T N^{-1} X\tilde{D}Y\right) \\ & \qquad \text{s.t.} \quad Q^T Q = I \end{aligned} \tag{2.17}$$

The above problem is equivalent to the following one:

$$\begin{aligned} & \arg\max_{Q} Tr\left(Y^T \tilde{D}X^T N^{-1} X\tilde{D}Y\right) \\ & \qquad \text{s.t.} \quad Q^T Q = I \end{aligned} \tag{2.18}$$

According to Sherman-Woodbury-Morrison formula, $N^{-1} = (M - \beta QQ^T)^{-1} = M^{-1} + \beta M^{-1} Q(I - \beta Q^T M^{-1} Q)^{-1}$
$Q^T M^{-1}$. Thus, (2.18) becomes:

$$\begin{aligned} & \arg\max_{Q} Tr\left(Y^T \tilde{D}X^T M^{-1} X\tilde{D}Y + \beta Y^T \tilde{D}X^T Q(I - \beta Q^T M^{-1} Q)^{-1} Q^T M^{-1} X\tilde{D}Y\right) \\ & \qquad \text{s.t.} \quad Q^T Q = I \end{aligned} \tag{2.19}$$

which is equivalent to:

$$\begin{aligned} & \arg\max_{Q} Tr\left(Y^T \tilde{D}X^T M^{-1} Q(I - \beta Q^T M^{-1} Q)^{-1} Q^T M^{-1} X\tilde{D}Y\right) \\ & \Rightarrow \arg\max_{Q} Tr\left(Y^T \tilde{D}X^T M^{-1} Q(Q^T Q - \beta Q^T M^{-1} Q)^{-1} Q^T M^{-1} X\tilde{D}Y\right) \\ & \Rightarrow \arg\max_{Q} Tr\left(Y^T \tilde{D}X^T M^{-1} Q[Q^T(I - \beta M^{-1})Q]^{-1} Q^T M^{-1} X\tilde{D}Y\right) \\ & \qquad \text{s.t.} \quad Q^T Q = I \end{aligned} \tag{2.20}$$

Algorithm 1: The algorithm for solving the SFUS objective function.

Input:

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

Output:

Optimized $W \in \mathbb{R}^{d\times c}$.

1: Set $t = 0$ and initialize $W_0 \in \mathbb{R}^{d\times c}$ randomly;

2: **repeat**

Compute $[z_t^1, ..., z_t^n]^T = X^T W_t - Y$;

Compute the diagonal matrix $\tilde{D}_t$ as: $\tilde{D}_t = \begin{bmatrix} \frac{1}{2\|z_t^1\|_2} & & \\ & ... & \\ & & \frac{1}{2\|z_t^n\|_2} \end{bmatrix}$;

Compute the diagonal matrix D_t as: $D_t = \begin{bmatrix} \frac{1}{2\|w_t^1\|_2} & & \\ & ... & \\ & & \frac{1}{2\|w_t^d\|_2} \end{bmatrix}$;

Compute $M_t = X\tilde{D}_t X^T + \alpha D_t + \beta I$;

Compute $A_t = I - \beta M_t^{-1}$;

Compute $B_t = M_t^{-1} X\tilde{D}_t YY^T \tilde{D}_t X^T M_t^{-1}$;

Obtain Q_t by the eigen-decomposition of $A_t^{-1}B_t$;

Update W_{t+1} according to (2.15);

$t = t + 1$.

until *Convergence*;

3: Return W.

As for any arbitrary matrices A, B and C, $Tr(ABC) = Tr(BCA)$, the above function becomes:

$$\begin{aligned} &\arg\max_Q Tr\left([Q^T(I-\beta M^{-1})Q]^{-1} Q^T M^{-1} X\tilde{D}YY^T\tilde{D}X^T M^{-1} Q\right) \\ \Rightarrow &\arg\max_Q Tr\left((Q^T A Q)^{-1} Q^T B Q\right) \\ &\text{s.t.} \quad Q^T Q = I, \end{aligned} \tag{2.21}$$

where $A = I - \beta M^{-1}$ and $B = M^{-1} X\tilde{D}YY^T\tilde{D}X^T M^{-1}$.

Equation (2.21) can be easily solved by the eigen-decomposition of $A^{-1}B$. However, as the solving of Q requires the input of $\tilde{D}$ and D which are related to W, it is still not straightforward to get Q and W. To solve this problem, we propose an iterative approach demonstrated in Algorithm 1. The complexity of the proposed algorithm is briefly discussed as follows. The complexity of calculating the inverse of a few matrices is $\mathcal{O}(d^3)$. To obtain Q, we need to conduct eigen-decomposition of $A^{-1}B$, which is also $\mathcal{O}(d^3)$ in complexity.

The proposed iterative approach in Algorithm 1 can be verified to converge to the optimal W by the following theorem.

Theorem 1 *The objective function value shown in (2.9) monotonically decreases in each iteration until convergence using the iterative approach in Algorithm 1.*

Proof. According to Algorithm 1, it can be inferred from (2.11) that:

$$\begin{aligned} W_{t+1} = \arg\min &Tr\left((X^T W - Y)^T \tilde{D}(X^T W - Y)\right) + \alpha Tr\left(W^T D W\right) + \beta \|W - QP\|_F^2 \\ &\text{s.t.} \quad Q^T Q = I \end{aligned}$$

Therefore, we have

$$
\begin{aligned}
&\mathrm{Tr}\left((X^{T}W_{t+1}-Y)^{T}\tilde{D}_{t}(X^{T}W_{t+1}-Y)\right)+\alpha\mathrm{Tr}\left(W_{t+1}^{T}D_{t}W_{t+1}\right)\\
&+\beta\left\|W_{t+1}-Q_{t+1}P_{t+1}\right\|_{F}^{2}\\
&\leqslant \mathrm{Tr}\left((X^{T}W_{t}-Y)^{T}\tilde{D}_{t}(X^{T}W_{t}-Y)\right)+\alpha\mathrm{Tr}\left(W_{t}^{T}D_{t}W_{t}\right)+\beta\left\|W_{t}-Q_{t}P_{t}\right\|_{F}^{2}\\
&\Rightarrow \sum_{i=1}^{n}\frac{\left\|x_{i}^{T}W_{t+1}-y_{i}\right\|_{2}^{2}}{2\left\|x_{i}^{T}W_{t}-y_{i}\right\|_{2}}+\alpha\sum_{i=1}^{d}\frac{\left\|w_{t+1}^{i}\right\|_{2}^{2}}{2\left\|w_{t}^{i}\right\|_{2}}+\beta\left\|W_{t+1}-Q_{t+1}P_{t+1}\right\|_{F}^{2}\\
&\leqslant \sum_{i=1}^{n}\frac{\left\|x_{i}^{T}W_{t}-y_{i}\right\|_{2}^{2}}{2\left\|x_{i}^{T}W_{t}-y_{i}\right\|_{2}}+\alpha\sum_{i=1}^{d}\frac{\left\|w_{t}^{i}\right\|_{2}^{2}}{2\left\|w_{t}^{i}\right\|_{2}}+\beta\left\|W_{t}-Q_{t}P_{t}\right\|_{F}^{2}\\
&\Rightarrow \sum_{i=1}^{n}\left\|x_{i}^{T}W_{t+1}-y_{i}\right\|_{2}-\sum_{i=1}^{n}\left\|x_{i}^{T}W_{t+1}-y_{i}\right\|_{2}+\sum_{i=1}^{n}\frac{\left\|x_{i}^{T}W_{t+1}-y_{i}\right\|_{2}^{2}}{2\left\|x_{i}^{T}W_{t}-y_{i}\right\|_{2}}\\
&+\alpha\sum_{i=1}^{d}\left\|w_{t+1}^{i}\right\|_{2}-\alpha\sum_{i=1}^{d}\left\|w_{t+1}^{i}\right\|_{2}+\alpha\sum_{i=1}^{d}\frac{\left\|w_{t+1}^{i}\right\|_{2}^{2}}{2\left\|w_{t}^{i}\right\|_{2}}+\beta\left\|W_{t+1}-Q_{t+1}P_{t+1}\right\|_{F}^{2}\\
&\leqslant \sum_{i=1}^{n}\left\|x_{i}^{T}W_{t}-y_{i}\right\|_{2}-\sum_{i=1}^{n}\left\|x_{i}^{T}W_{t}-y_{i}\right\|_{2}+\sum_{i=1}^{n}\frac{\left\|x_{i}^{T}W_{t}-y_{i}\right\|_{2}^{2}}{2\left\|x_{i}^{T}W_{t}-y_{i}\right\|_{2}}\\
&+\alpha\sum_{i=1}^{d}\left\|w_{t}^{i}\right\|_{2}-\alpha\sum_{i=1}^{d}\left\|w_{t}^{i}\right\|_{2}+\alpha\sum_{i=1}^{d}\frac{\left\|w_{t}^{i}\right\|_{2}^{2}}{2\left\|w_{t}^{i}\right\|_{2}}+\beta\left\|W_{t}-Q_{t}P_{t}\right\|_{F}^{2}\\
&\Rightarrow \sum_{i=1}^{n}\left\|x_{i}^{T}W_{t+1}-y_{i}\right\|_{2}+\alpha\sum_{i=1}^{d}\left\|w_{t+1}^{i}\right\|_{2}+\beta\left\|W_{t+1}-Q_{t+1}P_{t+1}\right\|_{F}^{2}\\
&-(\sum_{i=1}^{n}\left\|x_{i}^{T}W_{t+1}-y_{i}\right\|_{2}-\sum_{i=1}^{n}\frac{\left\|x_{i}^{T}W_{t+1}-y_{i}\right\|_{2}^{2}}{2\left\|x_{i}^{T}W_{t}-y_{i}\right\|_{2}})-\alpha(\sum_{i=1}^{d}\left\|w_{t+1}^{i}\right\|_{2}-\sum_{i=1}^{d}\frac{\left\|w_{t+1}^{i}\right\|_{2}^{2}}{2\left\|w_{t}^{i}\right\|_{2}})\\
&\leqslant \sum_{i=1}^{n}\left\|x_{i}^{T}W_{t}-y_{i}\right\|_{2}+\alpha\sum_{i=1}^{d}\left\|w_{t}^{i}\right\|_{2}+\beta\left\|W_{t}-Q_{t}P_{t}\right\|_{F}^{2}\\
&-(\sum_{i=1}^{n}\left\|x_{i}^{T}W_{t}-y_{i}\right\|_{2}-\sum_{i=1}^{n}\frac{\left\|x_{i}^{T}W_{t}-y_{i}\right\|_{2}^{2}}{2\left\|x_{i}^{T}W_{t}-y_{i}\right\|_{2}})-\alpha(\sum_{i=1}^{d}\left\|w_{t}^{i}\right\|_{2}-\sum_{i=1}^{d}\frac{\left\|w_{t}^{i}\right\|_{2}^{2}}{2\left\|w_{t}^{i}\right\|_{2}})
\end{aligned}
$$

It has been shown in [62] [95] that for any non-zero vectors $v_{t}^{i}|_{i=1}^{r}$:

$$
\sum_{i}\left\|v_{t+1}^{i}\right\|_{2}-\sum_{i}\frac{\left\|v_{t+1}^{i}\right\|_{2}^{2}}{2\left\|v_{t}^{i}\right\|_{2}}\leqslant\sum_{i}\left\|v_{t}^{i}\right\|_{2}-\sum_{i}\frac{\left\|v_{t}^{i}\right\|_{2}^{2}}{2\left\|v_{t}^{i}\right\|_{2}}
$$

where r is an arbitrary number. Thus, we can easily get the following inequality:

$$
\begin{aligned}
&\sum_{i=1}^{n}\left\|x_{i}^{T}W_{t+1}-y_{i}\right\|_{2}+\alpha\sum_{i=1}^{d}\left\|w_{t+1}^{i}\right\|_{2}+\beta\left\|W_{t+1}-Q_{t+1}P_{t+1}\right\|_{F}^{2}\\
&\leqslant \sum_{i=1}^{n}\left\|x_{i}^{T}W_{t}-y_{i}\right\|_{2}+\alpha\sum_{i=1}^{d}\left\|w_{t}^{i}\right\|_{2}+\beta\left\|W_{t}-Q_{t}P_{t}\right\|_{F}^{2}\\
&\Rightarrow \left\|X^{T}W_{t+1}-Y\right\|_{2,1}+\alpha\left\|W_{t+1}\right\|_{2,1}+\beta\left\|W_{t+1}-Q_{t+1}P_{t+1}\right\|_{F}^{2}\\
&\leqslant \left\|X^{T}W_{t}-Y\right\|_{2,1}+\alpha\left\|W_{t}\right\|_{2,1}+\beta\left\|W_{t}-Q_{t}P_{t}\right\|_{F}^{2}
\end{aligned}
$$

which indicates that the objective function value of (2.9) monotonically decreases until converging to the optimal W through the proposed approach in Algorithm 1□.

2.4 EXPERIMENTS

To validate the efficacy of our method when applied to automatic image annotation, we conduct several experiments particularly on image databases that are collected from the web image resources.

2.4.1 *Compared Methods*

We compare our method with one baseline and several feature selection algorithms on automatic image annotation to understand how our method progresses towards better annotation performance. The compared methods are enumerated as follows.

- Using all features (All-Fea): our baseline. It means that we use the original data without feature selection for annotation.
- Fisher Score (F-score) [22]: a classical method. It selects the most discriminative features by evaluating the importance of each feature individually.
- Sparse Multinomial Logistic Regression via Bayesian L1 Regularisation (SBMLR) [15]: a sparsity based state of the art method. It realizes sparse feature selection by using a Laplace prior.
- Spectral feature selection (SPEC) [106]: a state of the art method using spectral regression. It selects features one by one by leveraging the work of spectral graph theory. The supervised implementation is used in our experiments for fair comparison.
- Group Lasso with Logistic Regression (GLRR) [89]: a recently proposed method based on a sparse model. It utilizes group lasso extended with logistic regression to select both sparse and discriminative groups of homogeneous features.
- Feature Selection via Joint $l_{2,1}$-Norms Minimization (FSNM) [62]: a latest sparse feature selection algorithm. It employs joint $l_{2,1}$-norm minimization on both loss function and regularization for joint feature selection.

As our framework is expanded upon regularized least square regression, we use it as the classifier for all the compared approaches.

2.4.2 *Image Databases*

Web images cover almost all the concepts people are interested in, thus justifying their advantage to be used as research corpus for automatic image annotation. For the sake of the study on multimedia analysis, researchers have also managed to collect and process the web images to create good image databases for experimental purpose.

In our experiments, we select two large scale databases which are both made up of web images. The first one is the MSRA-MM 2.0 database which was created by Microsoft Research Asia [43]. This database was collected from the web through a commercial search engine and consists of 50,000 images belonging to 100 concepts. However, 7,734 images of the original database are not associated with any labels, we thus have removed these images and obtained a subset of 42,266 labeled images. In 2009, the Lab for Media Search in National University of Singapore proposed another large scale image database, *i.e.*, NUS-WIDE where all images are from Flickr [17]. NUS-WIDE includes 269,000 real-world images. The very large amount of NUS-WIDE, from our perspective, can well validate the scalability of our framework for real world annotation tasks. Hence, we choose

Table 1: Performance comparison (±Standard Deviation) on MSRA-MM 2.0 when $10 \times c$ images work as training data.

	MAP	MicroAUC	MacroAUC
All-Fea	0.062±0.001	0.840±0.001	0.655±0.006
F-score [22]	0.060±0.002	0.861±0.005	0.655±0.003
SBMLR [15]	0.056±0.002	0.869±0.003	0.643±0.006
SPEC [106]	0.058±0.001	0.852±0.002	0.650±0.004
FSNM [62]	0.061±0.002	0.875±0.002	0.658±0.006
GLRR [89]	0.060±0.001	0.846±0.001	0.653±0.005
SFUS	**0.063±0.001**	**0.878±0.002**	**0.662±0.005**

this database in our experiments as well. Nonetheless, 59,653 images within NUS-WIDE are unlabeled, we therefore have removed them and used the remaining 209,347 labeled images related to 81 concepts as experimental corpus.

Considering the computational efficiency, we combine three feature types, *i.e.*, Color Correlogram, Edge Direction Histogram and Wavelet Texture provided by the authors to represent the images of the two databases. As a consequence, the corresponding feature dimensions for MSRA-MM 2.0 and NUS-WIDE are 347 and 345 respectively [43] [17].

2.4.3 *Experiment Setup*

The procedure of our experiments can be generalized as follows. We first randomly generate a training set comprised of $m \times c$ images for each database similarly to the experimental setting in [16]. The remaining images are used as testing sets. To understand the performance variation *w.r.t* the number of training data, we set m as 10 and 20 respectively and report the corresponding results. We generate the training and testing sets for 5 times and report the average results for fair comparison with other methods.

Note that our objective function in (2.9) involves two parameters α and β. We tune both of them from $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$ and report the best results. The number of the selected features ranges from $\{100, 150, 200, 250, 300\}$ and we use the corresponding feature subset to represent the images. Then the regularized least square regression is applied as the classifier for image annotation.

To evaluate the annotation performance, we use three evaluation metrics, *i.e.*, Mean Average Precision (MAP), MicroAUC and MacroAUC which are all widely used for multi-label classification tasks [65] [89] [85] [26].

2.4.4 *Performance on Image Annotation*

Table 1 to Table 4 show the annotation results when using $10 \times c$ and $20 \times c$ training data. The results in bold indicate the best performance using the corresponding evaluation metric. According to the annotation results, we observe that our method demonstrates consistently superior performance on both databases.

Take MAP as an example. First, our method is better than All-Fea, *i.e.*, not using feature selection for annotation on both data sets. In particular, SFUS obtains notable improvement over All-Fea on NUS-WIDE. Second, our method has better annotation performance than the compared feature

Table 2: Performance comparison (±Standard Deviation) on NUS-WIDE when $10 \times c$ images work as training data.

	MAP	MicroAUC	MacroAUC
All-Fea	0.081±0.002	0.842±0.003	0.726±0.003
F-score [22]	0.080±0.002	0.851±0.003	0.728±0.004
SBMLR [15]	0.072±0.008	0.871±0.005	0.718±0.028
SPEC [106]	0.078±0.002	0.847±0.003	0.722±0.003
FSNM [62]	0.092±0.001	0.869±0.002	0.753±0.002
GLRR [89]	0.082±0.002	0.853±0.002	0.732±0.003
SFUS	**0.094±0.003**	**0.877±0.002**	**0.756±0.003**

Table 3: Performance comparison (±Standard Deviation) on MSRA-MM 2.0 when $20 \times c$ images work as training data.

	MAP	MicroAUC	MacroAUC
All-Fea	0.067±0.004	0.859±0.011	0.676±0.013
F-score [22]	0.066±0.002	0.876±0.004	0.680±0.004
SBMLR [15]	0.059±0.001	0.883±0.004	0.666±0.004
SPEC [106]	0.066±0.001	0.868±0.001	0.679±0.002
FSNM [62]	0.068±0.001	0.887±0.002	0.687±0.002
GLRR [89]	0.067±0.001	0.866±0.002	0.680±0.002
SFUS	**0.070±0.001**	**0.888±0.002**	**0.690±0.002**

Table 4: Performance comparison (±Standard Deviation) on NUS-WIDE when $20 \times c$ images work as training data.

	MAP	MicroAUC	MacroAUC
All-Fea	0.099±0.001	0.874±0.001	0.767±0.001
F-score [22]	0.098±0.004	0.880±0.005	0.770±0.006
SBMLR [15]	0.073±0.007	0.887±0.006	0.733±0.024
SPEC [106]	0.094±0.001	0.875±0.001	0.763±0.001
FSNM [62]	0.105±0.003	0.888±0.003	0.785±0.004
GLRR [89]	0.105±0.002	0.885±0.003	0.780±0.001
SFUS	**0.108±0.002**	**0.891±0.003**	**0.789±0.003**

Table 5: Performance comparison (±Standard Deviation) using Color Correlogram & Wavelet Texture on MSRA-MM 2.0 when $10 \times c$ training data are labeled.

	MAP	MicroAUC	MacroAUC
All-Fea	0.059±0.001	0.848±0.002	0.652±0.006
F-score [22]	0.059±0.001	0.861±0.006	0.651±0.003
SBMLR [15]	0.053±0.003	0.874±0.004	0.636±0.006
SPEC [106]	0.058±0.001	0.854±0.003	0.648±0.004
FSNM [62]	0.059±0.001	0.872±0.002	0.655±0.005
GLRR [89]	0.060±0.001	0.858±0.002	0.652±0.004
SFUS	**0.061±0.001**	**0.883±0.002**	**0.659±0.005**

selection methods. Using $10 \times c$ training data, SFUS outperforms the second best feature selection method by about 2.6% and it is better than other feature selection algorithms for both data sets; using $20 \times c$ training data, SFUS is better than the second best feature selection method by about 1.6% and 3% on MSRA-MM 2.0 and NUS-WIDE respectively and it demonstrates good advantage over other algorithms. Hence, we conclude that our algorithm is a good feature selection mechanism for web image annotation.

The good performance of SFUS for image annotation can be attributed to the appealing property that it can select features jointly across the whole feature space while simultaneously considering the correlation of multiple labels by exploring the shared feature subspace. The incorporation of the sparse model and shared subspace uncovering facilitates the feature selection by finding the most discriminative features, which can be used subsequently in annotation process.

2.4.5 *Influence of Feature Type*

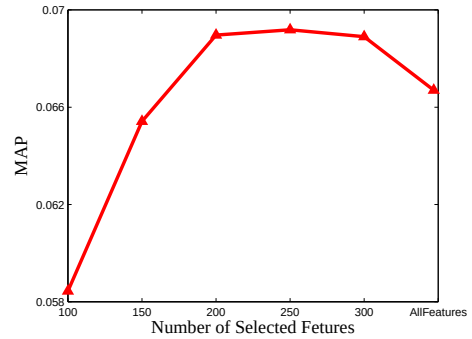
To evaluate the effectiveness of our method, we use a different original feature set, *i.e.*, only Color Correlogram and Wavelet Texture are combined to represent the images and we present the corresponding annotation results. The experiment is conducted on the MSRA-MM dataset with the results shown in Table 5.

It can be seen that our method still outperforms other feature selection algorithms when the images are represented by color histogram and wavelet texture. The results demonstrate that our algorithm is robust for the variance of the original feature set.

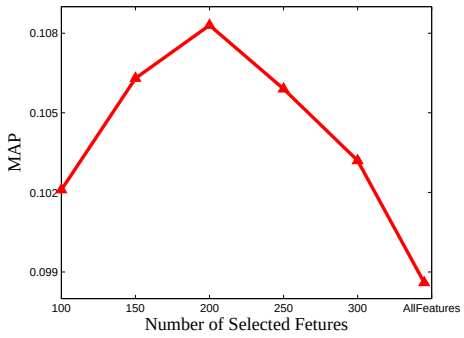
2.4.6 *Influence of Selected Features*

As feature selection is aimed at both accuracy and computational efficiency, we perform an experiment to study how the number of selected features can affect the annotation performance using $20 \times c$ training data. This experiment can present us the general trade-off between performance and computational efficiency for the two image databases.

Figure 2 shows the performance variation *w.r.t* the number of selected features in terms of MAP. We have the following observations: 1) When the number of selected features is too small, MAP is not competitive with using all features for annotation, which could be attributed to too much information loss. For instance, when using less than 150 features of MSRA-MM 2.0, MAP is worse than using all features for annotation. 2) MAP increases as the number of selected features increases up to 200. 3) MAP arrives at the peak level when using 200 features. 4) MAP keeps stable from using



(a) MSRA-MM



(b) NUS-WIDE

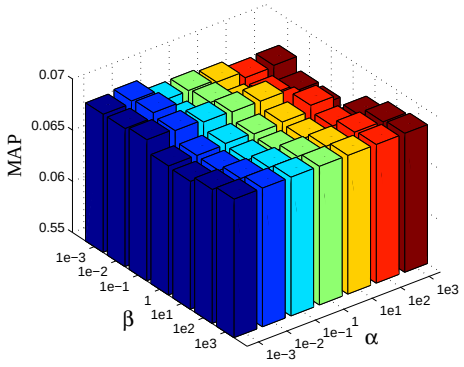
Figure 2: Performance variation *w.r.t* to the number of selected features using our feature selection algorithm.

200 features to using 300 features for MSRA-MM 2.0 while drops for NUS-WIDE. The different variance shown on the two datasets are supposed to be related to the properties of the datasets. 5) After all the features are selected, in other words, without feature selection, MAP is lower than selecting 200 features for MSRA-MM 2.0 and 100 features for NUS-WIDE. We conclude that, as MAP improves on both databases, our method reduces noise.

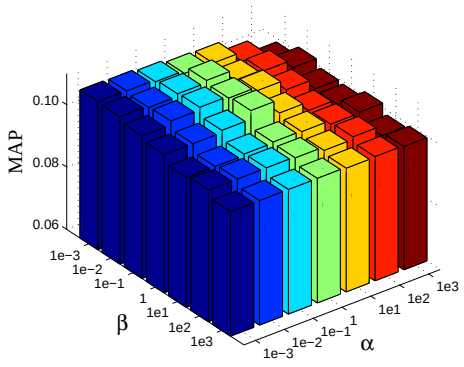
2.4.7 *Parameter Sensitivity Study*

Our method involves two regularization parameters, which are denoted as α and β in (2.9). To learn how they affect the feature selection and consequently the performance on image annotation, we conduct an experiment on the parameter sensitivity. Following the above experiment, we use $20 \times c$ training data for image annotation. MAP is used here to reflect the performance variation.

Figure 3 demonstrates the MAP variation *w.r.t* α and β on the two databases. From Figure 3 we notice that the annotation performance changes corresponding to different combinations of α and β. The impact of different values of the regularization parameters is supposed to be related to the trait of the database. On our experimental datasets, better results are generally obtained when α and β are comparable in value.



(a) MAP-MSRA



(b) MAP-NUS

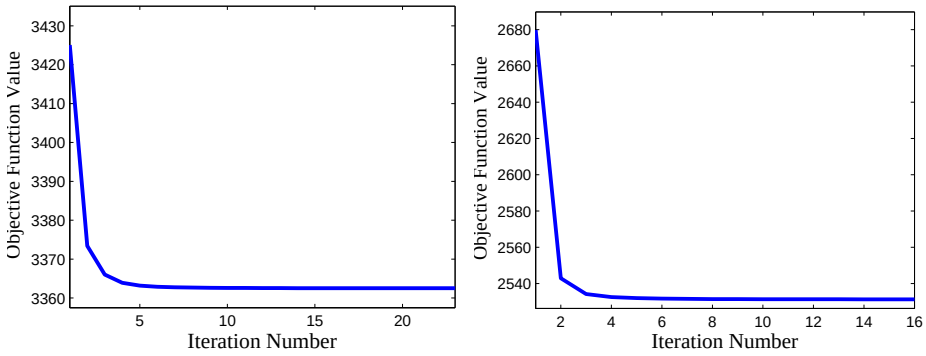
Figure 3: Performance variation *w.r.t* α and β when we fix the number of selected features at 200 for annotation. The figure shows different annotation results when using different values of α and β. With this setting, we get the best results when $\alpha = \beta = 10^{-2}$ for MSRA-MM 2.0 and when $\alpha = 1$ and $\beta = 10^{-2}$ for NUS-WIDE.

2.4.8 *Convergence Study*

As mentioned before, the proposed iterative approach monotonically decreases the objective function value in (2.9) until convergence. We conduct an experiment to validate our claim and to understand how the iterative approach works. Following the above experiments, we use $20 \times c$ training data in this experiment. The two parameters α and β are both fixed at 1 as that is the median value of the range from which the parameters are tuned. Figure 4 shows the convergence curves of our algorithm according to the objective function value in (2.9). It can be observed that the objective function value converges quickly. We also calculate the convergence time which is 17.6 and 10.9 seconds for MSRA-MM 2.0 and NUS-WIDE respectively on a personal PC with Intel Core 2 Quad 2.83GHz CPU. The convergence experiment demonstrates the efficiency of our algorithm.

2.5 CONCLUSION

In this chapter we have proposed a novel feature selection method and applied it to web image annotation. Our work integrates two state of the art innovations from shared feature subspace uncovering and joint feature selection with sparsity, thus endowing our method the following appealing proper-



(a) MSRA-MM **(b)** NUS-WIDE

Figure 4: Convergence curves of the objective function value in (2.9) using Algorithm 1. The figure shows that the objective function value monotonically decreases until convergence by applying the proposed algorithm.

ties. First, our method jointly selects the most discriminative features across the entire feature space. Additionally, our method considers the correlation between different labels, which has proved to be an effective way in multi-label learning tasks.

To validate the efficacy of our method for web image annotation, we conducted experiments on two popular image databases consisting of web images. It can be seen from the experimental results that our method outperforms classical and state-of-the-art algorithms for image annotation. Based on the observations from the experiments, we conclude that our method is robust and its feature subspace sharing foundation makes it particularly suitable for the multi-labeled web image sets used in this work. However, we would point out that our method may show different performance when different features or different datasets are used. This is because the hypothesis of our method is that the concepts of the target images are correlated and/or the original feature set is noisy and redundant. When the hypothesis does not hold, *i.e.*, the concepts have little correlation and/or the original feature set is already compact, we may not attain performance gain by using our method.

3

DISCRIMINATING JOINT FEATURE ANALYSIS FOR MULTIMEDIA DATA UNDERSTANDING[1]

In this chapter, we propose a novel semi-supervised feature analyzing framework for multimedia data understanding and apply it to three different applications: image annotation, video concept detection and 3D motion data analysis. Our method is built upon two advancements of the state of the art: (1) $l_{2,1}$-norm regularized feature selection which can jointly select the most relevant features from all the data points. This feature selection approach was shown to be robust and efficient in literature as it considers the correlation between different features jointly when conducting feature selection; (2) manifold learning which analyzes the feature space by exploiting both labeled and unlabeled data. It is a widely used technique to extend many algorithms to semi-supervised scenarios for its capability of leveraging the manifold structure of multimedia data. The proposed method is able to learn a classifier for different applications by selecting the discriminating features closely related to the semantic concepts. The objective function of our method is non-smooth and difficult to solve, so we design an efficient iterative algorithm with fast convergence, thus making it applicable to practical applications. Extensive experiments on image annotation, video concept detection and 3D motion data analysis are performed on different real-world data sets to demonstrate the effectiveness of our algorithm.

3.1 INTRODUCTION

The explosive increase of multimedia data, *i.e.*, text, image and video has brought the challenge of how to effectively index, retrieve and organize these resources. A common approach is to analyze the semantic concepts of multimedia data and to correlate concept labels with them for management tasks. Within the realm of multimedia data understanding, image and video concept understanding have obtained increasing research interest as both of them become prevalent with the popularity of the social web sites such as Flickr and YouTube. To effectively index, retrieve and manage these multimedia resources, it is necessary and beneficial to study concept analyzing techniques. Multimedia data are usually represented by different types of features. Previous works have shown that feature selection is able to reduce irrelevant and/or redundant information in the feature representation, thus facilitating subsequent analyzing tasks such as image annotation [89] [88].

Existing feature selection algorithms are achieved by different means. For instance, classical feature selection algorithms such as Fisher Score [22] compute the weights of different features and then select features one by one. These classical algorithms generally evaluate the importance of each feature individually but neglect the useful information of the correlation between different features.

1 Z. MA, F. NIE, Y. YANG, J. UIJLINGS, N. SEBE AND A. G. HAUPTMANN: "DISCRIMINATING JOINT FEATURE ANALYSIS FOR MULTIMEDIA CONTENT UNDERSTANDING". *IEEE TRANSACTIONS ON MULTIMEDIA*, 14(6): 1662-1672, 2012. IDEA PREVIOUSLY APPEARED IN: Z. MA, Y. YANG, F. NIE, J. UIJLINGS AND N. SEBE: "EXPLOITING THE ENTIRE FEATURE SPACE WITH SPARSITY FOR AUTOMATIC IMAGE ANNOTATION". IN *PROCEEDINGS OF THE ACM INTERNATIONAL CONFERENCE ON MULTIMEDIA*, PAGES 283-292, 2011.

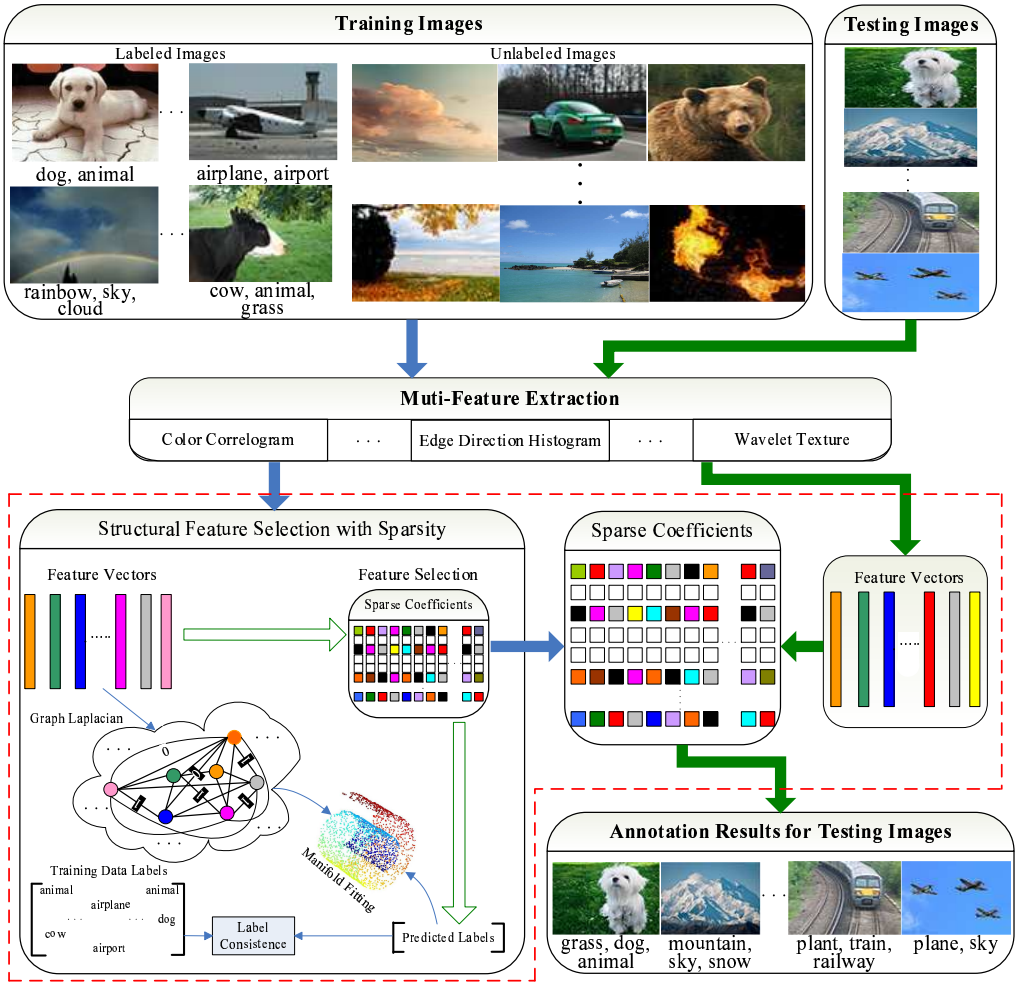


Figure 5: The general process of our method for image annotation. The red frame indicates the core part of our algorithm which analyzes the feature space for practical applications.

Another problem is that they only use labeled training samples for feature selection, which have an excessive cost in human labor. Semi-supervised learning has shown to be an effective tool for saving labeling cost by using both labeled and unlabeled data. Motivated by this fact, semi-supervised feature selection has also been proposed. For example, in [105], Zhao *et al.* have presented an algorithm based on the spectral graph theory but similarly to Fisher Score [22], their method selects features one by one. To overcome the disadvantage of selecting features individually, a plethora of state of the art approaches such as [89] [88] [62] have been proposed to extract features jointly across all data points. Nonetheless, [89] [88] [62] implement their methods in a supervised way.

Our semi-supervised feature selection method combines the strengths of joint feature selection [62] [89] [107] and semi-supervised learning [63] [75]. It utilizes both labeled and unlabeled data to select features while simultaneously consider the correlation between them. We name the proposed method Structural Feature Selection with Sparsity (SFSS).

In this chapter, we apply our method to three different multimedia analyzing tasks, *i.e.*, image annotation, video concept detection and human action analysis from 3D motion data. Image annotation correlates labels that describe semantic concepts to images. It is basically a classification problem

as it has to decide which classes an image may belong to. Annotation is realized by exploiting the correspondence between visual features and semantic concepts of the images. Video concept detection is another important tool for multimedia resource management. Similarly to image annotation, it aims to assign different concept labels to videos. We additionally apply SFSS to human action analysis from 3D motion data.

Taking image annotation as an example, we illustrate the general analyzing process of our method in Figure 5. All the training and testing images are first represented by different types of features, followed by the graph Laplacian construction. Then sparse feature selection and label prediction are conducted simultaneously by satisfying both label consistence with the training data labels and manifold fitting on the data structure. The obtained sparse coefficients can be applied to the feature vectors for selection and be directly leveraged for classification.

The main contributions are as follows:

- We combine the recent advances of feature selection and semi-supervised learning into a single framework.
- The advantage of manifold learning, which is known to be effective in exploring relationship among multimedia data, is incorporated into our framework.
- We apply our method to different applications for which we show promising performance. Our method is especially competitive when few labeled samples are available.
- A fast iterative algorithm is proposed to solve our objective function.

3.2 RELATED WORK

In this section, we briefly review the research on feature selection and semi-supervised learning.

3.2.1 *Feature Selection*

Feature selection is an effective tool in multimedia data understanding by selecting discriminating features and reducing the noise from the original data, resulting in more efficient and accurate multimedia analysis results.

In literature, there are many different feature selection algorithms. Some classical feature selection methods such as Fisher Score [22] evaluate the relevance of a feature according to the label distribution of the data. Although these classical methods have good performance when used in different applications, they have two major drawbacks. First, a lot of human labor is consumed as they require all the training data to be labeled to exploit the correlation between features and labels for feature selection. Second, their computational cost is high as they evaluate features one by one.

To progress beyond these classical methods, researchers have proposed sparsity-based feature selection to extract features jointly [107] [62] [95] [52], *i.e.*, each feature either has small scores or large scores over all data points, thus facilitating feature selection. Among various methods using this approach, $l_{2,1}$-norm regularization based algorithms have gained increasing interest for the sparsity, joint selection way and the ability to exploit the pairwise correlation among groups of features. For example, Zhao *et al.* use spectral regression with $l_{2,1}$-norm constraint to select features jointly and effectively remove redundant features in [107]. Nie *et al.* exploit joint $l_{2,1}$-norm minimization on both loss function and regularization for feature selection in [62]. Feature selection using $l_{2,1}$ models has shown its prominent performance. Therefore we propose to leverage it in our feature selection framework. However, the state of the art using $l_{2,1}$ models mostly conducts feature selection in a supervised scenario. Since in practice label information is expensive to obtain,

we design our $l_{2,1}$-norm based feature selection in a semi-supervised way which can utilize both labeled data and unlabeled data.

3.2.2 *Semi-supervised Learning*

Semi-supervised learning is widely used in many applications with the appealing feature that it can use both labeled and unlabeled data [109]. The benefit of utilizing semi-supervised learning is that we can save human labor cost for labeling a large amount of data because it can exploit unlabeled data to learn the data structure. Thus, the human labeling cost and accuracy are both considered which gives semi-supervised learning a great potential to boost the learning performance when properly designed [18].

Among the different methods, graph Laplacian based semi-supervised learning has gained most research interest. Yang *et al.* have proposed a semi-supervised approach for cross media retrieval in [96]. In [63], Nie *et al.* have proposed a Flexible Manifold Embedding framework built upon graph Laplacian and demonstrated its advantage for dimensionality reduction over other state of the art semi-supervised algorithms. In [94], a new semi-supervised algorithm based on a robust Laplacian matrix is proposed for relevance feedback. Semi-supervised learning has proved to be able to bring in promising performance by leveraging the whole data distribution for multimedia data understanding in these previous works [96] [63] [94].

3.3 METHODOLOGY

In this section, we illustrate the detailed approach of our algorithm.

3.3.1 *Problem Formulation*

We aim to select features that are mostly related to the concepts of the training data. Suppose that $X \in \mathbb{R}^{d\times n}$ indicate the training data, $Y \in \mathbb{R}^{n\times c}$ are the labels accordingly. d is the dimension of the original feature, n is the number of the training data, and c is the number of concepts. We propose to use a projection matrix W to correlate X with Y for feature selection. As W is used to select features from the original feature space and it is expected to be related to the semantic concepts, W is a $d \times c$ matrix. The problem is subsequently to design an objective function to obtain W for feature selection. In our method, we propose to exploit the $l_{2,1}$-norm based sparse feature selection due to its efficacy shown in recent works. The $l_{2,1}$-norm based methods select features by exploiting the correlations between different features and select them jointly [107] [62] [95] [52]. The boosted feature selection performance can consequently facilitate other applications. $l_{2,1}$-norm based algorithms can be generalized as the following objective function:

$$\min_{W} loss(W) + \gamma \|W\|_{2,1}, \tag{3.1}$$

where W is a projection matrix used for feature selection and $loss(W)$ is the loss function. γ is a regularization parameter. The definition of $\|W\|_{2,1}$ is:

$$\|W\|_{2,1} = \sum_{i=1}^{d} \sqrt{\sum\nolimits_{j=1}^{c} W_{ij}^2}. \tag{3.2}$$

The regularization term $\|W\|_{2,1}$ in the above function makes the optimal W sparse, according to [62] [95]. As a result, W can be regarded as the combination coefficients for the most discriminative features to achieve feature selection.

Our goal is to design a robust loss function of (3.1) through which we obtain the W for feature selection. In literature, most works built upon (3.1), *e.g.*, [8] [107] [62], are realized through supervised learning. However, we want to incorporate semi-supervised learning into (3.1) as it is known to be an effective tool for saving cost while simultaneously maintaining or enhancing the learning performance when properly designed [18]. To this end, we propose to leverage semi-supervised learning by using the widely adopted graph Laplacian.

To begin with, we have following notations. $X = [x_1, x_2, ..., x_n]$ is the training data matrix where m data are labeled. $x_i \in \mathbb{R}^d (1 \leqslant i \leqslant n)$ is the i-th datum and n is the total number of the training data. $Y = [y_1, y_2, ..., y_m, y_{m+1}, ..., y_n]^T \in \{0, 1\}^{n \times c}$ is the label matrix and c indicates the class number. $y_i \in \mathbb{R}^c (1 \leqslant i \leqslant n)$ is the label vector with c classes. Y_{ij} denotes the j-th datum of y_i and $Y_{ij} = 1$ if x_i is in the j-th class, while $Y_{ij} = 0$ otherwise. If x_i is not labeled, y_i is set to a vector with all zeros, *i.e.*, $\forall i > m$, $y_i|_{i=(m+1)}^{n} = 0^{c \times 1}$.

A typical way to construct the graph Laplacian is as follows: First, we define a matrix G whose element G_{ij} weighs the similarity between x_i and x_j as

$$G_{ij} = \begin{cases} 1 & x_i \text{ and } x_j \text{ are } k \text{ nearest neighbors;} \\ 0 & \text{otherwise.} \end{cases}$$

In (3.3), we use the Euclidean distance to evaluate whether the two samples x_i and x_j are within the k nearest neighbors in the original feature space. Second, a diagonal matrix D is formulated with $D_{ii} = \sum_{j=1}^{n} G_{ij}$. Finally, the graph Laplacian L is constructed through $L = D - G$.

The graph Laplacian is the basis of semi-supervised learning. We further leverage Manifold Regularization [11] built upon the graph Laplacian to extend our framework to a semi-supervised scenario. Manifold Regularization is adopted because multimedia data has been normally shown to possess a manifold structure [98] [44] and Manifold Regularization can explore it. Consequently, by applying Manifold Regularization to the loss function in (3.1) we obtain:

$$\arg\min_{W,b} Tr\left(W^T X L X^T W\right) + \mu \left\| X_l^T W + 1_n b^T - Y_l \right\|_F^2 + \gamma \|W\|_{2,1} . \tag{3.3}$$

where $Tr(\cdot)$ denotes the trace operator. X_l and Y_l denote the labeled training data and their ground truth labels respectively. $b \in \mathbb{R}^c$ is the bias term and $1_n \in \mathbb{R}^n$ denotes a column vector with all its n elements being 1. μ and γ are regularization parameters.

As can be seen, the optimal W obtained from (3.3) is affected by the known ground truth labels Y_l. However, inspired by the transductive classification algorithm proposed in [110] [109], we expect all the labels of the training data to contribute to the optimization of W. To achieve this goal, we denote a predicted label matrix as $F = [f_1, \dots f_n]^T \in \mathbb{R}^{n \times c}$ for all the training data in X. $f_i \in \mathbb{R}^c (1 \leqslant i \leqslant n)$ is the predicted label vector of $x_i \in X$. According to [63], F should satisfy the smoothness on both the ground truth labels of the training data and the manifold structure. Hence, it can be obtained as follows [110] [109]:

$$\arg\min_{F} Tr\left(F^T L F\right) + Tr\left((F - Y)^T U (F - Y)\right). \tag{3.4}$$

In the above function, we define a selecting diagonal matrix U whose diagonal element $U_{ii} = \infty$ if x_i is labeled and $U_{ii} = 1$ otherwise. This definition is to make the predicted labels F consistent with the ground truth labels Y. In practice, we can use a very large value, *e.g.* 10^{10} to approximate ∞.

Following the methodology in [63], we incorporate (3.4) into (3.3) and meanwhile consider all the training data with their labels (note that now we use X and F instead of X_l and Y_l respectively). Consequently, our objective function becomes:

$$\arg\min_{F,W,b} \mathrm{Tr}\left(F^T L F\right) + \mathrm{Tr}\left((F-Y)^T U (F-Y)\right) + \mu \left\| X^T W + 1_n b^T - F \right\|_F^2 + \gamma \|W\|_{2,1}. \tag{3.5}$$

From (3.5) we can see that we are able to get F, W and b simultaneously. Additionally, the optimal W obtained through (3.5) can be utilized directly for classification as W selects the features most related to the class labels.

3.3.2 *Solution*

Our objective function involves the $l_{2,1}$-norm which is non-smooth. Hence, it is not straightforward to optimize it. We propose to solve the problem as follows.

By setting the derivative of (3.5) *w.r.t.* b to zero, we obtain:

$$b^T = \frac{1}{n}(1_n^T F - 1_n^T X^T W). \tag{3.6}$$

Substituting b^T in (3.5) with (3.6), the problem becomes:

$$\arg\min_{F,W} \mathrm{Tr}\left(F^T L F\right) + \mathrm{Tr}\left((F-Y)^T U (F-Y)\right) + \mu \left\| (I - \frac{1}{n} 1_n 1_n^T) X^T W - (I - \frac{1}{n} 1_n 1_n^T) F \right\|_F^2 + \gamma \|W\|_{2,1}, \tag{3.7}$$

where I is an identity matrix. Let H represent $I - \frac{1}{n} 1_n 1_n^T$, the objective becomes:

$$\arg\min_{F,W} \mathrm{Tr}\left(F^T L F\right) + \mathrm{Tr}\left((F-Y)^T U (F-Y)\right) + \mu \left\| H X^T W - H F \right\|_F^2 + \gamma \|W\|_{2,1}. \tag{3.8}$$

Note that $H = H^T = H^2$. By setting the derivative of (3.8) *w.r.t.* F to zero, we have:

$$F = PQ, \tag{3.9}$$

where $P = (L + U + \mu H)^{-1}$ and $Q = UY + \mu H X^T W$. Substituting F in (3.8) with (3.9), we arrive at:

$$\begin{aligned} \arg\min_{W} \mathrm{Tr}\Big(& Q^T P^T (L+U) P Q - Q^T P^T U Y - Y^T U P Q + \mu W^T X H X^T W \\ & - \mu W^T X H P Q - \mu Q^T P^T H X^T W + \mu Q^T P^T H P Q\Big) + \gamma \|W\|_{2,1}. \end{aligned} \tag{3.10}$$

As $\mathrm{Tr}(Q^T P^T U Y) = \mathrm{Tr}(Y^T U P Q)$ and $\mathrm{Tr}(\mu W^T X H P Q) = \mathrm{Tr}(\mu Q^T P^T H X^T W)$, (3.10) becomes:

$$\arg\min_{W} \mathrm{Tr}\left(Q^T P^T Q - 2 Q^T P^T Q + \mu W^T X H X^T W\right) + \gamma \|W\|_{2,1}.$$

By substituting $Q = UY + \mu H X^T W$ in the above function, we get:

$$\arg\min_{W} \mathrm{Tr}\left(W^T (X H (\mu I - \mu^2 P) H X^T) W - 2\mu Y^T U P H X^T W\right) + \gamma \|W\|_{2,1}.$$

Denoting $A = XH(\mu I - \mu^2 P)HX^T$ and $B = \mu XHPUY$, the objective function becomes:

$$\arg\min_{W} \mathrm{Tr}\left(W^T A W\right) - 2\mathrm{Tr}\left(B^T W\right) + \gamma \|W\|_{2,1}. \tag{3.11}$$

3.3.3 *Algorithm*

(3.11) is a quadratic problem. First we have the following lemma to show that it is solvable.

Lemma 1 *The objective of our framework is convex.*

Proof. To prove *Lemma 1* is actually to prove that for any non-zero X, A defined in (3.11) is positive semi-definite. We therefore prove as follows:

$$\begin{aligned} A &= XH(\mu I - \mu^2 P)HX^T \\ &= \mu XHX^T - 2\mu^2 XHPHX^T + \mu^2 XHPP^{-1}PHX^T \\ &= \mu XHX^T - 2\mu^2 XHPHX^T + \mu^2 XHP(L+U+\mu H)PHX^T \\ &= \mu\left((X^T - \mu PHX^T)^T H(X^T - \mu PHX^T) + \mu XHP(L+U)PHX^T \right) \\ &= \mu\left(M^T HM + \mu XNX^T\right) \end{aligned} \tag{3.12}$$

where $M = X^T - \mu PHX^T$, $N = HP(L+U)PH$. As H and N are both larger than zero, we can easily draw the conclusion that $\mu M^T HM + \mu^2 XNX^T$ is greater than zero. Thus, $A = XH(\mu I - \mu^2 P)HX^T$ is positive semi-definite, demonstrating that the problem of our framework is convex□.

Algorithm 2: The optimization algorithm for SFSS.

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

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

1: Construct the graph Laplacian matrix $L \in \mathbb{R}^{n\times n}$;
2: Compute the selecting matrix $U \in \mathbb{R}^{n\times n}$;
3: $H = I - \frac{1}{n}1_n 1_n^T$;
4: $P = (L + U + \mu H)^{-1}$;
5: $A = XH(\mu I - \mu^2 P)HX^T$;
6: $B = \mu XHPUY$;
7: Set $t = 0$ and initialize $W_0 \in \mathbb{R}^{d\times c}$ randomly;
8: **repeat**
Compute the diagonal matrix D_t as: $D_t = \begin{bmatrix} 2\left\|w_t^1\right\|_2 & & \\ & \ldots & \\ & & 2\left\|w_t^d\right\|_2 \end{bmatrix}$;
Update W_{t+1} as: $W_{t+1} = (D_t A + \gamma I)^{-1} D_t B$;
$t = t + 1$.
until *Convergence*;
9: Return W.

To solve (3.11), we first reformulate it with the Lagrangian function as:

$$\mathcal{L}(W) = \mathrm{Tr}\left(W^T AW\right) - 2\mathrm{Tr}\left(B^T W\right) + \gamma \left\|W\right\|_{2,1}. \tag{3.13}$$

Denoting $W = [w^1, ..., w^d]^T$ with w^i as its i-th row, we define a diagonal matrix D whose diagonal elements $D_{ii} = 2\left\|w^i\right\|_2$. Then by setting the derivative of (3.13) *w.r.t.* W to zero, we obtain:

$$\begin{aligned} & 2AW - 2B + 2\gamma D^{-1}W = 0 \\ \Rightarrow\ & W = (A + \gamma D^{-1})^{-1}B = (DA + \gamma I)^{-1}DB. \end{aligned} \tag{3.14}$$

According to the mathematical deduction aforementioned, we propose an iterative approach to solve the problem in (3.11). The iterative algorithm is illustrated in Algorithm 2 and it converges. We briefly discuss the computational complexity. Computing the graph Laplacian is $\mathcal{O}(n^2)$. During the training, learning W involves calculating the inverse of a few matrices, among which the most complex part is $\mathcal{O}(n^3)$. Denote n_{te} as the number of testing data. Once we get W, it takes $c \times d \times n_{te}$ multiplications to predict the categories of the testing data. For large scale data sets $n_{te} \gg c$ and $n_{te} \gg d$. Thus, the classification complexity is approximately linear *w.r.t.* n_{te}, which is very efficient.

The convergence of Algorithm 2 can be proved following the work in [62] [95] [55].

3.4 EXPERIMENTS

We evaluate our method on image annotation, video concept detection and 3D motion data analysis respectively. Additional analyzing experiments are also performed to assess the overall performance of our method. These include a parameter sensitivity study and a convergence study.

3.4.1 *Compared Algorithms*

To evaluate the advantage of our method for multimedia data understanding, we compare it with the following algorithms:

- Fisher Score (FISHER) [22]: a classical method. It selects the most discriminative features by evaluating the importance of each feature individually.
- Sparse Multinomial Logistic Regression via Bayesian L1 Regularisation (SBMLR) [15]: a sparsity based state of the art method. It realizes sparse feature selection by using a Laplace prior.
- Group Lasso with Logistic Regression (GLLR) [89]: a recently proposed method based on a sparse model. It utilizes group lasso extended with logistic regression to select both sparse and discriminative groups of homogeneous features.
- Feature Selection via Joint $l_{2,1}$-Norms Minimization (FSNM) [62]: a recent sparse feature selection algorithm. It employs joint $l_{2,1}$-norm minimization on both loss function and regularization for joint feature selection.
- Semi-supervised Feature Selection via Spectral Analysis (sSelect) [105]: a semi-supervised feature selection method based on spectral analysis.
- Locality sensitive semi-supervised feature selection (LSDF) [104]: a semi-supervised approach based on two graph construction, *i.e.*, within-class graph and between-class graph.

We use the regularized least square regression for classification after FISHER, SBMLR, FSNM, sSelect and LSDF finish the feature selection. In contrast, GLLR and SFSS can learn the classifiers directly when performing feature selection.

Table 6 illustrates the different properties of each method used in our experiments.

Table 6: A brief comparison between the different methods.

Method	SS[a]	S[b]	J-FS[c]	I-FS[d]	One-Step[e]
FISHER [22]		√		√	
SBMLR [15]		√	√		
GLLR [89]		√	√		√
FSNM [62]		√	√		
sSelect [105]	√			√	
LSDF [104]	√			√	
SFSS	√		√		√

[a] semi-supervised.
[b] supervised.
[c] feature selection across all data points.
[d] feature selection one by one.
[e] simultaneous classifier learning.

3.4.2 *Experimental Data Sets*

Image Annotation

Three data sets, *i.e.*, Corel-5K [30] [29], MSRA-MM [43] and NUS-WIDE [17] are used in our experiments. The following is a brief description of the three data sets.

Corel-5K: In our experiment, we use the standard data set used in [30] [29]. Corel-5K consists of 5,000 images from 50 different categories. Three types of color features (color histogram, color moment, and color coherence) and three types of texture features (Tamura coarseness histogram, Tamura directionality, and MSRSAR texture) are used to represent the images.

MSRA-MM: The data set used in our experiments is a subset of the original MSRA-MM 2.0 data set, which includes 50,000 images related to 100 concepts. However, 7,734 images within it are not associated with any labels. We have removed these images and obtained a subset of 42,266 labeled images. Three feature types used in [89], namely Color Correlogram, Edge Direction Histogram and Wavelet Texture are combined in our experiments.

NUS-WIDE: It consists of 209,347 labeled real-world images collected from Flickr which are associated with 81 concepts. The images are also represented by the combination of Color Correlogram, Edge Direction Histogram and Wavelet Texture.

Video Concept detection

We choose the Kodak consumer video data set [46] and the CareMedia data set [1].

Kodak: It consists of 1,358 consumer video clips and 5,166 key-frames are extracted accordingly. Among these key-frames, 3590 ones are annotated. We use all the annotated key-frames belonging to 22 concepts in our experiments for video concept detection. Color Correlogram, Edge Direction Histogram and Wavelet Texture are used to represent the key-frames.

CareMedia: The video data set was collected by Carnegie Mellon University to provide useful statistics to help doctors' diagnosis and patients' heath status assessment. 15 geriatric patients' activities in public spaces were recorded in a nursing home [1]. We test the performance by annotating the following 5 concepts which are concerned with patients' detailed behaviors: 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). The MoSIFT fea-

ture [100] is used to represent each video sequence. In this experiment, we use a subset consisting of 3913 video sequences recorded by one camera in the dining room.

3D Motion Data Analysis

We choose the HumanEva 3D motion database [74]. There are five types of actions, namely boxing, gesturing, jogging, walking and throw-catch performed by different subjects in this database. We randomly sample 10,000 data of two subjects (5,000 per subject) similarly to [97] [64] in our experiment. The action of the two subjects is considered to be different. We simultaneously recognize the identities and actions, which comes to 10 semantic categories in total. Each action is encoded as a collection of 16 joint coordinates in 3D space, thus resulting in a 48 dimensional feature vector. On top of that, we compute the Joint Relative Features between different joints and get a feature vector with 120 dimensions. The two kinds of feature vectors are further combined to generate a 168 dimensional feature.

3.4.3 *Experimental Setup*

First, a training set for each data set is generated randomly consisting of n samples, among which $m\%$ samples are labeled. The detailed settings are given in Table 7. The remaining data of each data set work as the corresponding testing set. We generate the training and testing sets 5 times and report the average results with standard deviation.

Table 7: The settings of the training sets.

	Size (n)	Labeled Percentage (m)[2]
Corel-5K	2500	2, 5, 10, 25, 50, 100
MSRA-MM	10000	1, 5, 10, 25, 50, 100
NUS-WIDE	10000	1, 5, 10, 25, 50, 100
Kodak	2000	2, 5, 10, 25, 50, 100
CareMedia	1000	1, 5, 10, 25, 50, 100
HumanEva	3000	1, 5, 10, 25, 50, 100

In the experiments, we have to tune two types of parameters. One is the parameter k that specifies the k nearest neighbors used to compute the graph Laplacian. We fix it at 15 following the setting in our previous work [55]. The other one is the regularization parameters which are represented as μ and γ in (3.5). We tune them from $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^{2}, 10^{3}\}$ and report the best results.

To evaluate the classification performance, we use Mean Average Precision (MAP) as the evaluation metric for its stability and discriminating capability.

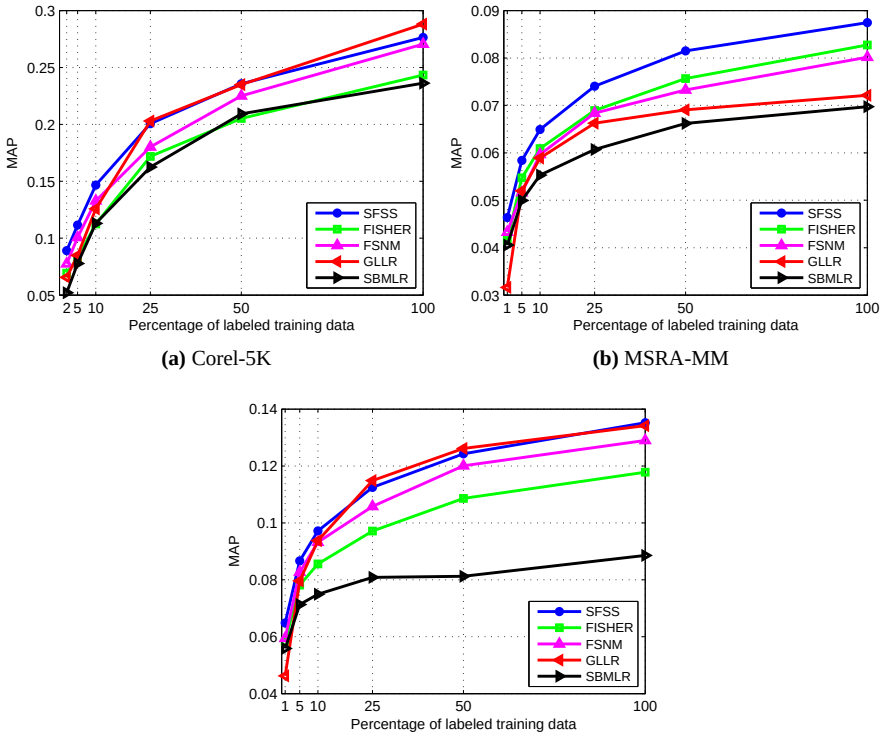
3.4.4 *Multimedia Understanding Performance*

In this section, we report the experimental results on image annotation, video concept detection and 3D motion data analysis respectively.

2 Note that the settings of the labeled training data on Corel-5K and Kodak are slightly different from others to guarantee that each concept class has at least one labeled training data.

Image Annotation

Figure 6 shows the annotation results when different percentages of data are labeled. Table 8 to Table 10 show the results when 2% (Corel-5K) or 1% (MSRA-MM&NUS-WIDE), 5% and 10% of the training data are labeled. We have the following observations from the experimental results: 1) As the number of labeled training data increases, the performance increases. 2) Our method is the only one which has consistently high scores on all three data sets. Other methods have varying degrees of success on each data set. 3) When 25% or more of the training data are labeled, our method is competitive with the best algorithms compared or better. Yet the more labeled data is available, the smaller our advantage is over other supervised algorithms. On the Corel-5K data set GLLR [89] slightly outperforms our method; on the NUS-WIDE data set our method is competitive with GLLR [89]; on the MSRA-MM data set our method outperforms all other methods. 4) Finally, when less than 25% of the data are labeled, our method consistently outperforms other methods on all three data sets. This is especially visible on the Corel-5K and MSRA-MM data sets.



(a) Corel-5K **(b)** MSRA-MM

(c) NUS-WIDE

Figure 6: Performance comparison of image annotation *w.r.t.* the percentage of labeled training data. When 10% or less of the data are labeled our method outperforms all other algorithms. When 25% or more of the data are labeled, our method yields top performance or, on the MSRA-MM data set significantly better performance.

Video Concept Detection

We illustrate the video concept detection results in Figure 7, Table 11 and Table 12. It can be seen from Figure 7 that our method has the top one performance over other algorithms. Table 11 and

Table 8: Performance comparison of image annotation (MAP±Standard Deviation) when 2% (Corel-5K) or 1% (MSRA-MM&NUS-WIDE) training data are labeled.

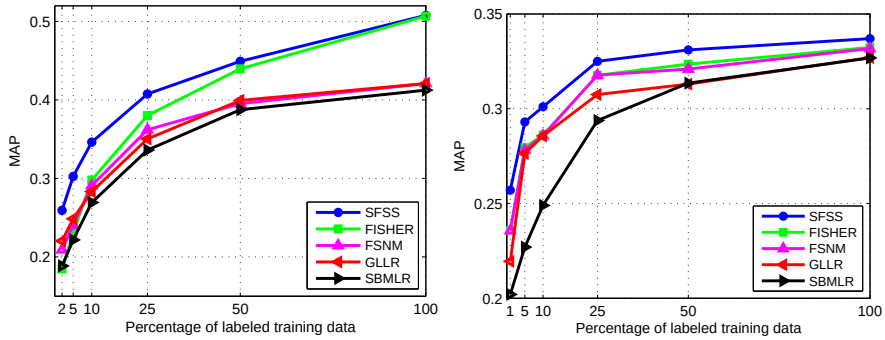
	Corel-5K	MSRA-MM	NUS-WIDE
SFSS	**0.090±0.008**	**0.047±0.002**	**0.065±0.002**
FISHER [22]	0.069±0.006	0.041±0.002	0.058±0.003
GLLR [89]	0.066±0.008	0.032±0.008	0.046±0.007
FSNM [62]	0.078±0.007	0.043±0.002	0.059±0.002
SBMLR [15]	0.052±0.004	0.040±0.002	0.056±0.003

Table 9: Performance comparison of image annotation (MAP±Standard Deviation) when 5% training data are labeled.

	Corel-5K	MSRA-MM	NUS-WIDE
SFSS	**0.112±0.009**	**0.059±0.002**	**0.087±0.003**
FISHER [22]	0.083±0.007	0.055±0.002	0.078±0.002
GLLR [89]	0.085±0.010	0.052±0.001	0.079±0.001
FSNM [62]	0.101±0.007	0.051±0.002	0.082±0.002
SBMLR [15]	0.078±0.005	0.050±0.002	0.071±0.003

Table 10: Performance comparison of image annotation (MAP±Standard Deviation) when 10% training data are labeled.

	Corel-5K	MSRA-MM	NUS-WIDE
SFSS	**0.147±0.009**	**0.065±0.001**	**0.097±0.002**
FISHER [22]	0.113±0.003	0.061±0.002	0.086±0.003
GLLR [89]	0.126±0.015	0.059±0.001	0.094±0.002
FSNM [62]	0.133±0.009	0.060±0.001	0.093±0.003
SBMLR [15]	0.113±0.013	0.055±0.002	0.075±0.007



(a) Kodak **(b)** CareMedia

Figure 7: Performance comparison of video concept detection *w.r.t.* the percentage of labeled training data. Our method is consistently better than other compared methods.

Table 11: Performance comparison of video concept detection (MAP±Standard Deviation) *w.r.t.* 2%, 5% and 10% labeled data on Kodak data set.

	2% labeled	5% labeled	10% labeled
SFSS	**0.259±0.015**	**0.303±0.023**	**0.346±0.027**
FISHER [22]	0.185±0.021	0.230±0.009	0.298±0.022
GLLR [89]	0.220±0.028	0.249±0.015	0.283±0.024
FSNM [62]	0.210±0.025	0.240±0.009	0.291±0.019
SBMLR [15]	0.189±0.029	0.222±0.009	0.269±0.026

Table 12: Performance comparison of video concept detection (MAP±Standard Deviation) *w.r.t.* 1%, 5% and 10% labeled data on CareMedia data set.

	1% labeled	5% labeled	10% labeled
SFSS	**0.257±0.018**	**0.293±0.009**	**0.301±0.014**
FISHER [22]	0.235±0.017	0.279±0.012	0.286±0.014
GLLR [89]	0.220±0.017	0.276±0.017	0.286±0.011
FSNM [62]	0.236±0.014	0.278±0.011	0.286±0.014
SBMLR [15]	0.202±0.003	0.227±0.004	0.249±0.007

Table 12 give the detailed results when 2% or 1%, 5% and 10% training data are labeled. We observe that our method is especially competitive when few training data are labeled.

Table 13: Performance comparison of 3D motion data analysis (MAP±Standard Deviation) *w.r.t.* 1%, 5% and 10% labeled data.

	1% labeled data	5% labeled data	10% labeled data
SFSS	**0.860±0.021**	**0.984±0.015**	**0.994±0.012**
FISHER [22]	0.453±0.016	0.608±0.022	0.678±0.019
GLLR [89]	0.559±0.037	0.645±0.024	0.666±0.013
FSNM [62]	0.480±0.013	0.615±0.024	0.696±0.018
SBMLR [15]	0.126±0.055	0.554±0.022	0.608±0.024

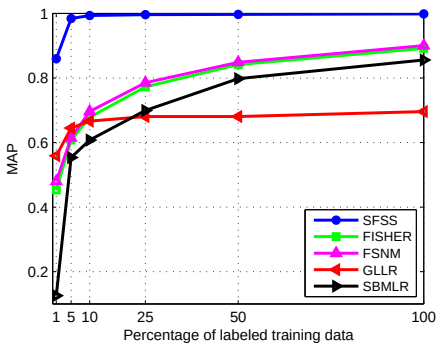
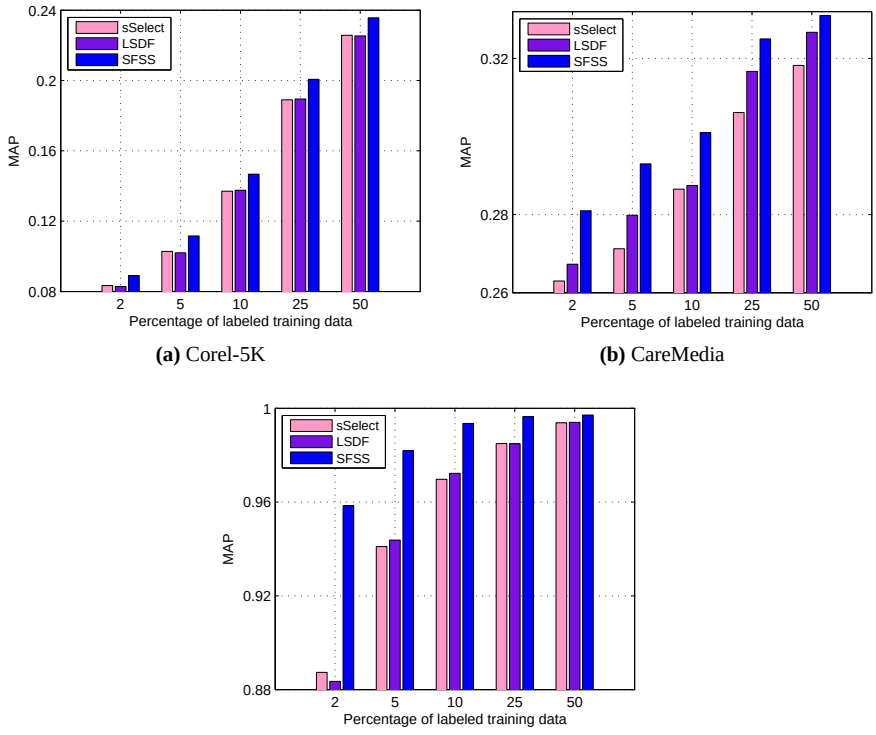


Figure 8: Performance comparison of 3D motion data analysis *w.r.t.* the percentage of labeled training data. Our method has much advantage over other algorithms. Good performance can be achieved even when very few training data are labeled.

3D Motion Data Analysis

The results of 3D motion data analysis are illustrated in Table 13 and Figure 8. From Table 13 and Figure 8 we observe that our method gains huge advantage over other compared approaches. We also notice that SFSS gets satisfactory performance when only 5% training data are labeled and it shows nearly perfect performance (close to 1 in terms of MAP) when over 10% training data are labeled. Intuitively, this indicates that the exploitation of the manifold structure has contributed considerably to the whole analyzing performance.

3.4.5 *Comparison with Other Semi-supervised Feature Selection Methods*



(a) Corel-5K **(b)** CareMedia

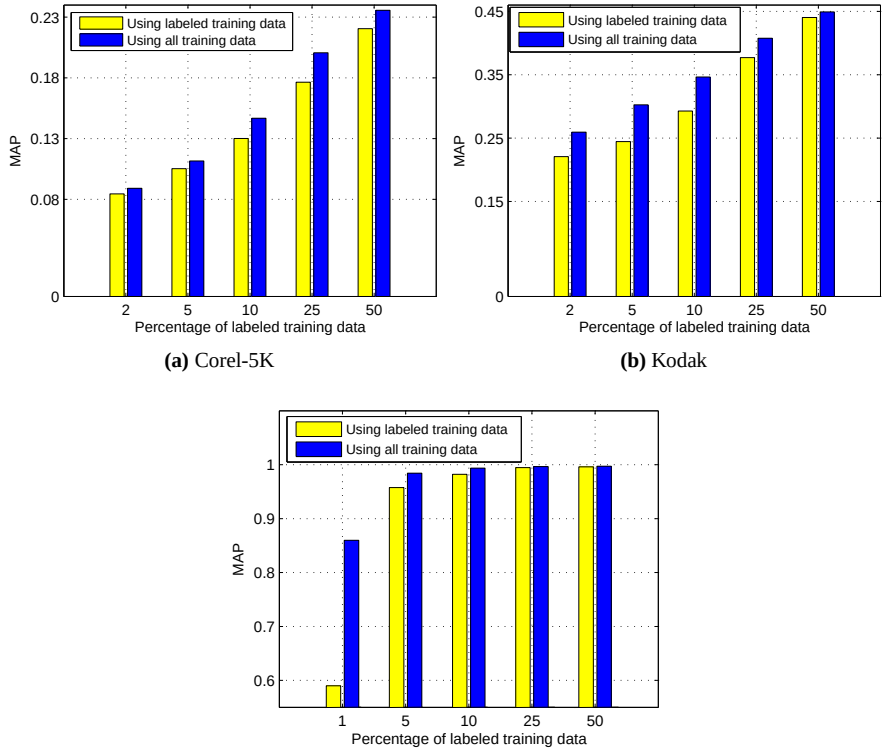
(c) HumanEva

Figure 9: Performance comparison with semi-supervised approaches on different applications *w.r.t.* the percentage of labeled training data. Our method outperforms sSelect and LSDF for all settings and has much advantage when few training data (2% and 5%) are labeled.

In this section, we compare SFSS with two state of the art semi-supervised feature selection algorithms, namely sSelect and LSDF. The experiments are conducted on Corel-5K, CareMedia and HumanEva data sets for different applications. To be consistent, 2%, 5%, 10%, 25% and 50% training data are labeled in this experiment for all data sets. The results are shown in Figure 9. It can be observed that our method consistently outperforms both sSelect and LSDF. The advantage is especially visible when only few training data are labeled, *i.e.*, 2% or 5%. Semi-supervised methods are used for the cases when we only have limited number of labeled training data. We thus conclude that SFSS is much better than sSelect and LSDF as it has much higher accuracy when only few labeled training data are available.

3.4.6 *Influence of the Unlabeled Data*

To study the influence of unlabeled training data on the multimedia understanding performance, we conduct an experiment correspondingly.



(a) Corel-5K

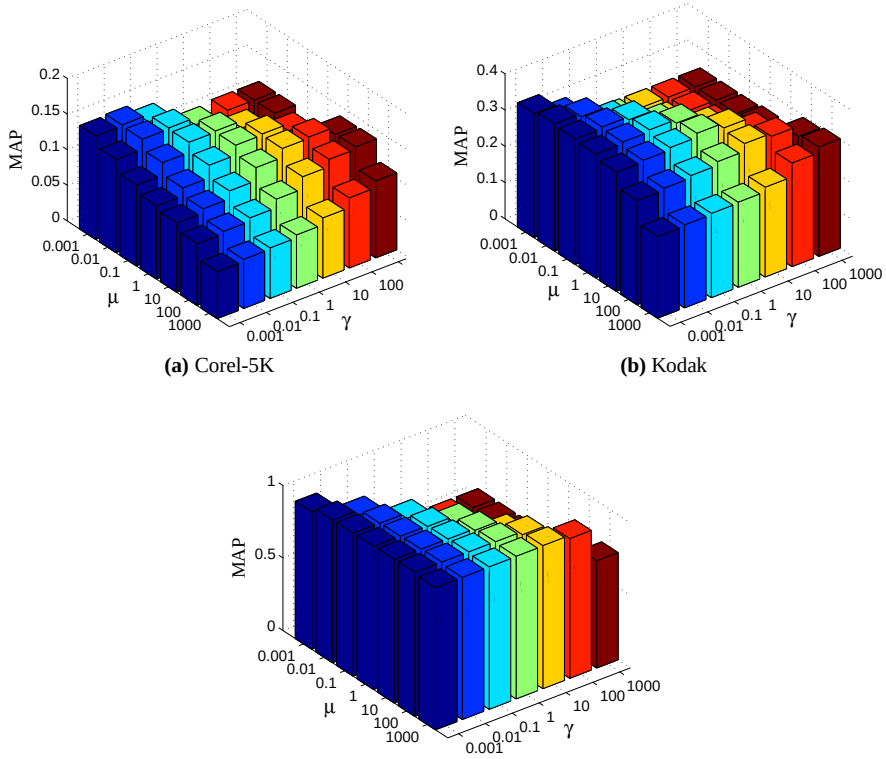
(b) Kodak

(c) HumanEva

Figure 10: The influence of unlabeled data on different multimedia analyzing tasks. The blue bar stands for the performance of SFSS. The yellow bar indicates the results that are obtained by using only labeled data (no unlabeled data). The comparisons between the two approaches show that using unlabeled data improves the analyzing performance.

The unlabeled data in the training set are left out and we only use labeled training data to conduct feature analysis. Then we compare the results with the ones that are achieved by using the entire training set including both labeled and unlabeled data. The experiment is performed on Corel-5K, Kodak and HumanEva data sets for each application respectively. 2% (Corel-5K, Kodak) or 1% (HumanEva), 5%, 10%, 25% and 50% training data are labeled as different settings. Figure 10 illustrates the comparisons.

It can be seen that using unlabeled data besides the labeled data yields better results over using the labeled data alone. When 10% of the data are labeled, by also using unlabeled data we obtain relative improvements of 13% on the Corel-5K data set and 18% on the Kodak data set. Yet the situation is different for the HumanEva data set. The largest improvement, 45%, is obtained when only 1% of the data are labeled. However, as the percentage of labeled training data grows, the performance by using only labeled training data increases dramatically. The reason could be that the HumanEva data set is clean and easy to analyze. Moreover, the MAP closes in on 1 after 5% training data are labeled, which makes the contribution of the unlabeled data on the performance limited. The improvements in semi-supervised learning are due to the learning of the manifold structure. In theory, the more



(a) Corel-5K **(b)** Kodak

(c) HumanEva

Figure 11: Performance variance *w.r.t.* μ and γ. The figure displays different results when using different μ and γ.

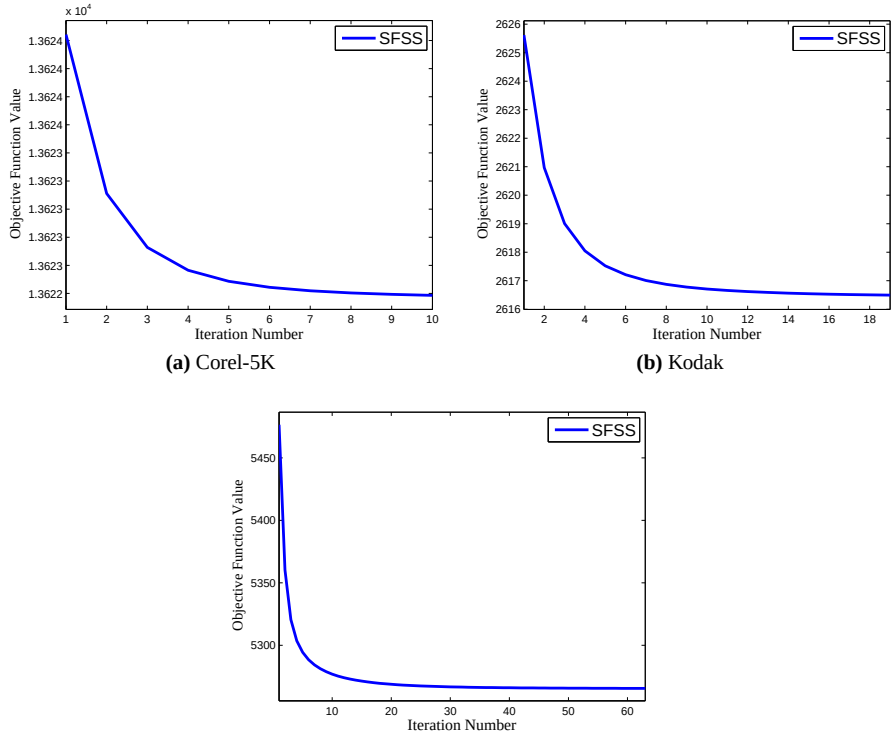
data points that one has, the better the manifold structure that can be learned. This saturates with enough data. The Corel-5K data set still has huge benefits from using all data instead of 50% for learning the manifold structure. For the HumanEva data set the manifold structure is very important as without this manifold the performance is much lower in general (see Figure 8). Figure 10c shows that this manifold is learned well using 25% of the data, after which performance is close to optimal for both the fully supervised and semi-supervised settings.

3.4.7 *Parameter Sensitivity Study*

In Figure 11, we show the influence of the two parameters μ and γ on the performance of different applications using Corel-5K, Kodak and HumanEva data sets when 10% training data are labeled. It can be seen that the MAP is generally higher when μ and γ are comparable for Corel-5K and Kodak data sets. In contrast, there is no analogous rule identifiable about when the optimal results are obtained for HumanEva data set. The phenomenon demonstrates that the parameter sensitivity is presumably related to the properties of the different data sets.

3.4.8 *Convergence Study*

In the previous section, we have proved that the objective function in (3.5) converges by using the proposed algorithm. For practical applications it is interesting how fast our algorithm converges.



(a) Corel-5K **(b)** Kodak

(c) HumanEva

Figure 12: Convergence curves of the objective function value in (3.5) using Algorithm 2. The figure shows that the objective function value monotonically decreases until convergence by applying the proposed algorithm.

Figure 12 shows the convergence curves of our optimization algorithm *w.r.t.* the objective function value in (3.5) on Corel-5K, Kodak and HumanEva when μ and γ are fixed at 1. It can be seen that our algorithm converges within as few as 10-20 iterations.

3.5 CONCLUSION

We have proposed a new multimedia analyzing method built upon feature analysis. The method takes advantage of joint feature selection with sparsity, manifold regularization and transductive classification. Additionally, to solve the non-smooth objective function of our algorithm, we have proposed an iterative approach. Our method is general and can be applied to different applications. In this chapter, we evaluate its performance on image annotation, video concept detection and 3D motion data analysis. The experimental results have demonstrated that our method consistently outperforms the other compared algorithms for different analyzing tasks. Our method considers the characteristic of multimedia data, the labeling cost, the computational efficiency and the adaptability. As shown in the experiments, our method is suitable for some multimedia understanding applications. It is, however, worth mentioning that when the dataset has no structured manifold, the manifold learning embedded in our algorithm may lose its power, thus leading to little performance gain. Additionally, if the original feature set is already discriminating enough, the feature analysis function in our method is likely to contribute less to the overall performance boost.

4

MULTIMEDIA EVENT DETECTION USING A CLASSIFIER-SPECIFIC INTERMEDIATE REPRESENTATION[1]

Multimedia event detection (MED) plays an important role in many applications such as video indexing and retrieval. Current event detection works mainly focus on sports and news event detection or abnormality detection in surveillance videos. Differently, our research aims to detect more complicated and generic events within a longer video sequence. In the past, researchers have proposed using intermediate concept classifiers with concept lexica to help understand the videos. Yet it is difficult to judge how many and what concepts would be sufficient for the particular video analysis task. Additionally, obtaining robust semantic concept classifiers requires a large number of positive training examples, which in turn has high human annotation cost. In this chapter, we propose an approach that exploits the external concepts-based videos and event-based videos simultaneously to learn an intermediate representation from video features. Our algorithm integrates the classifier inference and latent intermediate representation into a joint framework. The joint optimization of the intermediate representation and the classifier makes them mutually beneficial and reciprocal. Effectively, the intermediate representation and the classifier are tightly correlated. The classifier dependent intermediate representation not only accurately reflects the task semantics but is also more suitable for the specific classifier. Thus we have created a discriminative semantic analysis framework based on a tightly coupled intermediate representation. Extensive experiments on multimedia event detection using real-world videos demonstrate the effectiveness of the proposed approach.

4.1 INTRODUCTION

Research on video indexing and retrieval has long been faced with the challenge of semantic gap between low-level features and high-level semantic content description of videos [28] [77]. To bridge the semantic gap, various approaches have been proposed to help analyze the semantic content of videos, either at concept level or at event level.

According to [54], a “concept" means an abstract or general idea inferred from specific instances of objects, scenes and actions such as *fish*, *outdoor* and *boxing*. Concepts are lower level descriptions of multimedia data which usually can be inferred with a single image or a few video frames. An “event" refers to an observable occurrence that interests users. Compared with concepts, events are higher level descriptions of multimedia data. A meaningful event builds upon many concepts and is unlikely to be inferred with a single image or a few video frames. For example, the event *landing a fish* includes many concepts such as *people*, *fish*, *fishing rod* together with the action *landing*, and it

1 Z. MA, Y. YANG, N. SEBE, K. ZHENG, A. G. HAUPTMANN: “MULTIMEDIA EVENT DETECTION USING A CLASSIFIER-SPECIFIC INTERMEDIATE REPRESENTATION". *IEEE TRANSACTIONS ON MULTIMEDIA*, 15(7):1628-1637, 2013. IDEA PREVIOUSLY APPEARED IN: Z. MA, Y. YANG, A. G. HAUPTMANN AND N. SEBE: “CLASSIFIER-SPECIFIC INTERMEDIATE REPRESENTATION FOR MULTIMEDIA TASKS". IN *PROCEEDINGS OF THE ACM INTERNATIONAL CONFERENCE ON MULTIMEDIA RETRIEVAL*, 2012.

usually happens in a longer video sequence. We cannot tell if it is a *landing a fish* event if we only see a person sitting on a boat in one image or a few frames.

Annotation and detection are two different topics of both concept and event analysis [54]. Multimedia annotation, also known as recognition, aims to associate a datum with one or multiple semantic labels (tags) [54]. Many approaches have been proposed to improve annotation accuracy for both images and videos [79]. Detection identifies the occurrence of a class of interest in a large pool of data. In contrast with annotation for which both the training and testing data are from a fixed number of classes, the training and testing data in detection can be from an infinite number of classes [54]. Hence, detection is a more challenging problem.

The TREC Video Retrieval Evaluation (TRECVID) community has notably contributed to the research of video concept or event detection [4] [60] [76]. In the field of multimedia, many other works have also focused on *concept detection*, e.g., [78] [91] [45]. However, the research on video *event detection* is still in its infancy. Most existing research on event detection is limited to the sport events, news events, events with repetitive patterns like *running* or unusual events in surveillance videos [71] [73] [83] [5]. The "Event detection in Internet multimedia (MED)[2]" launched by TRECVID aims to encourage new technologies for detecting more complicated events, *e.g.*, *feeding an animal*. Ma *et al.* have made the first attempt on Ad Hoc detection of this type of events, for which only 10 positive example are available for training [54]. For this kind of events, there are huge intra-class variations. For example, an event "feeding an animal" can be either feeding a cat at home with cat food in a small container, or feeding a horse in a farm with a bundle of grass. Besides, they are usually characterized by long video sequences, which necessitates the exploration of all the sequences for analysis.

Recent research has shown that the performance of multimedia semantic analysis can be improved through proper machine learning approaches [41] [94] [102]. Therefore, it is reasonable to leverage good low-level features as well as effective machine learning algorithms on video data for MED. We propose a new algorithm for MED, which is extended from our previous work [51]. Our method has the following attributes:

1) Our algorithm learns an intermediate representation of videos by exploiting the *target videos* and *external video* archives together. In this chapter, the target videos are the videos depicting the event to be detected. The external videos are the auxiliary labeled video archives that are used to help learn the intermediate representation. The intermediate representation is a compact vector representation derived from the Bag-of-Words features of the videos through a transformation, during which the discriminative information is encoded.

2) Our algorithm integrates representation inference and classifier training into a joint framework. In this way, the intermediate representation is tightly coupled with the loss function used for the classifier.

3) A robust loss function is used in our objective function, making the performance more robust to outliers.

We name our method Semantic Analysis via Intermediate Representation (SAIR). The intermediate representation is dependent on the classifier while the classifier training benefits from the representation. The mutual benefit and reciprocality between the intermediate representation and the classifier endows the classification framework good capability for multimedia event detection.

4.2 RELATED WORK

In this section, we briefly review some related works, which cover multimedia representation and semantics understanding.

2 http://www.nist.gov/itl/iad/mig/med11.cfm

4.2.1 *Multimedia Low-level Feature Representation*

A common approach for low-level feature representation is to extract the key frames of videos and then generate features based on these frames. For example, traditional features include Color Correlogram, Edge Direction Histogram, Wavelet Texture, *etc.* Newly designed features, *e.g.*, SIFT draw more research interest for their discriminating capability [47]. Some other features can capture the spatial-temporal information, *e.g.*, STIP feature [39] and MoSIFT feature [100], and have shown promising performance in video semantic analysis.

Apart from visual features, some other modalities, which provide different yet complementary information, can also be used to represent videos. For example, textual representation based on Automatic Speech Recognition (ASR) and Optical Character Recognition (OCR), and auditory features based on Mel-frequency Cepstral Coefficients (MFCC) have also been frequently used to represent videos [108].

4.2.2 *Learning to Refine Multimedia Representation*

Multimedia representation refinement aims to obtain a more compact as well as accurate feature representation of multimedia data [31] [73] [98] [94] [82]. Shyu *et al.* propose a subspace based data mining framework for video concept/event detection [73]. To exploit the semantic relatedness among multiple modalities, Yang *et al.* propose a manifold learning based algorithm to infer a unified representation of different media types for cross media retrieval [98]. Based on users' feedbacks, a long term relevance feedback algorithm is proposed in [94] to refine the multimedia representation for better retrieval performance. In [82], a sparse projection method is proposed to infer a sparse representation for videos, by which the efficiency of video classification is improved. These research efforts have shown that multimedia data can be refined by proper machine learning algorithms, thus resulting in better performance for multimedia analysis. However, in most of these works, the refinement and the classifier training are independent from each other. As it is uncertain which classifiers benefit the most from these refinement algorithms, the performance improvement could be limited. Instead, we propose an integrated framework which learns a refined representation and a classifier jointly. As the refined representation is correlated with the loss function used in the classifier, the classifier dependent intermediate representation not only accurately reflects the task semantics but is also more suitable for the specific classifier, thus resulting in boosted classification accuracy.

4.2.3 *Concepts-based Representation*

Recently, some researchers suggest using concepts-based representation for video semantic understanding. A number of researchers have been building a variety of semantic concept detectors, such as those related to people (face, anchor), acoustic (speech, music), genre (weather, financial, sports), scene, *etc.* [28], and a series of concept lexica have been established, *e.g.*, LSCOM [61] and MediaMill [78]. 346 concepts have been defined for the TRECVID 2011 semantic indexing task. With these annotation corpora, different concept detectors can be trained. Therefore, videos can be represented by the concept detection results of those detectors [27]. If sufficient concept detectors are properly trained and appropriately applied, the concepts-based representation of videos, which is a set of textual descriptors, is more capable of reflecting video semantics. However, such approach is still confronted with some problems. First, it requires many labeled data to train intermediate concept classifiers, which costs much human labor. For example, while the full LSCOM set contains over 2600 concepts, many of them are unannotated or contain no positive instances [61]. Second, only concept-based archives have been used to infer the representation so far. In recent years, sev-

eral event-based video archives have been presented in the community. Effective usage of these event-based videos for learning intermediate representation could be another potential solution for improving multimedia event detection.

4.3 THE PROPOSED ALGORITHM

In this section, our algorithm is presented in details followed by an algorithm for solving the objective function. *Classifier-specific* in our method means being tightly coupled with the particular loss function used by the classifier.

4.3.1 *Learning An Intermediate Representation*

We first illustrate the traditional approach of concepts-based representation for multimedia analysis. Then we formulate our method which goes beyond the traditional approach.

Traditional Approach

Suppose there are n example videos, whose low-level features are $\{x_1, ..., x_n\}$. Here $x_i \in \mathbb{R}^d$ denotes the low-level feature of the video and d is the dimension of the feature. x_i is either a positive or negative example for a particular event, or an example of the external videos used to help learn the intermediate representation. Let y_i be the label of x_i, indicating category of x_i. A general approach to train a classifier f can be formulated as minimizing the following objective function

$$\min_f \sum_{i=1}^{n} \ell\left(f(x_i), y_i\right) + \alpha\Omega(f), \tag{4.1}$$

where $\ell(\cdot,\cdot)$ is a loss function and $\Omega(f)$ is a regularization function on f with α as a regularization parameter. Clearly, there are three main components to be properly designed, which are the feature representation x_i, the loss function $\ell(\cdot,\cdot)$, and the regularization function $\Omega(\cdot)$.

Using the concepts-based representation as in [27] [28] for multimedia event detection, we need another m annotated videos $\{x_{n+1}, ..., x_{n+m}\}$ from c classes with groundtruth labels $\{y_{n+1}, ..., y_{n+m}\}$. For the k-th class there are m_k positive examples. The videos $\{x_{n+1}, ..., x_{n+m}\}$ are used to pre-train c classifiers $g_k|_{k=1}^{c}$, one for each intermediate concept. For each training or testing video $x_i (1 \leqslant i \leqslant n)$, the classifiers $g_k|_{k=1}^{c}$ are applied to detect the intermediate concepts. In this way, $x_i (1 \leqslant i \leqslant n)$ is represented by a c dimensional vector, with each dimension corresponding to an intermediate concept. More specifically, the following two steps are taken. In the first step, c classifiers $\{g_1, ..., g_c\}$ are trained by minimizing the following objective function

$$\min_{g_1,...,g_c} \sum_{k=1}^{c} \sum_{j=1}^{m} \widetilde{\ell}\left(g_k(x_{n+j}), y_{n+j}\right) + \alpha\widetilde{\Omega}(g_k), \tag{4.2}$$

where $\widetilde{\ell}(\cdot,\cdot)$ and $\widetilde{\Omega}(f)$ are the loss function and the regularization function respectively and α is a parameter. Once the c classifiers $\{g_1, ..., g_c\}$ are obtained, we convert the original feature representation $x_i (1 \leqslant i \leqslant n)$ to the concepts-based representation $z_i = [z_{1i}, ..., z_{ci}] \in \mathbb{R}^c$ by $z_{ki} = g_k(x_i)$ $(1 \leqslant k \leqslant c)$. In the second step, the event detector f can be trained based on the new representation z_i $(1 \leqslant i \leqslant n)$ in the same way of (4.1), *i.e.*,

$$\min_f \sum_{i=1}^{n} \ell\left(f(z_i), y_i\right) + \alpha\Omega(f) \Rightarrow \min_f \sum_{i=1}^{n} \ell\left(f(g(x_i)), y_i\right) + \alpha\Omega(f), \tag{4.3}$$

where $g(x_i) = [g_1(x_i), ..., g_c(x_i)]$. For each testing video x_{te}, the decision score s_{te} indicating whether the event occurs in the video x_{te} is given by

$$s_{te} = f(g(x_{te})). \tag{4.4}$$

Although the traditional concepts-based representation [28] [27] is expected to be more precise than low-level features, this kind of approach suffers from some practical problems in implementation. First, it is time-consuming to find and annotate a large amount of positive examples to train many concept classifiers. Second, the number of concepts is limited and it remains unclear how many concepts (and what concepts as well) would be sufficient for some applications, *e.g.*, multimedia event detection. Third, the pre-trained concept classifiers are yet to be sufficiently reliable. Fourth, given a particular event to detect, only some concepts are discriminative while others are comparatively useless or even noisy. Taking “landing a fish" event as an example, some concepts like “fish" and “boat" are very discriminative, while “clouds" and “face" are less informative. It is a nontrivial task to define the ontology for different events, which are dynamic and diverse.

Joint Learning of Classifier and Representation with External Videos

In the traditional way of multimedia event detection using concepts-based representation, the concept classifiers $g_k|_{k=1}^c$ and multimedia event detector f are trained individually, as shown in (4.2) and (4.3). There is no guarantee, however, that the two are tightly correlated. Besides, training a large number of $g_k|_{k=1}^c$ is time consuming, while it remains unclear how large c should be. A question then comes up: Can we learn an intermediate representation closely related to a particular multimedia event, and the event detector without requiring many pre-labeled data? As demonstrated in [54], the classifier of external concepts-based videos and the event detector have shared components. Exploiting such information is beneficial for multimedia event detection. Different from [54],we assume that the external concepts-based videos and the event-based videos have a common intermediate representation. Specifically, we propose to simultaneously learn f and an intermediate representation built upon $g_k|_{k=1}^c$ from the external videos and g_{c+1}, g_{c+2} from the positive and negative examples of the particular event to be detected:

$$\min_{f,\{g_1,...,g_{c+2}\}} \sum_{i=1}^{n+m} \ell(f([g_1(x_i), ..., g_{c+2}(x_i)]), y_i) + \alpha\Omega(f), \tag{4.5}$$

where $x_i (1 \leqslant i \leqslant m+n)$ is either a positive or negative example of a particular event, or an example of external videos used to help learn the intermediate representation. In (4.5) the classifier and the intermediate representation are jointly optimized, which explicitly guarantees that the two are correlated. Inspired by [33], we define $f(x_i)$ and $g(x_i)$ as follows:

$$f(g(x_i)) = W^T g(x_i), \tag{4.6}$$

$$g(x_i) = [\theta_1^T x_i, ..., \theta_{c+2}^T x_i] = \Theta^T x_i. \tag{4.7}$$

Then we rewrite (4.5) as

$$\min_{W,\Theta} \sum_{i=1}^{n+m} \ell\left(W^T(\Theta^T x_i), y_i\right) + \alpha\|W\|_F^2. \tag{4.8}$$

In our previous work [51], we used the $\ell_{2,1}$-norm based loss function and obtained good performance for multimedia understanding. In this extension, we apply the $\ell_{2,p}$-norm ($0 < p < 2$) based

loss function as we can adjust the value of p to search for the optimal loss. In this way, our previous work is a special case of this new formula. For an arbitrary matrix $A \in \mathbb{R}^{d\times c}$, $\|A\|_{2,p}$ is defined as:

$$\|A\|_{2,p} = \left(\sum_{i=1}^{d} (\sum_{j=1}^{c} |A_{ij}|)^{\frac{p}{2}} \right)^{\frac{1}{p}}. \tag{4.9}$$

We propose our objective function as:

$$\begin{aligned} &\min_{W,\Theta,b} \left\| X\Theta W + 1_{n+m} b^T - Y \right\|_{2,p} + \alpha \|W\|_F^2 . \\ &\text{s.t.} \Theta^T \Theta = I \end{aligned} \tag{4.10}$$

In (4.10), $X = [x_1, x_2, ..., x_n, x_{n+1}, ..., x_{n+m}] \in \mathbb{R}^{(n+m)\times d}$ is the data matrix including the positive and negative examples $x_1, x_2, ..., x_n$ of a particular event together with the external videos $x_{n+1}, ..., x_{n+m}$. $Y = [y_1, y_2, ..., y_n, y_{n+1}, ..., y_{n+m}] \in \mathbb{R}^{(n+m)\times(c+2)}$ indicate their labels. Note that the external videos have c classes and the positive and negative examples for an event are treated as two classes so we have $c+2$ classes in total. $1_{n+m} \in \mathbb{R}^{n+m}$ is a column vector with all ones and $b \in \mathbb{R}^{c+2}$ is the bias. The bias is added for unbalanced data but we can preprocess the data by centering them. The orthogonal constraint $\Theta^T\Theta = I$ is added for two considerations: 1) to avoid arbitrary scaling of the intermediate representation; 2) to preserve as much information as possible [35]. Suppose the data are centered, (4.10) becomes:

$$\begin{aligned} &\min_{W,\Theta} \|X\Theta W - Y\|_{2,p} + \alpha \|W\|_F^2 . \\ &\text{s.t.} \Theta^T \Theta = I \end{aligned} \tag{4.11}$$

Note that although (4.11) looks similar to the objective function in [33], our proposed method is different from that of [33]. The primary difference is that the motivation of [33] is to address multi-label classification whereas ours manages to learn an intermediate representation coupled with the specific loss function. When the loss function changes, the intermediate representation, *i.e.*, Θ changes accordingly. Another difference is that we use an $\ell_{2,p}$-norm based loss function which is more robust.

Next, we discuss how the proposed approach tackles the four problems below (4.4) that are faced by the traditional concepts-based representation methods. First, to obtain good concept classifiers, it usually requires a large amount of labeled training data. Our method, however, does not directly use the concept classifiers but learns an intermediate representation so not many data are required, which is also validated by our experiment. To detect the event *feeding an animal*, traditional methods would train the concept classifier of "animal." However, it is hard to know what concepts else can be useful. If the event happens indoor, concepts such as "floor" would help. If the event happens outdoor, "grass land" is more informative. It is tricky to decide what concepts should be trained in advance. Differently, our method learns an intermediate representation, which does not directly use the pre-defined concept classifiers to perform MED. As can be seen, our method jointly optimizes the loss function and the intermediate representation. In this case, the loss function is optimized for *feeding an animal*. As this learning process is coupled with the detector, it is able to adjust $g(x)$ for the event. When the event is changed, X and Y in (4.11) will also be different. Consequently, the optimal Θ will be different, which means that different intermediate representations are learned for different events. However, traditional approach uses the same concept detection results for different events, and therefore the selection of concepts turns to a critical problem for the traditional concepts-based representation. Third, traditional methods directly use the output from trained concept classifiers as

input for event detection. If the output of the pre-trained classifiers is not reliable, the performance of MED degrades. Differently, our method learns a discriminative intermediate representation, which dose not directly use the output of concept clarifiers as input. Fourth, if we use traditional pre-trained concept classifiers for event detection, we have to decide in advance what concept classifiers to use. In contrast, our method learns g and f jointly with the assumption that concept classifiers and event detector have an intermediate representation. Consequently, we do not need to select the concepts for a particular event.

4.3.2 *Solution*

The $\ell_{2,p}$-norm in our framework is non-smooth which makes (4.11) difficult to solve. To deal with this problem, we propose the following solution. By denoting $X\Theta W - Y = [z^1, ..., z^{n+m}]^T$, the objective of (4.11) is equivalent to:

$$\begin{aligned} &\min_{W,\Theta} \mathrm{Tr}\left((X\Theta W - Y)^T \tilde{D}(X\Theta W - Y)\right) + \alpha \|W\|_F^2, \\ &\text{s.t.} \quad \Theta^T\Theta = I \end{aligned} \tag{4.12}$$

where $\tilde{D}$ is a matrix with its diagonal elements $\tilde{D}_{ii} = \frac{1}{\frac{2}{p}\|z^i\|_2^{2-p}}$. By setting the derivative *w.r.t.* W to 0, we have:

$$W = A^{-1}\Theta^T X^T \tilde{D} Y, \tag{4.13}$$

where $A = \Theta^T X^T \tilde{D} X\Theta + \alpha I$ and I is an identity matrix. The above procedure needs to calculate the inverse of A. $A = \Theta^T X^T \tilde{D} X\Theta + \alpha I = (X\Theta)^T \tilde{D}(X\Theta) + \alpha I$. As D is semi-positive, $(X\Theta)^T \tilde{D}(X\Theta)$ is semi-positive. I is positive definite. Thus, A is non-singular and invertible. Substituting (4.13) into (4.12), it becomes:

$$\begin{aligned} &\min_{\Theta} \mathrm{Tr}\left(Y^T \tilde{D} X\Theta A^{-1}(\Theta^T X^T \tilde{D} X\Theta - 2A + \alpha I)A^{-1}\Theta^T X^T \tilde{D} Y\right) \\ &\text{s.t.} \quad \Theta^T\Theta = I \end{aligned} \tag{4.14}$$

As $A = \Theta^T X^T \tilde{D} X\Theta + \alpha I$, (4.14) becomes:

$$\begin{aligned} &\max_{\Theta} \mathrm{Tr}\left((\Theta^T U\Theta)^{-1}\Theta^T V\Theta\right), \\ &\text{s.t.} \quad \Theta^T\Theta = I \end{aligned} \tag{4.15}$$

where $U = X^T \tilde{D} X + \alpha I$ and $V = X^T \tilde{D} Y Y^T \tilde{D} X$.

The objective function of (4.15) can be readily solved by the eigen-decomposition of $U^{-1}V$. However, the solving of Θ requires the input of $\tilde{D}$ that is related to W, so it is not handy to get Θ and W. Therefore, we propose an iterative approach demonstrated in Algorithm 3. It can be proved that the objective function value shown in (4.11) monotonically decreases in each iteration until convergence using the iterative approach in Algorithm 3. The complexity of calculating the inverse of a few matrices is $\mathcal{O}(d^3)$. To obtain Θ, we need to conduct eigen-decomposition of $U^{-1}V$, which is also $\mathcal{O}(d^3)$ in complexity.

4.3.3 *Nonlinear SAIR*

As nonlinear classifiers generally have better performance than linear ones for event detection [108], we extend our algorithm SAIR to a nonlinear classifier by utilizing kernel tricks. Assuming that

Algorithm 3: The SAIR algorithm.

Input:

The training data X and the label matrix Y;

Parameter α.

Output:

Converged Θ and W.

1: Set $t = 0$ and initialize Θ_0, W_0 randomly;

2: **repeat**

Compute $[z_t^1, ..., z_t^{n+m}]^T = X\Theta_t W_t - Y$;

Compute the diagonal matrix $\tilde{D}_t$ as: $\tilde{D}_t = \begin{bmatrix} \frac{1}{\frac{2}{p}\|z_t^1\|_2^{2-p}} & & \\ & ... & \\ & & \frac{1}{\frac{2}{p}\|z_t^{n+m}\|_2^{2-p}} \end{bmatrix}$;

Compute $U_t = X^T \tilde{D}_t X + \alpha I$;

Compute $V_t = X^T \tilde{D}_t Y Y^T \tilde{D}_t X$;

Obtain Θ_{t+1} by the eigen-decomposition of $U_t^{-1} V_t$;

Compute $A_t = \Theta_t^T X^T \tilde{D}_t X \Theta_t + \alpha I$;

Update W_{t+1} as $W_{t+1} = A_t^{-1} \Theta_t^T X^T \tilde{D}_t Y$;

$t = t + 1$.

until *Convergence*;

3: Return Θ and W.

there is a transformation function $\phi : \mathbb{R}^d \to \mathcal{H}$. Then, the objective function of the nonlinear SAIR can be written as:

$$\begin{aligned} &\min_{W,\phi(\Theta)} \|\phi(X)\phi(\Theta)W - Y\|_{2,p} + \alpha \|W\|_F^2, \\ &\text{s.t.} \quad \phi(\Theta)^T \phi(\Theta) = I \end{aligned} \tag{4.16}$$

It has been proved in [103] that if we map the data into a Hilbert space $\mathcal{H}$ by Kernelized Principal Component Analysis (KPCA) [72], (4.16) can be solved in a similar way as (4.11) using the representations in $\mathcal{H}$.

4.4 EXPERIMENTS

In this section, we present the experimental results. We use the nonlinear SAIR with χ^2 kernel. Our method is compared to the following algorithms: AdaBoost, TaylorBoost [70], SVM, Linear Discriminant Analysis (LDA) [23] followed by ridge regression and Semantic Concept Representation (SCR). For SCR, we use the existing concept-based video corpus to learn the representation of the event-based videos. Then SVM with χ^2 kernel is applied for classification.

4.4.1 *Datasets*

We use the TRECVID MED 2011 (MED11)[3] development set in our experiments, which includes 15 events: *Attempting a board trick* (E01), *Feeding an animal* (E02), *Landing a fish* (E03), *Wedding ceremony* (E04), *Working on a woodworking project* (E05), *Birthday party* (E06), *Changing a vehicle tire* (E07), *Flash mob gathering* (E08), *Getting a vehicle unstuck* (E09), *Grooming an animal*

3 http://www.nist.gov/itl/iad/mig/med11.cfm

(E10), *Making a sandwich* (E11), *Parade* (E12), *Parkour* (E13), *Repairing an appliance* (E14) and *Working on a sewing project* (E15). We perform event detection for these 15 events.

Another two video sets, *i.e.*, the TRECVID MED 2010 (MED10)[4] and the development set from TRECVID 2011 semantic indexing task are used as external video sources. We use them to help learn the intermediate representation for MED11. MED10 includes 3 events. The video set for semantic indexing task covers 346 concepts. We used 65 concepts suggested by [20]. These concepts are related to human, environment and object. For convenience, we denote the resulting dataset as Semantic Indexing dataset(SIN11). Recall that in (4.11) $Y \in \mathbb{R}^{(n+m)\times(c+2)}$ where $c = 3 + 65 = 68$ in our setting. According to the task definition from NIST, each event is detected independently. In our experiments, there are 15 individual detection tasks.

4.4.2 *Setup*

The training data comprise three parts. The first part consists of 100 positive examples and 500 negative examples randomly selected from MED11. The second part includes 309 positive examples from MED10. The third part is SIN11 which has 2529 video frames. The remaining videos in MED11 are our testing data.

We use a 4096 dimension Bag-of-Words feature to represent each video using SIFT, CSIFT [81] and MoSIFT separately. The three feature types are further concatenated. We ran our program on the Carnegie Mellon University Parallel Data Lab cluster, which contains 300 cores, to extract features and perform the bag-of-words mapping.

The parameters of all algorithms in our experiments are tuned by a "grid-search" strategy from $\{10^{-3}, 10^{-2}, \cdots, 10^{2}, 10^{3}\}$. We use two evaluation metrics. The first one, Minimum NDC (MinNDC) [2], is defined as follows:

$$\mathrm{MinNDC}(S,E) = \frac{C_M P_M(S,E) P_T + C_{FA} P_{FA}(S,E)\,(1 - P_{FA}(S,E))}{\mathrm{MINIMUM}\,(C_M P_T, C_M(1-P_T))}, \tag{4.17}$$

where $P_M(S,E)$ is the missed detection probability for system S, event E while $P_{FA}(S,E)$ is the false alarm probability for system S, event E. $CM = 80$ is the cost for missed detection, $C_{FA} = 1$ is the cost for false alarm and $PT = 0.001$. Lower MinNDC indicates better detection performance. The second one is Average Precision (AP). Higher AP indicates better performance.

4.4.3 *MED Results*

The MED results are displayed in Table 14 using the two evaluation metrics. It can be seen that our method SAIR is consistently competitive compared with other methods. Zooming into details, we have the following observations: 1) In terms of MinNDC, SAIR gains the best performance for 9 events and the second best performance for another 5 events. SAIR outperforms all other methods for the average accuracy over all the 15 events. 2) In terms of AP, SAIR is the best method for 8 events and the second best one for the other 7 events. SAIR obtains the top performance for the average accuracy over all the 15 events. Notably, it outperforms the runner-up SVM by 8%. 3) SVM and SCR have varying degree of success for some events. However, when considering the overall performance, they are not as consistently robust as SAIR. 4) As a linear approach, LDA has weak performance. Hence, it is preferable to use kernel methods. The better performance of SAIR indicates that leveraging other concept-based and/or event-based videos is beneficial for multimedia event detection.

4 http://nist.gov/itl/iad/mig/med10.cfm

Table 14: MED performance comparison. Note that LOWER MinNDC / HIGHER AP indicates BETTER performance. The best results are highlighted in bold.

Event	Metric	AdaBoost	TaylorBoost	SVM	LDA	SCR	SAIR
E01	MinNDC	1.218	0.995	0.826	0.998	**0.742**	0.775
	AP	0.086	0.094	0.225	0.131	**0.274**	0.248
E02	MinNDC	1.343	1.001	**0.963**	1.001	0.981	0.964
	AP	0.037	0.043	0.087	0.045	0.079	**0.089**
E03	MinNDC	1.119	0.932	0.665	0.938	0.704	**0.626**
	AP	0.065	0.097	0.260	0.103	0.234	**0.281**
E04	MinNDC	1.015	1.001	0.466	1.001	0.582	**0.441**
	AP	0.084	0.067	0.483	0.073	0.322	**0.493**
E05	MinNDC	1.203	1.001	0.726	1.001	0.940	**0.711**
	AP	0.055	0.046	**0.294**	0.096	0.091	0.283
E06	MinNDC	1.211	1.001	0.885	1.001	0.939	**0.882**
	AP	0.030	0.019	**0.079**	0.021	0.051	0.076
E07	MinNDC	1.187	1.001	0.670	1.001	0.862	**0.636**
	AP	0.006	0.006	0.023	0.006	0.013	**0.030**
E08	MinNDC	1.139	1.001	0.629	1.001	**0.509**	0.568
	AP	0.050	0.042	0.198	0.059	**0.291**	0.228
E09	MinNDC	1.031	0.902	0.802	0.970	**0.586**	0.711
	AP	0.019	0.027	0.051	0.018	**0.107**	0.083
E10	MinNDC	1.317	1.001	0.856	0.925	**0.814**	0.856
	AP	0.006	0.013	0.046	0.025	**0.056**	0.047
E11	MinNDC	1.355	1.001	**0.821**	1.001	0.843	0.858
	AP	0.008	0.009	**0.034**	0.010	0.029	0.030
E12	MinNDC	1.091	0.991	0.654	1.001	0.712	**0.632**
	AP	0.035	0.028	0.093	0.019	0.083	**0.108**
E13	MinNDC	1.156	0.955	0.570	1.001	0.566	**0.449**
	AP	0.014	0.005	0.047	0.009	0.050	**0.055**
E14	MinNDC	0.971	1.001	0.550	0.822	0.664	**0.508**
	AP	0.027	0.018	0.102	0.029	0.056	**0.109**
E15	MinNDC	1.188	1.001	0.706	0.974	0.833	**0.612**
	AP	0.012	0.008	0.037	0.016	0.027	**0.054**
Average	MinNDC	1.163	0.986	0.719	0.976	0.752	**0.682**
	AP	0.035	0.035	0.137	0.044	0.118	**0.148**

4.4.4 *Performance* w.r.t. *Fewer Concepts*

To study whether the number of concepts selected affects the MED performance, we conduct an experiment by reducing the 65 concepts to 30 concepts. The video frames related to these 30 concepts in SIN11 are used to help learn the intermediate representation. We also enlist the performance variance of SCR as it also leverages the SIN dataset to obtain a concepts-based representation for MED. The first three events, *i.e.*, *Attempting a board trick*, *Feeding an animal* and *Landing a fish* are used as showcases. Table 15 displays the corresponding results. It can be seen that the performance of SAIR does not vary much when using only 30 concepts for intermediate representation. However, the performance of SCR drops drastically. For example, SCR outperforms SAIR for the event *Attempting a board trick* when using 65 concepts but SAIR beats SCR when using 30 concepts. Thus, our method SAIR is more robust to the selection of concepts-based videos compared to SCR.

Table 15: Performance comparison between using 30 concepts and using 65 concepts from SIN11.

Event	Metric	SCR(30C)	SCR(65C)	SAIR(30C)	SAIR(65C)
E01	MinNDC	0.811	0.742	0.764	0.775
	AP	0.215	0.274	0.246	0.248
E02	MinNDC	0.976	0.981	0.961	0.964
	AP	0.071	0.079	0.091	0.089
E03	MinNDC	0.722	0.704	0.625	0.626
	AP	0.214	0.234	0.286	0.281

4.4.5 *Using More Negative Examples*

We further conduct an experiment to evaluate whether negative examples contribute much to the detection accuracy by increasing the number of negative examples to 1000. Figure 13 shows the performance comparison between using 500 negative examples and 1000 negative examples. It can be seen that using 1000 negative examples is clearly better than merely using 500 negative examples, which indicates that negative examples do help improve the detection accuracy. Since negative examples are quite easy to obtain in the real world, it is reasonable and beneficial to leverage such free resources for boosted detection accuracy.

4.4.6 *Parameter Sensitivity*

In our experiments we have tuned the regularization parameter α in (4.11). Thus, we conduct an experiment to study how the parameter α in (4.11) affects the detection performance. Similarly, we use *Attempting a board trick*, *Feeding an animal*, *Landing a fish* in this experiment. Figure 14 demonstrates the performance variation *w.r.t* α. For these three events, the best results are obtained when α is small.

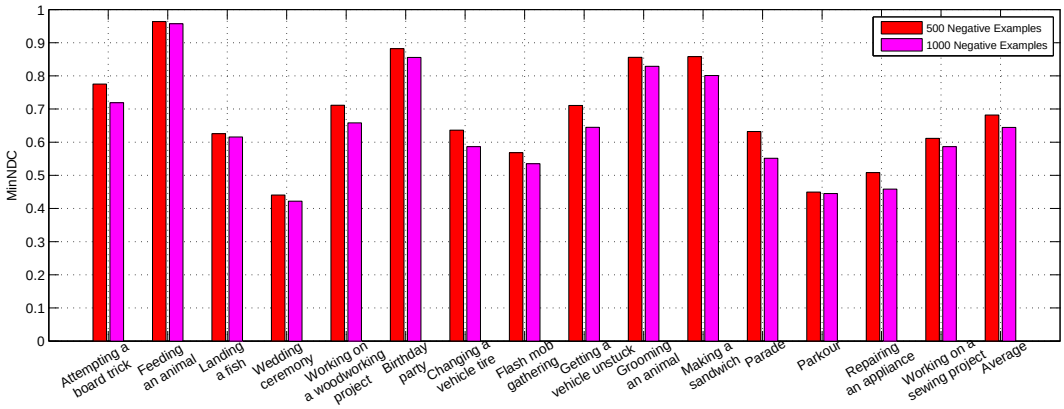
4.4.7 *Convergence*

In the previous section, we have proved that the objective function in (4.11) converges through the proposed algorithm. For practical applications it is interesting how fast our algorithm converges. In our convergence experiment we fix α at 1.

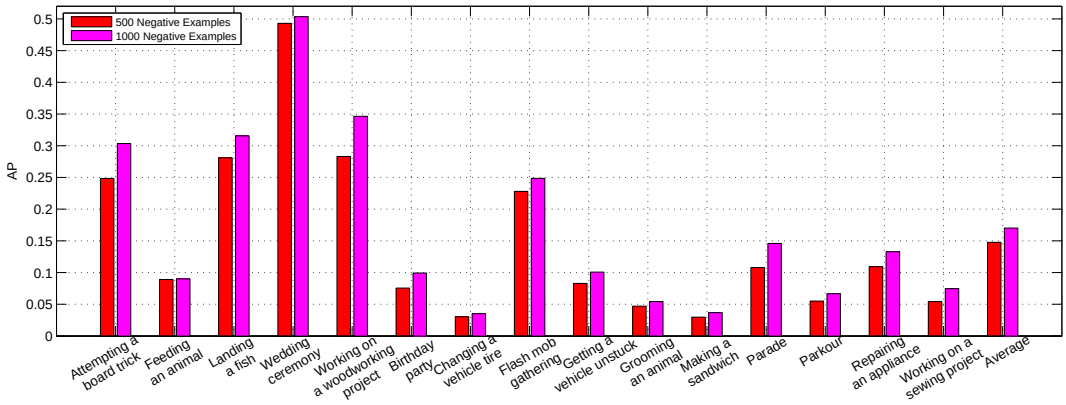
Figure 15 shows the convergence curve of our optimization algorithm. It can be seen that our algorithm converges within 10 iterations, which is efficient.

4.4.8 *Nonlinear SAIR vs Linear SAIR*

We have mentioned before that usually nonlinear classifiers obtain better performance than linear classifiers for event detection. For better performance, we have extended our algorithm SAIR to a nonlinear classifier. To understand the performance improvement from linear method to nonlinear method, we use the linear SAIR for MED. The comparison between the two approaches is displayed in Figure 16. It can be seen that nonlinear SAIR has remarkable advantage over linear SAIR in terms of MinNDC and AP. The result demonstrates that it is beneficial to implement our method as a nonlinear classifier for MED.



(a) *Performance Comparison in terms of MinNDC*

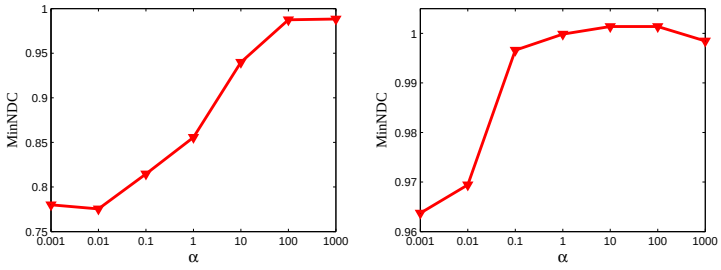


(b) *Performance Comparison in terms of AP*

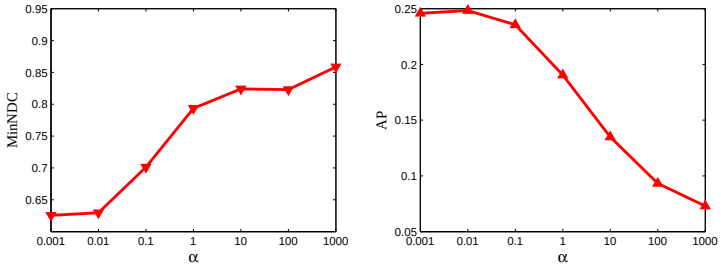
Figure 13: Performance comparison between using 500 negative examples and using 1000 negative examples. Note that LOWER MinNDC/HIGHER AP indicates BETTER performance.

4.5 CONCLUSION

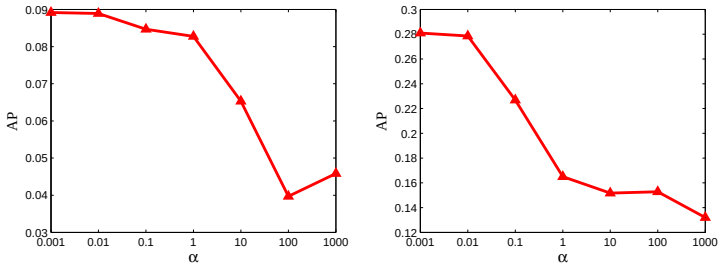
Multimedia event detection is important for video indexing and retrieval. We have proposed a new learning framework for multimedia event detection by leveraging the classifier-specific intermediate representation from low-level features. The intermediate representation of videos is automatically optimized together with the classifier. As a result, the intermediate representation is able to better reveal the video semantics and at the same time is preferable for the classifier learning. Specifically, we have used external videos in the learning process, which provide extra informative cues. The joint learning of the intermediate representation and the classifier results in a respectable framework for multimedia event detection. To validate its efficacy, we conducted several experiments using real-world video archives. The results showed that our method consistently yields competitive or better accuracy than other methods. However, it is important that the concepts from external videos for learning the intermediate representations should be related to the target event. If the concepts have little correlation with the event, we may be unable to find shared components in the subspace on which the intermediate representation is based. Consequently, little or no extra informative cues from the concepts can be incorporated into the classifier learning for event detection. Meaning: It is unlikely to witness much improvement for event detection. Another limit of our method is that it



(a) *Attempting a board trick* **(b)** *Feeding an animal*



(c) *Landing a fish* **(d)** *Attempting a board trick*



(e) *Feeding an animal* **(f)** *Landing a fish*

Figure 14: Performance variation *w.r.t.* different values of α. Note that LOWER MinNDC/HIGHER AP indicates BETTER performance.

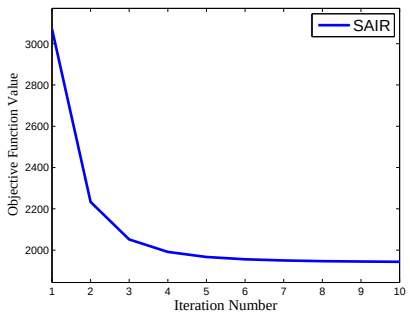
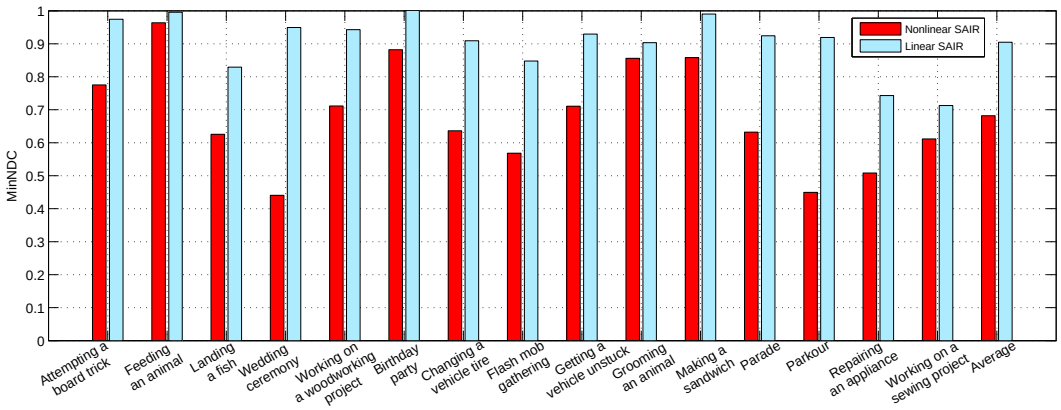
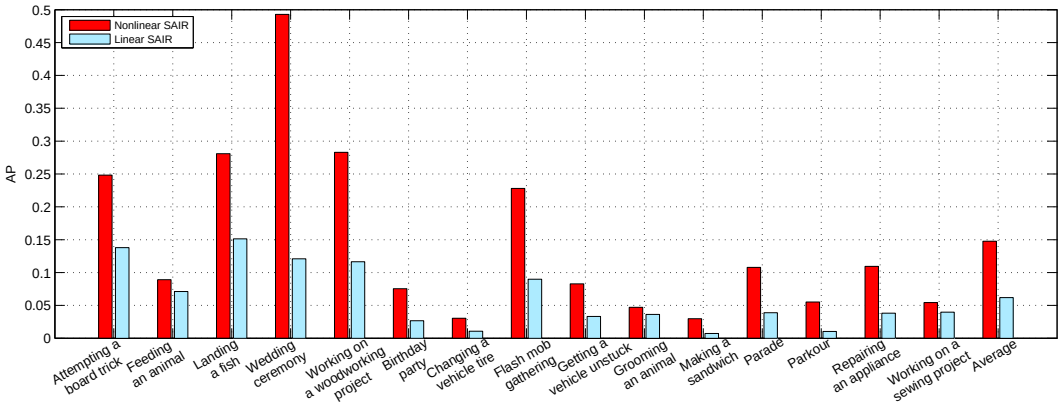


Figure 15: Convergence curve of the proposed algorithm.



(a) *Performance Comparison in terms of MinNDC*



(b) *Performance Comparison in terms of AP*

Figure 16: Performance comparison between using nonlinear SAIR and using linear SAIR. Note that LOWER MinNDC/HIGHER AP indicates BETTER performance.

is unable to help us understand the semantic meaning of an event as the intermediate representation is latent and uninformative. Therefore, our method is unsuitable for applications such as multimedia event recounting as no explicit concepts characterizing an event can be inferred by using our approach.

5

KNOWLEDGE ADAPTATION WITH PARTIALLY SHARED FEATURES FOR EVENT DETECTION USING FEW EXEMPLARS[1]

Multimedia event detection (MED) is an emerging area of research. Most related works mainly focus on sports and news event detection or abnormality detection in surveillance videos. In contrast, we focus on detecting more complicated and generic events that gain more users' interest, and we explore an effective solution for MED. Moreover, our solution only uses few positive examples since precisely labeled multimedia content is scarce in the real world. As the information from these few positive examples is limited, we propose using knowledge adaptation to facilitate event detection. Different from the state of the art, our algorithm is able to adapt knowledge from another source for MED even if the features of the source and the target are partially different, but overlapping. Avoiding the requirement that the two domains are consistent in feature types is desirable as data collection platforms change or augment their capabilities and we should be able to respond to this with little or no effort. We perform extensive experiments on real-world multimedia archives consisting of several challenging events. The results show that our approach outperforms several other state-of-the-art detection algorithms.

5.1 INTRODUCTION

With ever expanding multimedia collections, multimedia content analysis is becoming a fundamental research issue for many applications such as indexing and retrieval, *etc*. Multimedia content analysis aims to learn the semantics of multimedia data. To do so, it has to bridge the semantic gap between the low-level features and the high-level semantic content description [28] [94]. Different approaches have been proposed to bridge the semantic gap in the literature, either at concept level or event level.

We first highlight the difference between a concept and an event. A "concept" means an abstract or general idea inferred from specific instances of objects, scenes and actions such as *fish*, *outdoor* and *boxing*. Concepts are lower level descriptions of multimedia data which usually can be inferred with a single image or a few video frames. In multimedia research, a major thrust for multimedia content analysis is to learn the semantic concepts of the multimedia data and to use these concepts for multimedia indexing and retrieval. Multimedia concept analysis has been widely studied for images and videos [50] [78] [73]. However, as shared personal video collections, news videos and documentary videos have explosively proliferated these years, video event analysis is gradually attracting more research interest. An "event" refers to an observable occurrence that interests users,

1 Z. MA, Y. YANG, N. SEBE AND A. G. HAUPTMANN: "KNOWLEDGE ADAPTATION WITH PARTIALLY SHARED FEATURES FOR EVENT DETECTION USING FEW EXEMPLARS". PENDING MINOR REVISION IN *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 2013. IDEA PREVIOUSLY APPEARED IN: Z. MA, Y. YANG, Y. CAI, N. SEBE AND A. G. HAUPTMANN: "KNOWLEDGE ADAPTATION FOR AD HOC MULTIMEDIA EVENT DETECTION WITH FEW EXEMPLARS". IN *PROCEEDINGS OF THE ACM INTERNATIONAL CONFERENCE ON MULTIMEDIA*, PAGES 469-478, 2012.

e.g. celebrating the New Year. Compared with concepts, events are higher level descriptions of multimedia data. A meaningful event builds upon many concepts and is unlikely to be inferred with a single image or a few video frames. For example, the event *making a cake* consists of a combination of several concepts such as *cake*, *people*, *kitchen* together with the action *making* within a longer video sequence.

Annotation and detection are two different topics of both concept and event analysis. Multimedia annotation, also known as recognition, aims to associate a datum with one or multiple semantic labels (tags). Many approaches have been proposed to improve the annotation accuracy for both images and videos [50] [79] [59]. A typical annotation approach first pre-trains a series of classifiers, one for each class, and then applies the pre-trained classifiers to predicting the class label of each testing datum. In contrast to annotation, detection identifies the occurrence of a class of interest. The main difference between annotation and detection is that in annotation each testing datum is guaranteed to be a positive sample of one of the predefined classes while the negative examples in detection are from a set of infinite classes. In other words, both the training and testing data in annotation tasks are from a fixed number of classes but the training and testing data in detection tasks can be from an infinite number of classes. We have no clue about all the concepts or events these negative examples include. This provides very limited training information for obtaining a robust detector, thus making detection a challenging problem.

Figure 17: Some sample frames from two videos of the event *landing a fish*.

The TREC Video Retrieval Evaluation (TRECVID) community [4] has notably contributed to the research of video concept and event detection by providing a common testbed for evaluating different detection approaches [60]. In the field of multimedia, many other works have also focused on *concept detection*, e.g., [78] [91] [45]. However, the research on video *event detection* is still in its infancy. Before 2011, most existing research on event detection was limited to the events in sports [71] [90] [73] and news video archives [84], or those with repetitive patterns like *running* [83] or unusual events in surveillance videos [5] [101] [68]. In 2010, the TRECVID community launched the task of “Event detection in Internet multimedia (MED)" which aims to encourage new technologies for detecting more generic and complicated events, *e.g.*, *landing a fish*. For this kind of events, there are huge intra-class variations. Besides, they can only be characterized by long video

sequences, which necessitates the exploration of all the sequences for analysis. Figure 17 shows some frames from two videos of the same event *landing a fish*. At the first glance, we may consider Video 1 to be *skiing* as it contains the *concept* of “outdoor with snow" which is not a typical scene for *landing a fish*. The scene of Video 2 is more typical, in contrast, though it can also be a scene for *sailing*. The comparison of these two videos aims to demonstrate the huge intra-class variation of complex events. On the other hand, the information from only a few frames is patchy, as shown in Figure 17. Thus, the entire video is needed for analysis.

SVM has been used in few systems designed for the MED task and proved to be highly effective [10] [13] [108]. These systems commonly use sufficient positive examples (about 100) for reliable performance. Recently, NIST has proposed a problem of how to attain respectable detection accuracy when there are very few positive examples since precisely labeled multimedia content is scarce in the real world. In this paper, we focus on developing an effective method for MED with few exemplars. Though SVM is effective in current systems, its performance would likely be less robust when there are only a few positive examples for training. Humans often adapt knowledge obtained from previous experiences to improve learning of new tasks. Therefore, in the same manner, it is advantageous to leverage and adapt knowledge from other related domains or tasks to address the problem of an insufficient number of labeled examples. In the multimedia community, there are some available video archives with annotated concept labels, which can be leveraged to facilitate MED with few exemplars. Inspired by [91] [34] [21], we propose to adapt the knowledge from concept level to assist in our task. Specifically, we use the available video corpora with annotated concepts as our auxiliary resource and MED is performed on the target videos. The concepts are supposed to be relevant to the event to be detected.

Currently, most knowledge adaptation algorithms require that the features extracted from the raw data in the source domain and the target domain must be of exactly the same type. In many applications, such a requirement may be too restrictive, as data collection platforms change or augment their capabilities. In practice, the data in MED and those in the available concept-based video archives usually only have partially shared data features. For example, many video archives are key-frame based so they cannot be represented by audio features such as MFCC. These kinds of features are commonly used for MED and provide additional information for event detection. Hence, we propose to study how to effectively adapt knowledge from one domain to another when the available feature sets are partially different, but overlapping, for example if new or different features have more or better instrumentation for observations.

This chapter is the extension of our previous work [54]. We summarize the main contributions of this chapter are as follows:

- We perform the first exploration of MED with few exemplars by proposing a novel approach built atop knowledge adaptation.

- Unlike many knowledge adaptation methods, our approach does not require that auxiliary videos have the same events as the target videos. We exploit videos with several semantic *concepts* to facilitate the *Event* Detection on the target videos; the event differs from the concepts and the video collections are different from each other.

- Another merit is that our method is able to adapt knowledge from other sources to the target videos when only parts of the feature space are shared by the two domains. This is an intrinsic difference from most state-of-the-art knowledge adaptation algorithms.

5.2 RELATED WORK

In this section, we briefly review the related works on video event detection and knowledge adaptation.

5.2.1 *Video Event Detection*

Event detection is a challenging problem that has not been yet sufficiently studied. Based on its difficulty, event detection can be roughly categorized into simple event detection, predefined MED and Ad Hoc MED.

Simple Event Detection

Much effort has been dedicated to the detection of sports events, news events, unusual surveillance events or those with repetitive patterns. For example, Xu *et al.* propose using web-casting text and broadcast video to detect events from live sports game [90]. In [84], a model based on a multi-resolution, multi-source and multi-modal bootstrapping framework has been developed for events detection in news videos. Adam *et al.* present an algorithm using multiple local monitors which collect low-level statistics to detect certain types of unusual events in surveillance videos [5]. Wang *et al.* have proposed a new motion feature by using motion relativity and visual relatedness for event detection [83]. Their approach primarily applies to events that have repetitive motion attributes and are usually describable by a single shot, *e.g. walking* and *dancing*. The aforementioned events are usually simple, well-defined and describable by a short video sequence.

Multimedia Event Detection

In 2010, “Event detection in Internet multimedia (MED)" was initialized in the TRECVID competition by NIST for detecting more complicated events. Compared to the simple events mentioned above, the events in MED usually contain many people and/or objects, various human actions, multiple scenes and have significant intra-class variations. Additionally, these events take place in much longer and more complex video clips. For instance, *making a cake* includes objects such as water and bowl; can happen either in the kitchen or outdoor; is accompanied by specific motions such as getting the flour, adding water and baking within a longer video sequence. Though MED is an arduous problem, researchers have been making steady effort on it [10] [13] [108] [57] [93].

NIST introduced the predefined MED competition as follows: Each team is given the event kits about 5 months before the submission of the detection system. Hence, there is enough time for the system to be tailored particularly for a specific event. SVM is widely used and shows good performance for predefined MED. We may also use some recent state-of-the-art classifiers for MED. For example, a new family of boosting algorithms is proposed in [70] and demonstrates prominent performance on a variety of applications. In predefined MED, we can identify some event-specific rules or templates to facilitate detection of the particular event.

To address the generalizability of the MED system, NIST introduced Ad Hoc MED competition[2] in 2012. Ad Hoc MED differs from predefined MED in the sense that we should not tailor the system for a specific event. For this purpose, NIST releases the event kits to each team only about 12 days before the submission of the detection system. In this case, we know the testing events when we build the system but the short time period does not allow for special tuning for a specific event.

For both predefined MED and Ad Hoc MED, NIST has introduced an even more challenging problem, *i.e.*, using few labeled positive exemplars to build a detection system to deal with the

2 http://www.nist.gov/itl/iad/mig/med12.cfm

scarcity of labeled multimedia content. Our work focuses on this problem by adapting knowledge from auxiliary concept-based data. As we do not select auxiliary concepts for a particular event, our work is different from predefined MED. Moreover, the time needed for building our system satisfies the time constraint regulated by NIST. Consequently, our work gets as close as possible to Ad Hoc MED in the intended understanding of NIST.

5.2.2 *Knowledge Adaptation for Multimedia Analysis*

Knowledge adaptation, also known as transfer learning, aims to propagate the knowledge from an auxiliary domain to a target domain [91] [34] [21]. Several algorithms have been proposed but most of them require that: 1) the auxiliary domain and the target domain have the same classes; 2) the features extracted from the raw data in the source domain and the target domain must be using the exact same raw sensor output. However, MED deals with very complicated events that come from an unlimited semantic space. Furthermore, the requirement of feature consistency may be too restrictive, as data collection platforms change or augment their capabilities. Hence, most existing methods are not capable of adapting knowledge for MED when we have heterogeneous feature type between the source and the target. For example, Yang *et al.* have proposed to use Adaptive SVMs for cross-domain video concept detection [91]. The method obtained encouraging results but has some shortcomings. The proposed approach requires that the auxiliary videos and the target videos have the same video concepts. However, in MED the events are complicated and collecting many auxiliary videos with the same event description as the target videos within limited time is impractical. Jiang *et al.* [34] have used the image context of Flickr to select concept detectors. These pre-selected detectors are then refined by the semantic context transfer from the target domain. In this way, more precise concept detectors are obtained for video search. The proposed method is interesting but the selected concept detectors cannot be handily used for event detection without other sophisticated algorithms. Besides, as in our problem we only have very few positive examples, using these examples to refine the concept detectors is not reliable. Another algorithm proposed by Duan *et al.* [21] realizes event recognition of consumer videos by leveraging web videos. Their method does not require that the auxiliary domain and the target domain have the same events. However, the approach is very time consuming. Luo *et al.* have presented an object classification method by casting prior features learned from auxiliary images into their multiple kernel learning framework and obtained advantageous performance [49]. Yet this approach works in a two-step fashion, *i.e.*, training prior features using auxiliary data and then incorporating them into the following step. In contrast, our method works in a unified framework which can jointly optimize the knowledge from the auxiliary domain and the target domain. Besides those limitations mentioned above, existing knowledge adaptation algorithms mostly require that the features in the source domain and the target domain be of exactly the same type. However, in practice, this requirement may be too restrictive as MED videos can be represented by different types of features in contrast with the auxiliary video archives. Our previous work in [54] has some advantages compared to the existing knowledge adaptation algorithms such as no requirement for the same classes between the auxiliary domain and the target domain, efficiency, *etc.* But it still ignores the reality that the auxiliary domain and the target domain possibly have heterogenous feature type.

To progress beyond these aforementioned works, we propose a new knowledge adaptation method for MED with few exemplars from heterogeneous features. During the training phase, the partially shared features of the source domain and target domain will be exploited to establish a correspondence between the two domains. Meanwhile, the instrumentation obtained from the particular MED features is incorporated into our framework. The two kinds of aforementioned knowledge are then integrated to refine the detector of the target videos.

5.3 FRAMEWORK OVERVIEW

Figure 18 illustrates our framework for MED with few exemplars. The video archive where the MED is to be conducted is our target domain. The homogeneous features of the auxiliary and target videos, denoted by Modality A, are transformed to nonlinear representations based on which the shared knowledge between them is to be explored. Specifically, we perform KPCA [72] to complete the mapping. The video concept classifier and the video event detector obtained from the homogeneous features presumably have common components which contain irrelevance and noise. We propose to remove such components by optimizing the concept classifier and the event detector jointly, thereby bringing discriminating knowledge for the event detector. On the other hand, we have the heterogeneous features for MED videos and they are combined with the homogeneous features to form Modality B as indicated in Figure 18. Another event detector of MED videos is subsequently trained based on Modality B. Then we integrate the two event detectors for optimization, after which the decision values from both are fused for the final prediction.

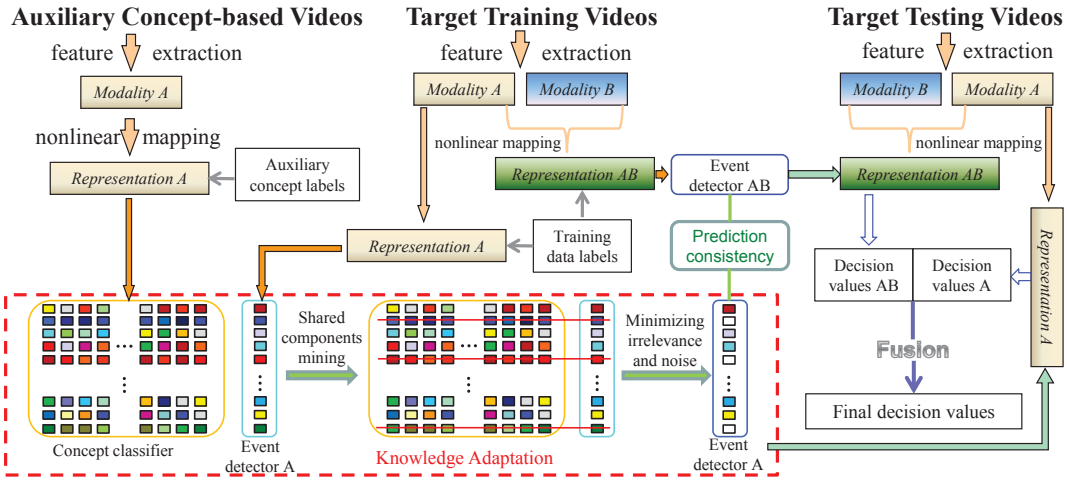


Figure 18: The illustration of our framework. We first map the homogeneous features of the auxiliary and target videos, *i.e.*, Modality A into a Hilbert Space. The video concept classifier and the video event detector obtained from the homogeneous features presumably have common components which contain irrelevance and noise. We propose to remove such negative information by optimizing the concept classifier and the event detector jointly. Meanwhile, another event detector of MED videos is trained based on Modality B. Then we integrate the two event detectors for optimization, after which the decision values from both are fused for the final prediction.

5.4 CONCEPTS ADAPTATION ASSISTED EVENT DETECTION

Next, we explain how we adapt knowledge for MED with few exemplars when the two domains have heterogeneous features. Our approach is grounded on two components: one is the knowledge from the available target training examples and the other one is the knowledge propagated from the auxiliary concepts-based videos.

We first demonstrate how to exploit the knowledge from the target training examples. Denote the nonlinear representations of the target training videos using Modality B as $\tilde{Z}_t = [\tilde{z}_t^1, \tilde{z}_t^2, ..., \tilde{z}_t^{n_t}] \in \mathbb{R}^{d_z \times n_t}$ where t stands for the target, d_z is the feature dimension and n_t is the number of the

training data. $y_t = [y_t^1, y_t^2, ..., y_t^{n_t}]^T \in \{0,1\}^{n_t \times 1}$ are the labels for the target training videos. $y_t^i = 1$ if the i^{th} video z_t^i is a positive example whereas $y_t^i = 0$ otherwise. To begin with, we associate the low-level representations and high-level semantics of videos by a decision function f which, for an input video sequence z, predicts an output y. In this paper, we define f_t as:

$$f_t(\tilde{Z}_t) = \tilde{Z}_t^T P_t + 1_t b_t, \tag{5.1}$$

where $P_t \in \mathbb{R}^{d_z \times 1}$ is an event detector which correlates $\tilde{Z}_t$ with their labels y_t, $b_t \in \mathbb{R}^1$ is a bias term and $1_t \in \mathbb{R}^{n_t \times 1}$ denotes a column vector with all ones. f_t is decided by minimizing the following objective based on the training examples Z_t and their labels y_t:

$$\min_{f_t} loss\left(f_t(Z_t), y_t\right). \tag{5.2}$$

$loss(\cdot,\cdot)$ is a loss function. Different loss functions such as the hinge loss and the least square loss can be used. In this paper, we use the $\ell_{2,1}$-norm based loss function because it is robust to outliers [52]. Thus, Eq. (5.2) is reformulated as:

$$\min_{P_t, b_t} \left\| \tilde{Z}_t^T P_t + 1_t b_t - y_t \right\|_{2,1}. \tag{5.3}$$

Now we show how to adapt the knowledge from auxiliary videos which are associated with different concepts and are represented only by the homogeneous features, *i.e.*, Modality A to assist in MED with few exemplars. Denote the nonlinear representations of the auxiliary videos as $\tilde{X}_a = [\tilde{x}_a^1, \tilde{x}_a^2, ..., \tilde{x}_a^{n_a}] \in \mathbb{R}^{d_h \times n_a}$ where a stands for the auxiliary domain, d_h is the feature dimension and n_a is the number of the auxiliary videos. $Y_a = [y_a^1, y_a^2, ..., y_a^{n_a}]^T \in \{0,1\}^{n_a \times c_a}$ is their label matrix where c_a indicates that there are c_a different concepts. Y_a^{ij} denotes the j^{th} class of y_a^i and $Y_a^{ij} = 1$ if x_a^i belongs to the j^{th} concept, while $Y_a^{ij} = 0$ otherwise. The fundamental step is to mine the correlation between the low-level representations and high-level semantics of the auxiliary concepts-based videos. Similarly to Eq. (5.3), we realize that by the following objective function:

$$\min_{W_a, b_a} \left\| \tilde{X}_a^T W_a + 1_a b_a - Y_a \right\|_{2,1} \tag{5.4}$$

where a concept classifier $W_a \in \mathbb{R}^{d_h \times c_a}$ is used to correlate $\tilde{X}_a$ with their labels Y_a, $b_a \in \mathbb{R}^{1 \times c_a}$ is a bias term and $1_a \in \mathbb{R}^{n_a \times 1}$ is a column vector with all ones.

Next, we illustrate how to adapt knowledge from the auxiliary concepts-based videos for a more discriminating event detector. To begin with, we also use Modality A for the target videos in accordance with the auxiliary videos. Denote the corresponding nonlinear representations as $\tilde{X}_t = [\tilde{x}_t^1, \tilde{x}_t^2, ..., \tilde{x}_t^{n_t}] \in \mathbb{R}^{d_h \times n_t}$. We can similarly find an event detector W_t based on $\tilde{X}_t$. $W_t \in \mathbb{R}^{d_h \times 1}$ is used to correlate $\tilde{X}_t$ with their labels y_t.

Considering each domain separately, it is reasonable to assume that for classification purposes some noisy and irrelevant features will not be used, which in turn makes the corresponding rows of the projection matrix W_a or W_t identically equal to zero. Considering the two domains together, the auxiliary concept videos and the event videos can be correlated in the semantic level, *e.g.*, the concepts *fish*, *water*, *people* are basic elements of the event *landing a fish*. Previous work on multi-task learning has suggested that this kind of correlation usually results in common components in the feature level shared across related tasks [9] [66] [92]. In our scenario, the semantically related auxiliary videos and event videos can be treated as related tasks because the events build upon the related concepts. When we represent videos from both domains with the same type of feature such as SIFT Bag-of-Words using the same centroid, they would have some shared components. For

example, assuming that the event video *landing a fish* has SIFT Bag-of-Words of *fish*, we may find similar SIFT Bag-of-Words in an image of *fish*. Hence, some shared components in the features between them need to be uncovered. Note that the event detector is actually a mapping function from features to event labels. Intuitively, not all the Bag-of-words are related to semantic labels. Given certain Bag-of-Words, if they are irrelevant to all the concepts, it is very likely that these Bag-of-Words are also irrelevant to the events, because the event is built on top of the concepts. Recalling that the corresponding rows of W_a or W_t are identically equal to zero for the irrelevant or noisy features, we should be able to find similar patterns in the distribution of these rows by learning W_a and W_t jointly. Thus, we exploit the concept classifier W_a to help remove the noise in W_t for a more discriminative event detector.

Denote $W_a = [w_a^1, ..., w_a^{d^h}]^T$, $W_t = [w_t^1, ..., w_t^{d^h}]^T$. Then we combine them and define a joint analyzer $W = [w^1, ..., w^{d^h}]^T$ where w^i is the vertical concatenation of w_a^i and w_t^i, *i.e.*, $w^i = [w_a^i; w_t^i]$. In this sense, w^i reflects the joint information from the auxiliary videos and the target training videos. Through proper optimization of w_i, we can remove the shared irrelevant or noisy components. Previous work has shown that sparse models are useful for feature selection by eliminating redundancy and noise [9] [55] [52]. The sparse models are used to make some of the feature coefficients shrink to zeros to achieve feature selection. The "shrinking to zero" idea can be applied to uncover the common distribution of the "identically equal to zero" rows of W_a and W_t discussed before. In this way, we can remove the shared irrelevance and noise, thus obtaining a more discriminative W_t.

Now we introduce the technical details of our joint sparsity model. Specifically, we propose to exploit $\|W\|_{2,p} = \left(\sum_{i=1}^{d_h} (\sum_{j=1}^{c_a+1} |W_{ij}|)^{\frac{p}{2}} \right)^{\frac{1}{p}}$ to achieve that goal. $\|\cdot\|_{2,p}$ denotes the $\ell_{2,p}$-norm ($0 < p < 2$). By minimizing $\|W\|_{2,p}^p$, we can reduce the negative impact of the irrelevant or noisy w_i's. Our model has the flexibility of characterizing different degree of relevance between concepts and events. p is used to control the degree of shared structures. The lower p is, the more semantically correlated are the auxiliary concepts and the target event. By contrast, when the auxiliary concepts and the target event have less relevance, we can use a larger p. When we increase p to 2, we do not impose sharing on the two domains. To step further, it is expected that the predicted labels of W_t on $\tilde{X}_t$ be consistent with those of P_t on $\tilde{Z}_t$, thus resulting in more accurate P_t and W_t. In this way, P_t from the heterogeneous features of the target and W_t from the knowledge adaptation would jointly augment the observations for MED. We achieve this by minimizing $\left\| \tilde{X}_t^T W_t - \tilde{Z}_t^T P_t \right\|_F^2$ where $\|\cdot\|_F^2$ indicates the Frobenius norm of a matrix.

To this end, we propose the following objective function for MED with few exemplars:

$$\begin{aligned} \min_{P_t, W_t, W_a, b_t, b_a} & \left\| \tilde{Z}_t^T P_t + 1_t b_t - y_t \right\|_{2,1} + \left\| \tilde{X}_t^T W_t - \tilde{Z}_t^T P_t \right\|_F^2 \\ & + \left\| \tilde{X}_a^T W_a + 1_a b_a - Y_a \right\|_{2,1} + \alpha \|W\|_{2,p}^p + \beta (\|W_t\|_F^2 + \|W_a\|_F^2) \end{aligned} \tag{5.5}$$

where $(\|W_a\|_F^2 + \|W_t\|_F^2)$ is added to avoid over-fitting. α and β are regularization parameters.

Once P_t and W_t are obtained, we apply them to the nonlinear representations of the testing videos for event detection. The decision values of them are normalized and then their weighted sum based on the feature numbers are the final decision values of the testing videos. Our method builds upon 1) the knowledge adaptation from concepts-based videos to event-based videos by leveraging the shared structures between them; and 2) the augmented observation from the particular features that are only owned by MED videos. We therefore name our method Heterogenous Features based Structural Adaptive Regression (HF-SAR).

5.5 OPTIMIZING THE EVENT DETECTOR

In this section, we present our solution for obtaining the target event detector. Our problem in Eq. (5.5) involves the $\ell_{2,1}$-norm and the $\ell_{2,p}$-norm which are both non-smooth and cannot be solved in a closed form. We propose to solve it as follows.

Denote $\tilde{Z}_t^T P_t - y_t = [u^1, ..., u^{n_t}]^T$, $\tilde{X}_a^T W_a - Y_a = [v^1, ..., v^{n_a}]^T$. Next, we define three diagonal matrices D_t, D_a and D with their diagonal elements $D_t^{ii} = \frac{1}{2\|u^i\|_2}$, $D_a^{ii} = \frac{1}{2\|v^i\|_2}$, $D^{ii} = \frac{1}{\frac{2}{p}\|w^i\|_2^{2-p}}$ respectively. The objective in Eq. (5.5) is equivalent to:

$$\begin{aligned}
&\min_{P_t, W_t, W_a, b_t, b_a} \mathrm{Tr}\left((\tilde{Z}_t^T P_t + 1_t b_t - y_t)^T D_t (\tilde{Z}_t^T P_t + 1_t b_t - y_t)\right) \\
&+ \left\|\tilde{X}_t^T W_t - \tilde{Z}_t^T P_t\right\|_F^2 + \mathrm{Tr}\left((\tilde{X}_a^T W_a + 1_a b_a - Y_a)^T D_a (\tilde{X}_a^T W_a + 1_a b_a - Y_a)\right) \\
&+ \alpha \mathrm{Tr}\left(W^T D W\right) + \beta(\|W_a\|_F^2 + \|W_t\|_F^2)
\end{aligned} \tag{5.6}$$

where $\mathrm{Tr}(\cdot)$ denotes the trace operator. By setting the derivative of Eq. (5.6) *w.r.t.* b_a to zero, we get:

$$b_a = \frac{1}{n_a} 1_a^T Y_a - \frac{1}{n_a} 1_a^T \tilde{X}_a^T W_a. \tag{5.7}$$

Similarly, we obtain b_t as:

$$b_t = \frac{1}{n_t} 1_t^T y_t - \frac{1}{n_t} 1_t^T \tilde{Z}_t^T P_t. \tag{5.8}$$

Substituting Eq. (5.7) and Eq. (5.8) into Eq. (5.6), it becomes:

$$\begin{aligned}
&\min_{P_t, W_t, W_a} \mathrm{Tr}\left((H_t \tilde{Z}_t^T P_t - H_t y_t)^T D_t (H_t \tilde{Z}_t^T P_t - H_t y_t)\right) \\
&+ \left\|\tilde{X}_t^T W_t - \tilde{Z}_t^T P_t\right\|_F^2 + \mathrm{Tr}\left((H_a \tilde{X}_a^T W_a - H_a Y_a)^T D_a (H_a \tilde{X}_a^T W_a - H_a Y_a)\right) \\
&+ \alpha \mathrm{Tr}\left(W^T D W\right) + \beta(\|W_a\|_F^2 + \|W_t\|_F^2)
\end{aligned} \tag{5.9}$$

where $H_t = I_t - \frac{1}{n_t} 1_t 1_t^T$, $H_a = I_a - \frac{1}{n_a} 1_a 1_a^T$ and $I_t \in \mathbb{R}^{n_t \times n_t}$, $I_a \in \mathbb{R}^{n_a \times n_a}$ are two identity matrices. Setting the derivative of Eq. (5.9) *w.r.t.* W_a to zero, we get:

$$W_a = (\tilde{X}_a H_a D_a H_a \tilde{X}_a^T + \alpha D + \beta I_d)^{-1} \tilde{X}_a H_a D_a H_a Y_a \tag{5.10}$$

where $I_d \in \mathbb{R}^{d_h \times d_h}$ is an identity matrix. Note that D is treated as a constant in this step as we adopt an alternating optimization approach here. In the same manner, we obtain the event detector W_t as:

$$W_t = A^{-1} \tilde{X}_t \tilde{Z}_t P_t \tag{5.11}$$

where $A = \alpha D + \beta I_d + \tilde{X}_t \tilde{X}_t^T$.

To optimize P_t, the problem equals to:

$$\begin{aligned}
&\min_{P_t} \mathrm{Tr}(P_t^T \tilde{Z}_t H_t D_t H_t Z_t^T P_t - 2 P_t^T \tilde{Z}_t H_t D_t H_t y_t) + \left\|\tilde{X}_t^T W_t - Z_t^T P_t\right\|_F^2 \\
&+ \alpha \mathrm{Tr}(W_t^T D W_t) + \beta \mathrm{Tr}(W_t^T W_t)
\end{aligned} \tag{5.12}$$

Substituting Eq. (5.11) into Eq. (5.12) and defining

$$J = \tilde{Z}_t H_t D_t H_t Z_t^T - \tilde{Z}_t \tilde{X}_t^T A^{-1} \tilde{X}_t \tilde{Z}_t^T + \tilde{Z}_t \tilde{Z}_t^T \tag{5.13}$$

$$K = 2\tilde{Z}_t H_t D_t H_t y_t, \tag{5.14}$$

the problem becomes:

$$\min_{P_t} Tr(P_t^T J P_t - P_t^T K) \tag{5.15}$$

By setting the derivative of the above function *w.r.t.* P_t to zero, we get:

$$P_t = \frac{1}{2} J^{-1} K \tag{5.16}$$

Algorithm 4: Optimizing the event detector.

Input:

The target training data $\tilde{Z}_t \in \mathbb{R}^{d_z \times n_t}$, $\tilde{X}_t \in \mathbb{R}^{d_h \times n_t}$, $y_t \in \mathbb{R}^{n_t \times 1}$;

The auxiliary data $\tilde{X}_a \in \mathbb{R}^{d_h \times n_a}$, $Y_a \in \mathbb{R}^{n_a \times c_a}$;

Parameters α, β and p.

Output:

Optimized $P_t \in \mathbb{R}^{d_z \times 1}$, $W_t \in \mathbb{R}^{d_h \times 1}$ and $b_t \in \mathbb{R}^1$.

1: Set $t = 0$, initialize $P_t \in \mathbb{R}^{d_z \times 1}$, $W_t \in \mathbb{R}^{d_h \times 1}$ and $W_a \in \mathbb{R}^{d_h \times c_a}$ randomly;

2: **repeat**

Compute $\tilde{Z}_t^T P_t - y_t = [u^1, ..., u^{n_t}]^T$, $\tilde{X}_a^T W_a - Y_a = [v^1, ..., v^{n_a}]^T$ and $W = [w^1, ..., w^d]^T$;

Compute the diagonal matrix D_t^t, D_a^t and D^t according to $D_t^{ii} = \frac{1}{2\|u^i\|_2}$, $D_a^{ii} = \frac{1}{2\|v^i\|_2}$, and $D^{ii} = \frac{1}{\frac{2}{p}\|w^i\|_2^{2-p}}$ respectively;

Update W_a^{t+1} as: $W_a^{t+1} = (\tilde{X}_a H_a D_a^t \tilde{X}_a^T + \alpha D^t + \beta I_d)^{-1} \tilde{X}_a H_a D_a H_a Y_a^T$;

Update b_a^{t+1} as: $b_a^{t+1} = \frac{1}{n_a} 1_a^T Y_a - \frac{1}{n_a} 1_a^T \tilde{X}_a^T W_a^{t+1}$;

Update P_t^{t+1} according to Eq. (5.13), Eq. (5.14) and Eq. (5.16);

Update W_t^{t+1} as: $W_t^{t+1} = A^{-1} \tilde{X}_t \tilde{Z}_t^T P_t^{t+1}$;

Update b_t^{t+1} as: $b_t^{t+1} = \frac{1}{n_t} 1_t^T y_t - \frac{1}{n_t} 1_t^T \tilde{Z}_t^T W_t^{t+1}$;

$t = t + 1$.

until *Convergence*;

3: Return P_t, W_t and b_t.

Next, we propose Algorithm 4 to solve the objective function in Eq. (5.5). The computational complexity of Algorithm 4 is as follows. For training, it is $\mathcal{O}(d_z^3)$ as $d_z > d_h$. Note that $d_z \gg n_t$ because usually there are few training examples in Ad Hoc MED. Thus, the training process is not very computationally expensive. During testing, computing kernels between the testing data and

the training data is the most expensive process. Suppose there are n_{te} testing videos, we need to compute $n_t n_{te}$ kernels. Each datum is d_z dimensional so the complexity is $\mathcal{O}(d_z n_t n_{te})$.

It can be proved by the following Theorem that the objective function value of Eq. (5.5) monotonically decreases in each iteration converging to local optimum using Algorithm 4.

5.6 EXPERIMENTS

In this section, we present the experiments which evaluate the performance of our Heterogenous Features based Structural Adaptive Regression (HF-SAR) for MED with few exemplars.

5.6.1 *Datasets*

NIST has provided so far the largest video corpora for MED. Our experiments on MED with few exemplars are conducted on the TRECVID MED 2010 (MED10) and TRECVID MED 2011 (MED11) development set. MED10[3] includes 3 events defined by NIST, which are *Making a cake*, *Batting a run*, and *Assembling a shelter*. MED11[4] includes 15 events, *i.e.*, *Attempting a board trick*, *Feeding an animal*, *Landing a fish*, *Wedding ceremony*, *Working on a woodworking project*, *Birthday party*, *Changing a vehicle tire*, *Flash mob gathering*, *Getting a vehicle unstuck*, *Grooming an animal*, *Making a sandwich*, *Parade*, *Parkour*, *Repairing an appliance* and *Working on a sewing project*. The two datasets are combined together (MED10-11 for short) in our experiments so we have a dataset of 9746 video clips.

We first use the development set from TRECVID 2012 semantic indexing task (SIN12) as the auxiliary videos. SIN12 covers 346 concepts but some of them have few positive examples. Additionally, “events" usually refer to “semantically meaningful human activities, taking place within a selected environment and containing a number of necessary objects" [42]. Hence, we removed the concepts with few positive examples and selected 65 concepts that are related to human, environment and objects. We thus use a subset with 3244 video frames. On the other hand, multimedia events are usually accompanied by human actions, which suggests that we may find similar motion features between event videos and basic human action videos. Hence, we additionally use UCF50 dataset [67] to test whether it is able to facilitate multimedia event detection.

We ran our program on the Carnegie Mellon University Parallel Data Lab cluster, which contains 300 cores, to extract features and perform the Bag-of-Words mapping for all the videos. When utilizing SIN12 dataset, we extract SIFT [47] and CSIFT [81] features for the videos in MED10-11 and SIN12. Then we use 1x1, 2x2 and 3x1 spatial grids to generate the spatial BoW representation [58]. For each grid, we use a standard BoW representation with 4,096 dimensions, thus resulting in a 32,768 dimension spatial BoW feature for SIFT/CSIFT to represent each video. When utilizing UCF50 dataset, we extract STIP [39] feature for the videos in MED10-11 and UCF50 since STIP has proved to be robust for analyzing action videos. A similar procedure is followed to generate the spatial BoW representation. Apart from visual features, some other features, which provide different yet complementary information, can also be used to represent videos. For example, auditory features based on Mel-frequency Cepstral Coefficients (MFCC) have also been frequently used [108]. We additionally use this feature for MED videos and the dimension is 4096. Thus, when using the SIN12 dataset, our two domains have SIFT and CSIFT as shared feature type while MFCC works as the heterogeneous feature for MED videos; when using UCF50 dataset, our two domains have STIP as shared feature type while MFCC is the heterogeneous feature for MED videos.

3 http://nist.gov/itl/iad/mig/med10.cfm
4 http://www.nist.gov/itl/iad/mig/med11.cfm

According to the MED task definition from NIST, each event is detected independently. Therefore, there are 18 individual detection tasks. NIST has defined that the number of positive training examples is 10 for MED with few exemplars [3]. However, there is no standard training and testing set partition provided by NIST. Hence, we randomly split the MED10-11 dataset into two subsets, one as the training set and the other one as the testing set. We follow the definition given by NIST and randomly select 10 positive examples for each event. Other 1000 negative examples are selected and combined with the positive examples as the training data. The remaining 8736 videos are our testing data. The experiments are independently repeated 5 times with randomly selected positive and negative examples. The average results are reported.

We use three evaluation metrics. The first one, Minimum NDC (MinNDC), is officially used by NIST in TRECVID MED 2011 evaluation [2]. Lower MinNDC indicates better detection performance. The second one is the Probability of Miss-Detection based on the Detection Threshold 12.5. This evaluation metric is used by NIST in TRECVID MED 2012 [3] to evaluate MED performance. We denote it as Pmd@TER=12.5 for short. Likewise, lower Pmd@TER=12.5 indicates better performance. For more details about the above two evaluation metrics, please see the TRECVID 2011 and 2012 evaluation plans [2] [3]. The third one is Average Precision (AP). Higher AP indicates better performance.

5.6.2 *Comparison Algorithms*

In this section, we show the MED results using Heterogenous Features based Structural Adaptive Regression (HF-SAR) and other state-of-the-art algorithms. A brief introduction of the comparison algorithms is as follows:

- HF-SAR: the proposed new method which is designed for knowledge adaptation based on heterogeneous features. The χ^2 kernel is used for its advantageous performance on video analysis.
- Structural Adaptive Regression (SAR) [54]: our previous algorithm on knowledge adaptation for MED with few exemplars. Similarly, the χ^2 kernel is used.
- Adaptive Multiple Kernel Learning (A-MKL) [21]: a recent knowledge adaptation algorithm built upon SVM.
- Multiple Kernel Transfer Learning (MKTL) [49]: a recent multi-class transfer learning algorithm built within a multiple kernel learning framework. The original algorithm in [49] has used RBF kernel. For fair comparison, we implement it with χ^2 kernel.
- SAR&SVM: We use SAR based on SIFT+CSIFT features between the auxiliary domain and the target domain. In addition, we use SVM based on MFCC feature in the target domain. Then we fuse the decision values obtained by both of them. In this way, we can evaluate the performance of combining homogeneous transfer learning and the classifier on the heterogeneous feature.
- SVM: the most widely used and robust event detector for MED [108] [10] [28] [83]. Similarly, we use the χ^2 kernel for it.
- TaylorBoost [70]: a state-of-the-art classifier extended from AdaBoost.

For SVM, we use LIBSVM, and for A-MKL, MKTL and TaylorBoost we use the code shared by the authors. During the training and predicting, we combine SIFT, CSIFT and MFCC features of the MED10-11 dataset for SVM and TaylorBoost. SAR, A-MKL and MKTL are knowledge adaptation

Table 16: Average detection accuracy of different methods. Better results are highlighted in bold.

Metric	SAR	A-MKL	MKTL	SAR&SVM	SVM	TaylorBoost	HF-SAR
MinNDC	0.860	0.881	0.873	0.841	0.850	0.902	**0.817**
Pmd@12.5	0.601	0.617	0.610	0.572	0.575	0.677	**0.549**
AP	0.162	0.144	0.153	0.183	0.181	0.080	**0.201**

based algorithms, which utilize the SIN12 dataset as auxiliary data. However, they require that the target domain and the auxiliary domain have the homogeneous feature representation so only SIFT and CSIFT are used for them. HF-SAR leverages SIN12 for MED with few exemplars on MED10-11 and it is capable of using SIFT, CSIFT and MFCC together. All the regularization parameters are tuned from $\{0.001, 0.1, 10, 1000\}$, and the parameter p of HF-SAR and SAR is tuned from $\{0.5, 1, 1.5\}$. We report the best results for each algorithm.

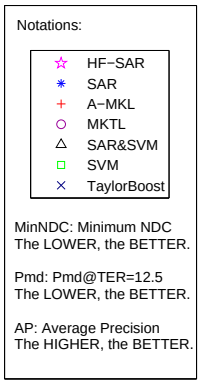
5.6.3 *MED Results*

The detection performance of different algorithms is displayed in Figure 19, Figure 20, Figure 21 and Table 16 where all the knowledge adaptation methods have exploited SIN12 dataset. Note that LOWER MinNDC and Pmd@TER=12.5 indicate BETTER performance; HIGHER AP indicates BETTER performance. The proposed method HF-SAR is consistently competitive for all the events. Zooming into details, we have the following observations: 1) when using MinNDC as metric, HF-SAR gains the best performance for 17 events; 2) when using Pmd@TER=12.5 as metric, HF-SAR gains the best performance for 15 events;2) when using AP as metric, HF-SAR is the best method for 14 events; 3) HF-SAR obtains the top performance for the average accuracy over all the 18 events; 4) SAR&SVM is generally the second competitive algorithm. This indicates that incorporating the additional information contained in the heterogenous feature into a robust knowledge adaptation algorithm based on homogeneous features is beneficial. However, it is unclear which algorithms should be combined together for the best performance as they may work with different mechanisms; 5) SAR, A-MKL and SVM have varying degrees of success on some events. However, they are generally worse than HF-SAR and SAR&SVM. It means knowledge adaptation based on homogeneous features loses useful information from the heterogenous feature, and SVM utilizes all the features but it cannot leverage knowledge from other sources. In contrast, the newly proposed method HF-SAR transfer knowledge between homogeneous features while simultaneously exploits the heterogeneous feature to get boosted performance.

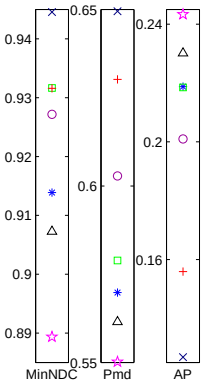
Next we show the detection results by exploiting UCF50 dataset. As HF-SAR has already shown its advantage over other knowledge adaptation algorithms and this experiments aims to show that we can even adapt useful action knowledge for Ad Hoc MED, we only compare HF-SAR to the best baseline classifier SVM. Similarly, we combine STIP and MFCC features of the MED10-11 dataset for SVM. The detailed results are illustrated in Table 17. As can be seen, HF-SAR beats SVM on 17, 17, 15 events and the average performance over all the 18 events in terms of MinNDC, Pmd@TER=12.5, AP respectively. Moreover, for those events on which HF-SAR is better, we can observe noticeable performance improvement.

5.6.4 *Influence of Knowledge Adaptation*

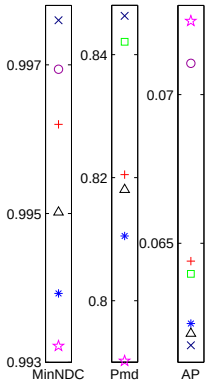
It is interesting to understand how the knowledge adaptation from the auxiliary concept-based videos impacts the Ad Hoc MED. We base our study on two scenarios: First, we set α in Eq (5.5) to 0 so



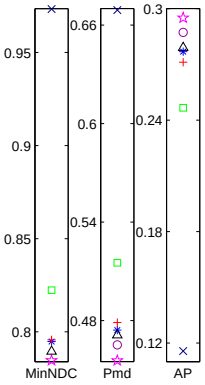
(a) Notations



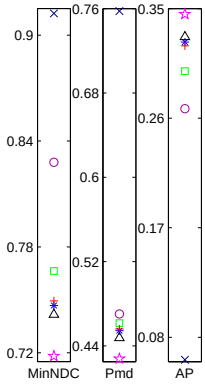
(b) *Attempting a board trick*



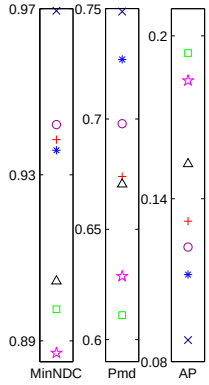
(c) *Feeding an animal*



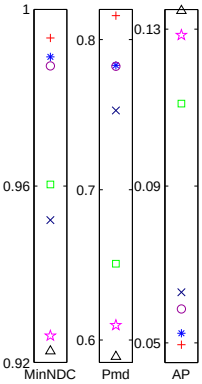
(d) *Landing a fish*



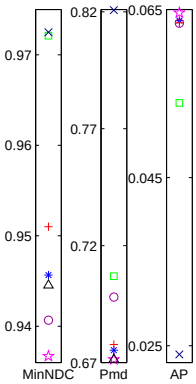
(e) *Wedding ceremony*



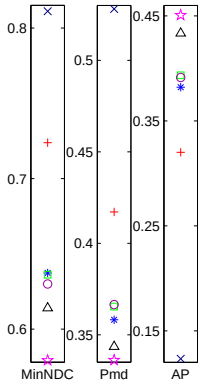
(f) *Working on a wood-working project*



(g) *Birthday party*

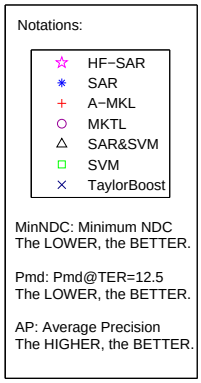


(h) *Changing a vehicle tire*

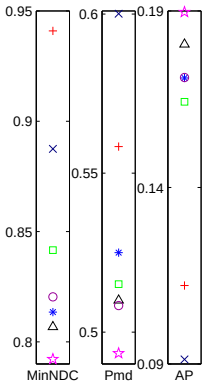


(i) *Flash mob gathering*

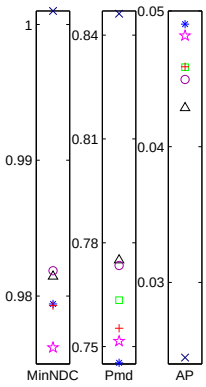
Figure 19: Performance Comparison on MED with few exemplars.



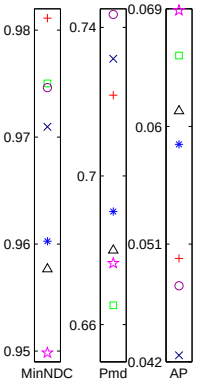
(a) Notations



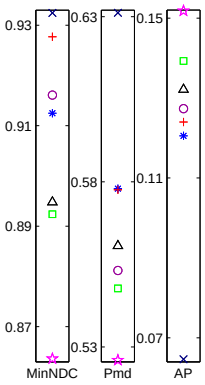
(b) *Getting a vehicle unstuck*



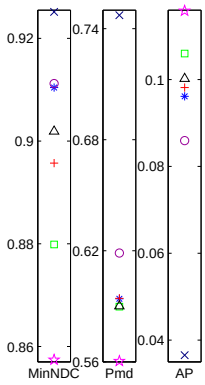
(c) *Grooming an animal*



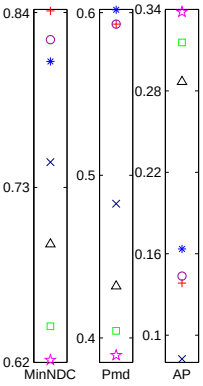
(d) *Making a sandwich*



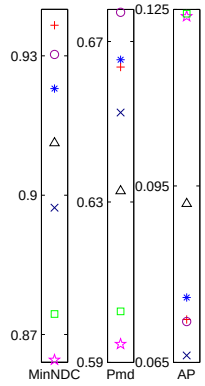
(e) *Parade*



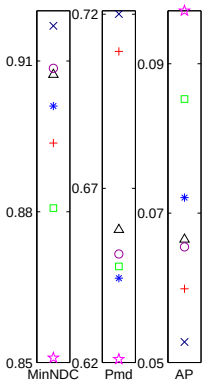
(f) *Parkour*



(g) *Repairing an appliance*



(h) *Working on a sewing project*

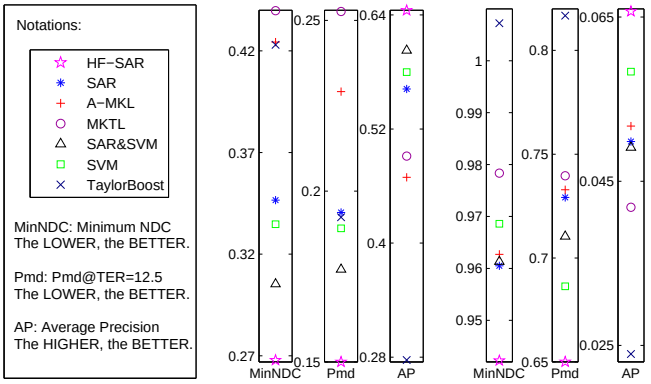


(i) *Making a cake*

Figure 20: Performance Comparison on MED with few exemplars.

Table 17: Detection results by exploiting UCF50 dataset in comparison with SVM.

Event	Metric	SVM	HF-SAR	Relative Improvement
E01	MinNDC	0.884	0.922	N/A
	Pmd@TER=12.5	0.546	0.569	N/A
	AP	0.247	0.206	N/A
E02	MinNDC	1.000	0.999	0.1%
	Pmd@TER=12.5	0.938	0.877	7.0%
	AP	0.037	0.046	24.3%
E03	MinNDC	1.000	0.990	1.0%
	Pmd@TER=12.5	0.928	0.815	13.9%
	AP	0.035	0.061	74.3%
E04	MinNDC	0.936	0.912	2.6%
	Pmd@TER=12.5	0.870	0.770	13.0%
	AP	0.044	0.132	200%
E05	MinNDC	0.975	0.946	3.1%
	Pmd@TER=12.5	0.914	0.817	11.9%
	AP	0.061	0.097	59.0%
E06	MinNDC	0.992	0.973	2.0%
	Pmd@TER=12.5	0.917	0.797	15.1%
	AP	0.049	0.077	57.1%
E07	MinNDC	1.000	0.992	0.8%
	Pmd@TER=12.5	0.944	0.881	7.2%
	AP	0.033	0.032	N/A
E08	MinNDC	0.945	0.833	13.4%
	Pmd@TER=12.5	0.862	0.676	27.5%
	AP	0.094	0.173	84.0%
E09	MinNDC	0.970	0.928	4.5%
	Pmd@TER=12.5	0.804	0.703	14.4%
	AP	0.072	0.093	29.2%
E10	MinNDC	0.997	0.991	0.6%
	Pmd@TER=12.5	0.933	0.862	8.2%
	AP	0.035	0.043	22.9%
E11	MinNDC	0.995	0.982	1.3%
	Pmd@TER=12.5	0.904	0.835	8.3%
	AP	0.037	0.041	10.8%
E12	MinNDC	0.975	0.940	9.4%
	Pmd@TER=12.5	0.889	0.770	4.5%
	AP	0.052	0.077	13.7%
E13	MinNDC	0.970	0.957	3.7%
	Pmd@TER=12.5	0.711	0.689	3.2%
	AP	0.094	0.078	N/A
E14	MinNDC	0.919	0.819	12.2%
	Pmd@TER=12.5	0.840	0.687	22.3%
	AP	0.083	0.191	130.1%
E15	MinNDC	0.964	0.945	2.0%
	Pmd@TER=12.5	0.880	0.794	10.8%
	AP	0.059	0.066	11.9%
E16	MinNDC	0.975	0.937	4.1%
	Pmd@TER=12.5	0.864	0.796	8.5%
	AP	0.045	0.053	17.8%
E17	MinNDC	0.893	0.736	21.3%
	Pmd@TER=12.5	0.766	0.585	30.9%
	AP	0.125	0.253	102.4%
E18	MinNDC	0.982	0.967	1.6%
	Pmd@TER=12.5	0.922	0.836	10.3%
	AP	0.036	0.041	13.9%
Average	MinNDC	0.965	0.932	3.5%
	Pmd@TER=12.5	0.857	0.764	12.2%
	AP	0.069	0.098	42.0%



(a) Notations **(b)** *Batting a run* **(c)** *Assembling a shelter*

Figure 21: Performance Comparison on MED with few exemplars.

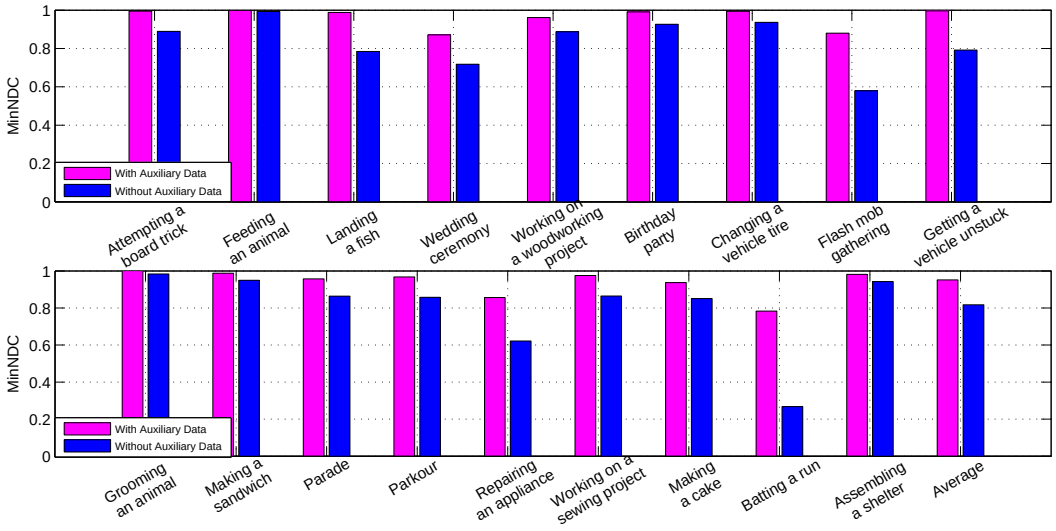
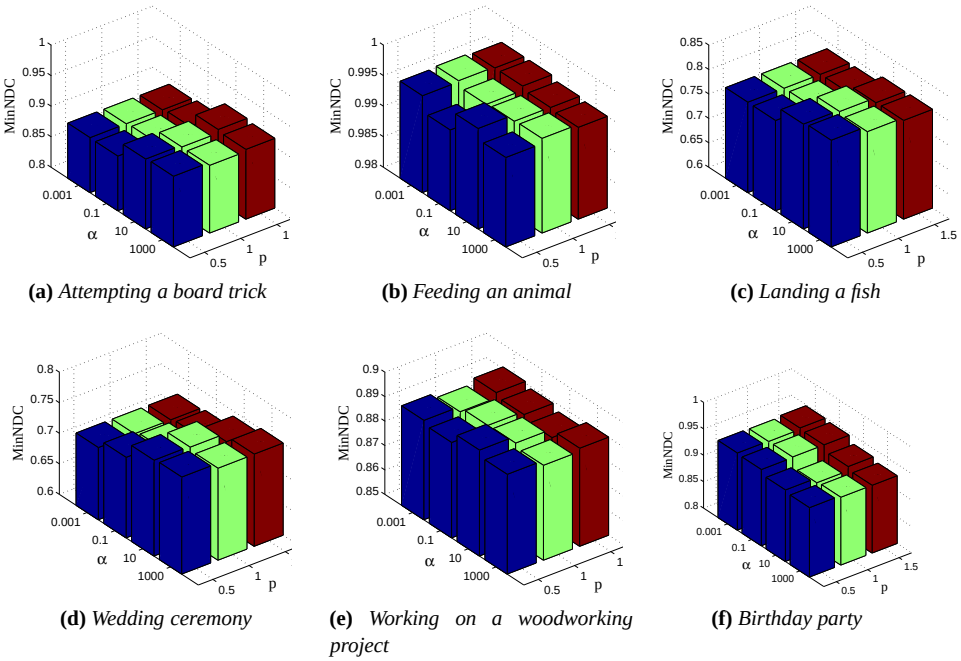


Figure 22: Performance comparison between using auxiliary knowledge and not using auxiliary knowledge.

there is no knowledge adaptation; Second, since in our objective function the item $\alpha \|W\|_{2,p}^{p}$ controls the effect of the knowledge adaptation, we investigate the influence by varying the parameter α and p after fixing β at its optimal values.

For the first scenario, we show the performance comparison between using auxiliary data and not using it in Figure 22. MinNDC is used as metric due to the space limit. It can be seen that using auxiliary data has clear advantage over not using it, which demonstrates that through proper design, the auxiliary knowledge contributes notably to the MED with few exemplars.

For the second scenario, we similarly use MinNDC as metric to show the performance variation. We show the results on the first 6 events in Figure 23 as showcases. We observe from Figure 23 that the best results are generally obtained when $p = 0.5$ or $p = 1$. For the other parameter α there is no obvious rule, which is presumably data-dependent. Lower p indicates that the model is more sparse, thereby eliminating more redundancy and noise.



(a) *Attempting a board trick* **(b)** *Feeding an animal* **(c)** *Landing a fish*

(d) *Wedding ceremony* **(e)** *Working on a woodworking project* **(f)** *Birthday party*

Figure 23: The detection performance variance *w.r.t.* α and p.

5.6.5 *Using Fewer Concepts*

In this experiment, we test the performance variance of the proposed algorithm by varying the number of auxiliary concepts as 5, 10, 20, 30, 50 and 65. Figure 24 displays the corresponding results in terms of Minimum NDC. We have the following observations: 1) Generally, the performance of using only 5 auxiliary concepts is noticeably worse than using all the 65 auxiliary concepts; 2) From using 5 auxiliary concepts to using 30 auxiliary concepts, the performance is gradually improved for most events; 3) From using 30 auxiliary concepts to using 65 auxiliary concepts, the performance does not vary much, which suggests that the performance saturates at the point when 30 auxiliary concepts are used. Our observation indicates that when the number of auxiliary concepts is very few, which also means few auxiliary videos, the performance gain is limited. To get more performance boost, we may want to incorporate more auxiliary videos with more concepts. However, how to decide the optimal number of auxiliary concepts is still an open problem and is out of the scope of this chapter. It would be an interesting topic in our future work.

5.6.6 *Do Negative Examples Help?*

We further conduct an experiment to evaluate whether negative examples contribute much to the detection accuracy by reducing the number of negative examples to 500 and 100. Figure 25 shows the performance comparison between using 100, 500 and 1000 negative examples. Similarly, Minimum NDC is chosen as the evaluation metric.

From Figure 25 we have the following observations: 1). Using 1000 or 500 negative examples is better than using only 100 negative examples. 2). The performance difference between using 1000 and using 500 negative examples is quite small. This experiment indicates that negative examples are helpful in improving the detection accuracy in some degree. For example, when 500 or 1000

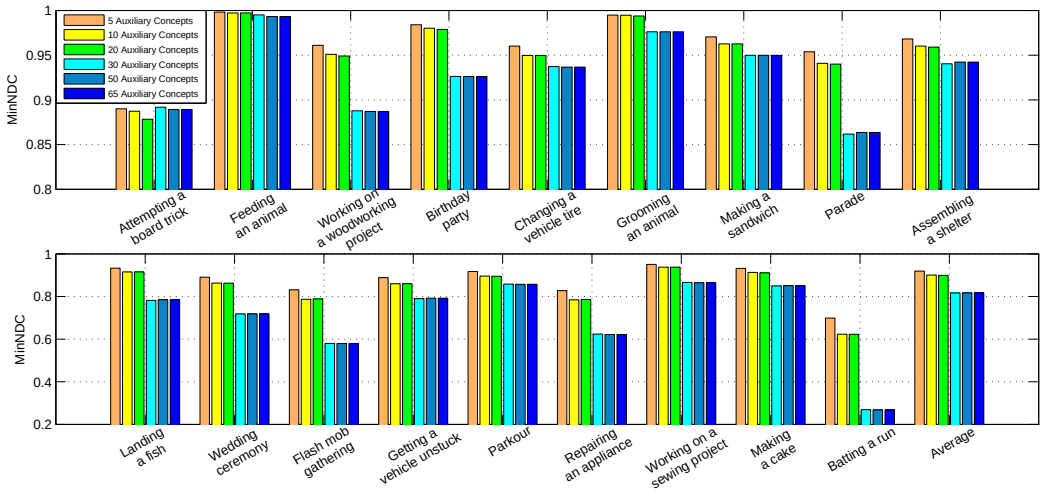


Figure 24: Performance comparison between using 5, 10, 20, 30, 50 and 65 auxiliary concepts.

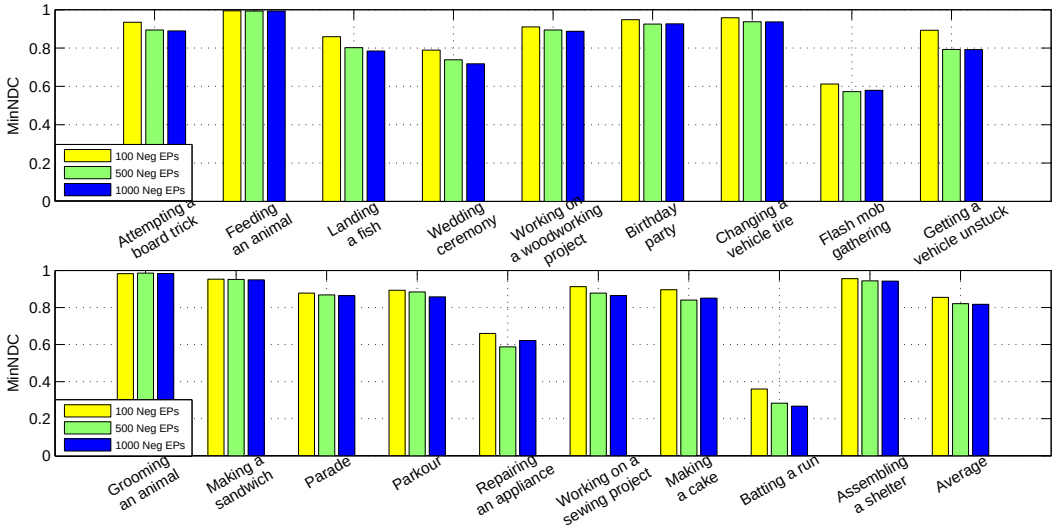
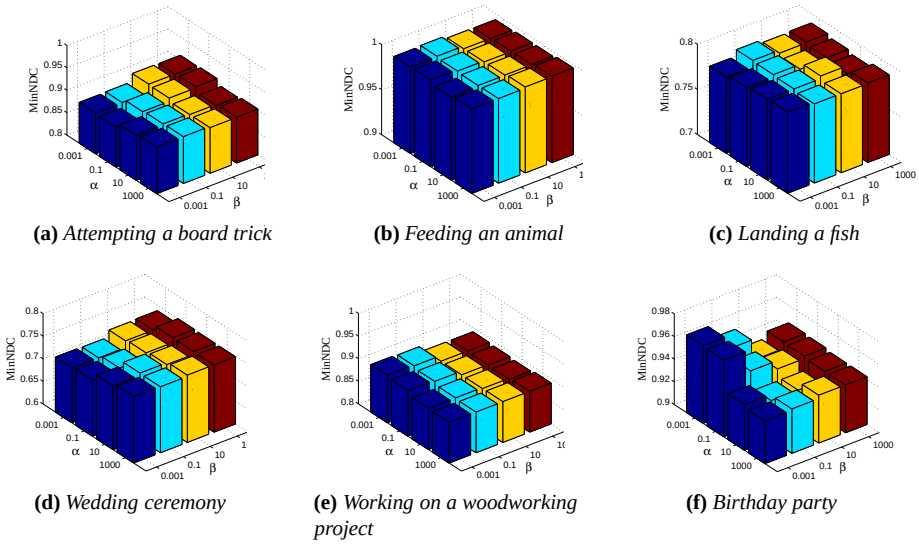


Figure 25: Performance comparison between using 100, 500 and 1000 negative examples.

negative examples are used, the performance is generally better than using 100 negative examples only. However, as the number of negative examples used increases, the performance gain does not increase accordingly, *e.g.*, from using 500 negative examples to using 1000 negative examples. How many negative examples would bring in the most performance gain is still an open problem, which is out of the scope of this chapter. However, since negative examples are quite easy to obtain in the real world, it is natural to leverage such cheap resources for boosted detection accuracy.

5.6.7 *Parameter Sensitivity Study*

There are two regularization parameters, denoted as α and β in Eq. (5.5). To learn how they affect the performance on image annotation, we conduct an experiment on the parameter sensitivity. We still only show the results on the first 6 events in Figure 26. From Figure 26 we notice that for



(a) *Attempting a board trick* **(b)** *Feeding an animal* **(c)** *Landing a fish*

(d) *Wedding ceremony* **(e)** *Working on a woodworking project* **(f)** *Birthday party*

Figure 26: The detection performance variance *w.r.t.* α and β.

some events, *e.g.*, *Birthday party*, the performance is sensitive to the two parameters. For some other events like *Feeding an animal* the performance does not change much. However, we can generally obtain good performance for these events when α and β are comparable. For example, good performance is obtained when $\alpha = \beta = 0.001$ for *Attempting a board trick*, *Feeding an animal*, *Landing a fish* and *Wedding ceremony*, and $\alpha = \beta = 10$ for *Birthday party*. Similar pattern is observed for other events as well.

5.6.8 *Convergence Study*

We solve our objective problem using an iterative approach. In practice, how fast our algorithm converges is crucial for the whole computational efficiency. Hence, we conduct an experiment to show the convergence curve of our algorithm. As we have similar results on all the 18 events, we only present the convergence curve on the first event. All the parameters involved are fixed at 1. Figure 27 shows the convergence curve. It can be seen that the objective function value converges within 10 iterations. The convergence experiment demonstrates the efficiency of our iterative algorithm.

5.7 COMPLEMENTARY EXPERIMENT ON MULTI-CLASS CLASSIFICATION

Our proposed algorithm can be easily extended to other applications such as multi-class classification. In this section, we conduct a complementary experiment on image annotation to show its effectiveness for multi-class classification.

We use the Animals with Attributes (AwA) dataset [38] for evaluation. The reason is that the dataset has both animal categories and the associated attributes. Similarly to our assumption, different animal categories may share common attributes. Thus, we use the 10 animal categories specified in [38] as our target annotation categories and the rest as our auxiliary data. Note that for the auxiliary data we use their attribute labels since these attributes are the shared components with the target animal categories. The 10 target categories are *persian cat*, *hippopotamus*, *leopard*, *humpback whale*, *seal*, *chimpanzee*, *rat*, *giant panda*, *pig* and *raccoon*. For the 10 classes to be annotated,

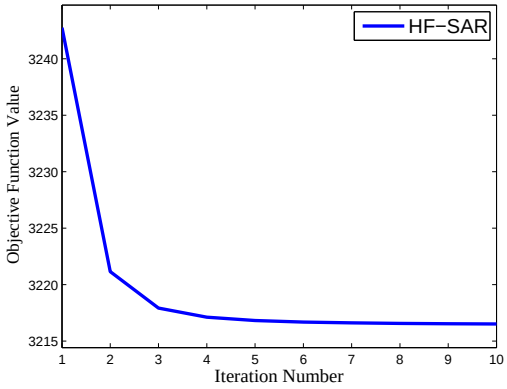


Figure 27: Convergence curve of the objective function value in Eq. (5.5) using Algorithm 4. The figure shows that the objective function value monotonically decreases until convergence by applying the proposed algorithm.

Table 18: Performance comparison of different methods on image annotation. The best result is highlighted in bold.

Evaluation Metric	SAR	A-MKL	MKTL	SAR&SVM	SVM	TaylorBoost	HF-SAR
Accuracy	0.257	0.248	0.232	0.265	0.310	0.264	**0.373**

we randomly select 10 samples per category to form the training set and the remaining data of these categories are our testing data. We use the SIFT feature as the homogeneous feature and the Locality Similarity Histogram (LSH) feature as the heterogeneous feature for image representation. In other words, the images of the 10 target categories are represented by SIFT and LSH while those of the auxiliary categories are represented by SIFT only.

The annotation comparison between different algorithms is displayed in Table 18. We can see that HF-SAR is much better than other comparison algorithms. SVM is second best algorithm. Especially, other transfer learning algorithms have weaker performance as only one feature is exploited.

The reported accuracy in [38] is 40.5%. But we point out that in [38] six features have been used whereas we only use two features. We did not use all the features used in [38] because we were concerned with the computational efficiency, *e.g.*, the comparison algorithms A-MKL and TaylorBoost are computationally expensive. On the other hand, to be consistent with our previous experiment on MED with few exemplars, we select 10 samples from each target category to form the training set, making our training set also different from that in [38]. Thus, we cannot directly compare the annotation accuracy of our method and that of [38].

This complementary experiment has demonstrated that our method also has the potential for other applications.

5.8 CONCLUSION

In this paper, we have introduced the research exploration of MED with few exemplars. This is an important research issue as it focuses on more generic, complicated and meaningful events that reflect our daily activities. In addition, the situation we are faced in the real world requires that only few positive examples are used. To achieve good performance, we have proposed to borrow

strength from available concepts-based videos for MED with few exemplars. A notable difference between our proposed algorithm and most existing knowledge adaptation algorithms is that it is built upon heterogeneous features, *i.e.*, the features of the source and the target are partially different, but overlapping. Specifically, we first mine the shared irrelevance and noise between the auxiliary videos and the target videos based on the homogeneous features. Then a sophisticated method is exerted to alleviate the negative impact of the irrelevance and noise to optimize the event detector. Meanwhile, another event detector of MED videos is trained based on the heterogeneous feature. Then we integrate the two event detectors for optimization, after which the decision values from both are fused for the final prediction. Extensive experiments using real-world multimedia archives were conducted with results showing that our method outperforms all the comparison algorithms. The results validate that: 1) it is beneficial to leverage auxiliary knowledge for MED when we do not have sufficient positive examples; and 2) the capability of knowledge adaptation based on heterogeneous features is realistic and advantageous.

We would like to point out that the effectiveness of our method is grounded on the condition that the auxiliary concepts be relevant to the target events. If the concepts and the events are not related, it is unlikely for us to mine the shared noise and redundancy, thus improving the event detection. This is a limitation as it is difficult to generalize our method to Ad Hoc MED as we are not supposed to look into the events and tune our system accordingly. That said, the selected auxiliary concepts would possibly have little correlation to an Ad Hoc event, thus limiting its helpfulness for event detection. However, a possible solution to address this problem is to enlarge the repository of the auxiliary concepts to thousands of concepts covering a wide range of objects, scenes and actions. This approach, for sure, would cause computational burden but would be worth a try with the fast development of our computing facilities. Our method is also based on the hypothesis that the feature representations from both domains are noisy and redundant. If even more discriminating features with little noise or redundancy is developed in the future, our method would lose its capability of harnessing the shared noise or redundancy. Hence, the performance gain from using our method would be presumably limited.

6

CONCLUSION

In this thesis we have addressed multimedia analysis with the focus on different applications, *i.e.*, image and video annotation, and multimedia event detection.

Multimedia analysis is a fundamental tool for many applications. The primary problem of this topic is to overcome the semantic gap between the low-level features and high-level semantics. That said, features work as the basis for understanding multimedia content. Extracting discriminating features, therefore, plays an important role in attaining good performance. What else can we do except the feature design itself? The generation of feature representations would presumably bring in noise and redundancy. Hence, the feature representation can be improved from two aspects. By removing the noise we can get more accurate representation whereas by removing the redundancy we can reduce the dimension of the representation. As a result, it is helpful for attaining better analyzing accuracy and efficiency. A widely used technique for this purpose is feature selection. Though plenty of feature selection algorithms have been proposed in the literature, most of them select the features in a one-by-one fashion. Meaning: They evaluate the feature importance independently, thus ignoring the correlation between different features. Aiming to address this shortcoming, we have proposed to do feature selection in a batch mode with a sparse model in Chapter 2. In this way, the correlation between different features is taken into account. As we have focused on Web image annotation where many of the images have multiple semantic labels, *i.e.*, one image can depict multiple concepts, we further incorporate the subspace learning scheme to uncover the correlation between different labels. The experimental results on Web images have validated the effectiveness of our method.

Following the progress in the feature level, we have considered making some effort for multimedia analysis in the classifier level. Particularly, we asked ourselves one question: what can be a common problem for both image and video analysis? There probably exist many common problems but the scarcity of precise labeled images and videos gets our attention. Tons of images and videos are uploaded to the Internet every day. Users tend to add descriptions to their uploaded images and videos but such descriptions can be subjective and noisy or even irrelevant to the real semantic concepts. However, learning a robust analyzing model requires precise labels to associate the low-level features with the high-level semantic concepts. For sure we can manually label images and videos but it requires expertise and much human labor. Previous work has shown that semi-supervised learning is a good way to handle the paucity of accurate labels as it simultaneously exploits labeled and unlabeled training data. Hence, we have developed a novel semi-supervised feature analysis algorithm for image and video annotation in Chapter 3. Our method has integrated manifold learning, inductive learning and a sparse model, thus resulting in its capability of utilizing discriminative features for classifying out-of-sample data when only few labeled training data are provided. The method has been applied to image and video annotation with encouraging performance.

The videos focused in Chapter 3 contain mostly simple objects, scenes and activities. Yet in our daily life, users are more interested in complicated multimedia events such as *Dog show*. Having noticed that, we decided to work towards more complicated event-based video analysis. A multi-

media event builds upon several related concepts such as objects and actions. It is more difficult to understand multimedia event as it usually exists in long video clips with huge intra-class variations. Furthermore, we have focused on multimedia event detection which is way more difficult than annotation since the we have to detect a particular event from an infinite pool of unknown classes. We have leveraged the fact that events contain concepts by learning an intermediate representation from both event videos and auxiliary concept-based videos. The intermediate representation is optimized together with the event detector so we would expect improved detection accuracy. The proposed approach has been evaluated on a large-scale multimedia event video archive. The experimental results show that our approach works better than the main-stream classifier SVM.

Having achieved encouraging progress on multimedia event detection, we further pushed the research on this topic to an even more challenging problem, *i.e.*, detection with only few positive exemplars in Chapter 5. This problem also corresponds to the paucity of precisely labeled multimedia data. In contrast with the semi-supervised approach we used in Chapter 3, we have investigated the efficacy of transfer learning for our problem. The reasons are: First, multimedia events are higher-level multimedia contents based on objects, scenes and actions, which means the two domains have certain shared components; Second, the research community has already contributed many precisely labeled multimedia archives related to objects, scenes and actions; Third, the assumption of transfer learning is that we have abundant labeled data in the auxiliary domain while few labeled data in the target domain. Technically, we assumed that videos including objects, scenes and actions and those including complex events have shared noise and irrelevance. To this end, we have taken advantage of novel sparse models on both domains to jointly remove the noise and irrelevance. On top of that, we have investigated another meaningful direction for transfer learning. Most existing transfer learning algorithms require that the features of the target and auxiliary domains are of the same type. Nonetheless, in many applications such a requirement may be too restrictive. In practice, the data of multimedia event videos and those in the available concept-based video archives usually only have partially shared data features. Hence, we have extended our algorithm to be able to effectively adapt knowledge from one domain to another when the available feature sets are partially different, but overlapping, for example if new or different features have more or better instrumentation for observations. Our newly proposed method was tested on large-scale multimedia event videos and the results have shown that it outperforms mainstream classifiers such as SVM and several other state-of-the-art transfer learning algorithms.

In summary, in this thesis we have studied different machine learning techniques for multimedia analysis. Our work suggests that proper usage of feature selection, semi-supervised learning and transfer learning does help improve the overall understanding of multimedia contents. Hence, in the future we will continue our research on this direction with the following possible pursuits:

- With the advance of computer vision research, a variety of features have been proposed to represent images and videos. Focusing on different characteristics of multimedia data, these features intuitively should complement each other. That said, it is highly possible to further boost the analysis performance by proper use of multiple features. Hence, it would be interesting to study on novel algorithms that are capable of harnessing different features jointly as symbiotic solutions.

- New automatic methodologies will be developed for effective exploitation of knowledge in large-scale sensor data with emphasis on spatial information. We will still focus on the common problem that when systems are creating knowledge from complex data, there are not enough examples of the phenomena interest that have been labeled by analysts for an automated system to accurately classify and label. Our research will facilitate knowledge adaptation that leverages unlabeled data through exploitation of knowledge in multiple related domains and knowledge adaptation between two domains that have partially shared data features.

- All the research effort on multimedia analysis is essentially for serving users. Therefore, we will be interested in investigating user-centric research problems. How to conduct user behavior analysis, user emotional analysis, user perspective understanding, user attention understanding and user need mining would be of great benefit to health care, commercial, art, esthetics and *etc*.

ACKNOWLEDGEMENT

This thesis cannot be done without many important people in my work and life. My good friend Dr. Yi Yang, is my mentor who led me into the realm of machine learning and multimedia analysis. Through our collaboration, I have learnt a lot of theoretic knowledge and research skills. My supervisor, Prof. Nicu Sebe, is not only a good teacher but also a good friend, a big brother to us. He is always there to help us in our work and our life. He has created a diverse, intimate and creative M-Hug group. I have really enjoyed working with him and I am so proud to have been a member of M-Hug. Dr. Alexander G. Hauptmann was my supervisor when I visited Carnegie Mellon University. I am always impressed by the way he thinks about every research problem and how he leads and unifies the whole team with his personal charm working toward the goal. He taught me to think about my work laterally and in an overall situation. Dr. Shuicheng Yan was my supervisor when I visited National University of Singapore. His self-motivation, hard-working spirit and rigorous research attitude has set an excellent example for me. Another good friend Dr. Feiping Nie is a master of maths. I always admire how knowledgeable he is on maths. He has taught me a lot of mathematical theory and skills that are very useful in my work. I am also very pleased to have the opportunity to work with Dr. Jasper R. R. Uijlings. What I have learnt from him is also his meticulous thinking style on our work. He is always trying to figure out every detail of the research problem and can give me quite insightful advice. My beloved M-Huggers are my source of happiness and joy. Everyone is unique, friendly and helpful. They have made my life in Trento, Italy so enjoyable and it will definitely be an unforgettable experience in my life. I truly appreciate all the people mentioned above and wish everyone happiness and bright future.

Lastly, I would like to thank my parents for their continuous support for my academic pursuit. I wish them health and happiness.

BIBLIOGRAPHY

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

[2] http://www.nist.gov/itl/iad/mig/upload/med11-evalplan-v03-20110801a.pdf.

[3] http://www.nist.gov/itl/iad/mig/upload/med12-evalplan-v01.pdf.

[4] Trec video retrieval evaluation. National Institute of Standards and Technology. In *http://www-nlpir.nist.gov/projects/trecvid/*.

[5] Amit Adam, Ehud Rivlin, Ilan Shimshoni, and David Reinitz. Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(3):555–560, 2008.

[6] Jaume Amores, Nicu Sebe, and Petia Radeva. Context-based object-class recognition and retrieval by generalized correlograms. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(10):1818–1833, 2007.

[7] Rie Kubota Ando and Tong Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853, 2005.

[8] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Multi-task feature learning. In *NIPS*, pages 41–48, 2006.

[9] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

[10] Lei Bao, Longfei Zhang, Shoou-I Yu, Zhenzhong Lan, Jiang Lu, Arnold Overwijk, Qin Jin, Shohei Takahashi, Brian Langner, Yuanpeng Li, Michael Garbus, Susanne Burger, Florian Metze, and Alexander G. Hauptmann. Informedia @ trecvid2011. In *TRECVID*, 2011.

[11] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006.

[12] Deng Cai, Chiyuan Zhang, and Xiaofei He. Unsupervised feature selection for multi-cluster data. In *KDD*, pages 333–342, 2010.

[13] Liangliang Cao, Shih-Fu Chang, Noel Codella, Courtenay Cotton, Dan Ellis, Leiguang Gong, Matthew Hill, Gang Hua, John Kender, Michele Merler, Yadong Mu, Apostol Natsev, and John R. Smith. IBM Research and Columbia University TRECVID-2011 Multimedia Event Detection (MED) System. In *NIST TRECVID Workshop*, 2011.

[14] Gustavo Carneiro, Antoni B. Chan, Pedro J. Moreno, and Nuno Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(3):394–410, 2007.

[15] Gavin C. Cawley, Nicola L. C. Talbot, and Mark Girolami. Sparse multinomial logistic regression via bayesian l1 regularisation. In *NIPS*, pages 209–216, 2006.

[16] Bin Cheng, Jianchao Yang, Shuicheng Yan, Yun Fu, and Thomas S. Huang. Learning with ℓ^1-graph for image analysis. *IEEE Transactions on Image Processing*, 19(4):858–866, 2010.

[17] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *CIVR*, 2009.

[18] Ira Cohen, Fabio Gagliardi Cozman, Nicu Sebe, Marcelo Cesar Cirelo, and Thomas S. Huang. Semisupervised learning of classifiers: Theory, algorithms, and their application to human-computer interaction. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(12):1553–1567, 2004.

[19] Alberto del Bimbo. Visual information retrieval. *Morgan Kaufmann*, 1999.

[20] Duo Ding, Florian Metze, Shourabh Rawat, Peter Franz Schulam, Susanne Burger, Ehsan Younessian, Lei Bao, Michael G. Christel, and Alexander G. Hauptmann. Beyond audio and video retrieval: towards multimedia summarization. In *ICMR*, page 2, 2012.

[21] Lixin Duan, Dong Xu, Ivor Wai-Hung Tsang, and Jiebo Luo. Visual event recognition in videos by learning from web data. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(9):1667–1680, 2012.

[22] Richard Duda, David Stork, and Peter Hart. Pattern classification (2nd ed.). *Wiley-Interscience, New York, USA*, 2001.

[23] Keinosuke Fukunaga. Introduction to statistical pattern recognition (2nd ed.). *Academic Press Professional, San Diego, USA*, 1990.

[24] Yuli Gao, Jianping Fan, Xiangyang Xue, and Ramesh Jain. Automatic image annotation by incorporating feature hierarchy and boosting to scale up svm classifiers. In *ACM Multimedia*, pages 901–910, 2006.

[25] AmirHossein Habibian, Koen E. A. van de Sande, and Cees G. M. Snoek. Recommendations for video event recognition using concept vocabularies. In *ICMR*, pages 89–96, 2013.

[26] Yahong Han, Fei Wu, Jinzhu Jia, Yueting Zhuang, and Bin Yu. Multi-task sparse discriminant analysis (mtsda) with overlapping categories. In *AAAI*, 2010.

[27] Alexander G. Hauptmann, Michael G. Christel, and Rong Yan. Video retrieval based on semantic concepts. *Proceedings of the IEEE*, 96(4):602–622, 2008.

[28] Alexander G. Hauptmann, Rong Yan, Wei-Hao Lin, Michael G. Christel, and Howard D. Wactlar. Can high-level concepts fill the semantic gap in video retrieval? a case study with broadcast news. *IEEE Transactions on Multimedia*, 9(5):958–966, 2007.

[29] Steven C. H. Hoi, Wei Liu, and Shih-Fu Chang. Semi-supervised distance metric learning for collaborative image retrieval and clustering. *TOMCCAP*, 6(3), 2010.

[30] Steven C. H. Hoi, Michael R. Lyu, and Rong Jin. A unified log-based relevance feedback scheme for image retrieval. *IEEE Trans. Knowl. Data Eng.*, 18(4):509–524, 2006.

[31] Weiming Hu, Nianhua Xie, Li Li, Xianglin Zeng, and Stephen J. Maybank. A survey on visual content-based video indexing and retrieval. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 41(6):797–819, 2011.

[32] Shuiwang Ji, Lei Tang, Shipeng Yu, and Jieping Ye. A shared-subspace learning framework for multi-label classification. *TKDD*, 4(2), 2010.

[33] Shuiwang Ji and Jieping Ye. Linear dimensionality reduction for multi-label classification. In *IJCAI*, pages 1077–1082, 2009.

[34] Yu-Gang Jiang, Chong-Wah Ngo, and Shih-Fu Chang. Semantic context transfer across heterogeneous sources for domain adaptive video search. In *ACM Multimedia*, pages 155–164, 2009.

[35] Effrosini Kokiopoulou and Yousef Saad. Orthogonal neighborhood preserving projections: A projection-based dimensionality reduction technique. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(12):2143–2156, 2007.

[36] Igor Kononenko. Estimating attributes: Analysis and extensions of relief. In *ECML*, pages 171–182, 1994.

[37] Balaji Krishnapuram, Alexander J. Hartemink, Lawrence Carin, and Mário A. T. Figueiredo. A bayesian approach to joint feature selection and classifier design. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(9):1105–1111, 2004.

[38] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, pages 951–958, 2009.

[39] Ivan Laptev and Tony Lindeberg. Space-time interest points. In *ICCV*, pages 432–439, 2003.

[40] Martin H. C. Law, Mário A. T. Figueiredo, and Anil K. Jain. Simultaneous feature selection and clustering using mixture models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(9):1154–1166, 2004.

[41] Michael S. Lew, Nicu Sebe, Chabane Djeraba, and Ramesh Jain. Content-based multimedia information retrieval: State of the art and challenges. *TOMCCAP*, 2(1):1–19, 2006.

[42] Fei-Fei Li and Li-Jia Li. What, where and who? telling the story of an image by activity classification, scene recognition and object categorization. In *Computer Vision: Detection, Recognition and Reconstruction*, pages 157–171. 2010.

[43] Hao Li, Meng Wang, and Xian-Sheng Hua. Msra-mm 2.0: A large-scale web multimedia dataset. In *ICDM Workshops*, pages 164–169, 2009.

[44] Yen-Yu Lin, Tyng-Luh Liu, and Hwann-Tzong Chen. Semantic manifold learning for image retrieval. In *ACM Multimedia*, pages 249–258, 2005.

[45] Ken-Hao Liu, Ming-Fang Weng, Chi-Yao Tseng, Yung-Yu Chuang, and Ming-Syan Chen. Association and temporal rule mining for post-filtering of semantic concept detection in video. *IEEE Transactions on Multimedia*, 10(2):240–251, 2008.

[46] Alexander Loui, Jiebo Luo, Shih-Fu Chang, Dan Ellis, Wei Jiang, Lyndon Kennedy, Keansub Lee, and Akira Yanagawa. Kodak's consumer video benchmark data set: Concept definition and annotation. In *Proceedings of the International Workshop on Multimedia Information Retrieval*, pages 245–254, 2007.

[47] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[48] Yijuan Lu and Qi Tian. Discriminant subspace analysis: An adaptive approach for image classification. *IEEE Transactions on Multimedia*, 11(7):1289–1300, 2009.

[49] Jie Luo, Tatiana Tommasi, and Barbara Caputo. Multiclass transfer learning from unconstrained priors. In *ICCV*, pages 1863–1870, 2011.

[50] Jiebo Luo, Jie Yu, Dhiraj Joshi, and Wei Hao. Event recognition: viewing the world with a third eye. In *ACM Multimedia*, pages 1071–1080, 2008.

[51] Zhigang Ma, Alexander G. Hauptmann, Yi Yang, and Nicu Sebe. Classifier-specific intermediate representation for multimedia tasks. In *ICMR*, page 50, 2012.

[52] Zhigang Ma, Feiping Nie, Yi Yang, Jasper R. R. Uijlings, and Nicu Sebe. Web image annotation via subspace-sparsity collaborated feature selection. *IEEE Transactions on Multimedia*, 14(4):1021–1030, 2012.

[53] Zhigang Ma, Feiping Nie, Yi Yang, Jasper R. R. Uijlings, Nicu Sebe, and Alexander G. Hauptmann. Discriminating joint feature analysis for multimedia data understanding. *IEEE Transactions on Multimedia*, 14(6):1662–1672, 2012.

[54] Zhigang Ma, Yi Yang, Yang Cai, Nicu Sebe, and Alexander G. Hauptmann. Knowledge adaptation for ad hoc multimedia event detection with few exemplars. In *ACM Multimedia*, pages 469–478, 2012.

[55] Zhigang Ma, Yi Yang, Feiping Nie, Jasper R. R. Uijlings, and Nicu Sebe. Exploiting the entire feature space with sparsity for automatic image annotation. In *ACM Multimedia*, pages 283–292, 2011.

[56] Zhigang Ma, Yi Yang, Nicu Sebe, Kai Zheng, and Alexander G. Hauptmann. Multimedia event detection using a classifier-specific intermediate representation. *IEEE Transactions on Multimedia*, 15(7):1628–1637, 2013.

[57] Zhigang Ma, Yi Yang, Zhongwen Xu, Shuicheng Yan, Nicu Sebe, and Alexander G. Hauptmann. Complex event detection via multi-source video attributes. In *CVPR*, pages 2627–2633, 2013.

[58] Marcin Marszałek, Cordelia Schmid, Hedi Harzallah, and Joost van de Weijer. Learning object representations for visual object class recognition. In *Visual Recognition Challange workshop, in conjunction with ICCV*, 2007.

[59] Emily Moxley, Tao Mei, and Bangalore S. Manjunath. Video annotation through search and graph reinforcement mining. *IEEE Transactions on Multimedia*, 12(3):184–193, 2010.

[60] Milind R. Naphade and John R. Smith. On the detection of semantic concepts at trecvid. In *ACM Multimedia*, pages 660–667, 2004.

[61] Milind R. Naphade, John R. Smith, Jelena Tesic, Shih-Fu Chang, Winston H. Hsu, Lyndon S. Kennedy, Alexander G. Hauptmann, and Jon Curtis. Large-scale concept ontology for multimedia. *IEEE MultiMedia*, 13(3):86–91, 2006.

[62] Feiping Nie, Heng Huang, Xiao Cai, and Chris H. Q. Ding. Efficient and robust feature selection via joint ;2, 1-norms minimization. In *NIPS*, pages 1813–1821, 2010.

[63] Feiping Nie, Dong Xu, Ivor Wai-Hung Tsang, and Changshui Zhang. Flexible manifold embedding: A framework for semi-supervised and unsupervised dimension reduction. *IEEE Transactions on Image Processing*, 19(7):1921–1932, 2010.

[64] Huazhong Ning, Wei Xu, Yihong Gong, and Thomas S. Huang. Discriminative learning of visual words for 3d human pose estimation. In *CVPR*, 2008.

[65] Stefanie Nowak, Ainhoa Llorente, Enrico Motta, and Stefan M. Rüger. The effect of semantic relatedness measures on multi-label classification evaluation. In *CIVR*, pages 303–310, 2010.

[66] Guillaume Obozinski, Ben Taskar, and Michael I. Jordan. Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, 20(2):231–252, 2010.

[67] Kishore K. Reddy and Mubarak Shah. Recognizing 50 human action categories of web videos. *Mach. Vis. Appl.*, 24(5):971–981, 2013.

[68] Elisa Ricci, Gloria Zen, Nicu Sebe, and Stefano Messelodi. A prototype learning framework using emd: Application to complex scenes analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(3):513–526, 2013.

[69] Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, and William T. Freeman. Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision*, 77(1-3):157–173, 2008.

[70] Mohammad J. Saberian, Hamed Masnadi-Shirazi, and Nuno Vasconcelos. Taylorboost: First and second-order boosting algorithms with explicit margin control. In *CVPR*, pages 2929–2934, 2011.

[71] David A. Sadlier and Noel E. O'Connor. Event detection in field sports video using audio-visual features and a support vector machine. *IEEE Trans. Circuits Syst. Video Techn.*, 15(10):1225–1233, 2005.

[72] Bernhard Schölkopf, Alex J. Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

[73] Mei-Ling Shyu, Zongxing Xie, Min Chen, and Shu-Ching Chen. Video semantic event/concept detection using a subspace-based multimedia data mining framework. *IEEE Transactions on Multimedia*, 10(2):252–259, 2008.

[74] Leonid Sigal and Michael J. Black. Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion. *Technical Report CS-06-08, Brown University, Department of Computer Science*, 2006.

[75] Vikas Sindhwani, Partha Niyogi, and Mikhail Belkin. Linear manifold regularization for large scale semi-supervised learning. In *ICML Workshop on Learning with Partially Classified Training Data*, 2005.

[76] Alan F. Smeaton, Paul Over, and Wessel Kraaij. Evaluation campaigns and trecvid. In *Multimedia Information Retrieval*, pages 321–330, 2006.

[77] Arnold W. M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1349–1380, 2000.

[78] Cees Snoek, Marcel Worring, Jan van Gemert, Jan-Mark Geusebroek, and Arnold W. M. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *ACM Multimedia*, pages 421–430, 2006.

[79] Vincent S. Tseng, Ja-Hwung Su, Jhih-Hong Huang, and Chih-Jen Chen. Integrated mining of visual features, speech features, and frequent patterns for semantic video annotation. *IEEE Transactions on Multimedia*, 10(2):260–267, 2008.

[80] Adrian Ulges, Marcel Worring, and Thomas M. Breuel. Learning visual contexts for image annotation from flickr groups. *IEEE Transactions on Multimedia*, 13(2):330–341, 2011.

[81] Koen E. A. van de Sande, Theo Gevers, and Cees G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1582–1596, 2010.

[82] Shiv Naga Prasad Vitaladevuni, Pradeep Natarajan, Rohit Prasad, and Prem Natarajan. Efficient orthogonal matching pursuit using sparse random projections for scene and video classification. In *ICCV*, pages 2312–2319, 2011.

[83] Feng Wang, Yu-Gang Jiang, and Chong-Wah Ngo. Video event detection using motion relativity and visual relatedness. In *ACM Multimedia*, pages 239–248, 2008.

[84] Gang Wang, Tat-Seng Chua, and Ming Zhao. Exploring knowledge of sub-domain in a multi-resolution bootstrapping framework for concept detection in news video. In *ACM Multimedia*, pages 249–258, 2008.

[85] Meng Wang, Xian-Sheng Hua, Jinhui Tang, and Richang Hong. Beyond distance measurement: Constructing neighborhood similarity for video annotation. *IEEE Transactions on Multimedia*, 11(3):465–476, 2009.

[86] Xin-Jing Wang, Lei Zhang, Xirong Li, and Wei-Ying Ma. Annotating images by mining image search results. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(11):1919–1932, 2008.

[87] Hua-Liang Wei and Stephen A. Billings. Feature subset selection and ranking for data dimensionality reduction. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(1):162–166, 2007.

[88] Fei Wu, Yahong Han, Qi Tian, and Yueting Zhuang. Multi-label boosting for image annotation by structural grouping sparsity. In *ACM Multimedia*, pages 15–24, 2010.

[89] Fei Wu, Ying Yuan, and Yueting Zhuang. Heterogeneous feature selection by group lasso with logistic regression. In *ACM Multimedia*, pages 983–986, 2010.

[90] Changsheng Xu, Jinjun Wang, Kongwah Wan, Yiqun Li, and Lingyu Duan. Live sports event detection based on broadcast video and web-casting text. In *ACM Multimedia*, pages 221–230, 2006.

[91] Jun Yang, Rong Yan, and Alexander G. Hauptmann. Cross-domain video concept detection using adaptive svms. In *ACM Multimedia*, pages 188–197, 2007.

[92] Yi Yang, Zhigang Ma, Alexander G. Hauptmann, and Nicu Sebe. Feature selection for multimedia analysis by sharing information among multiple tasks. *IEEE Transactions on Multimedia*, 15(3):661–669, 2013.

[93] Yi Yang, Zhigang Ma, Zhongwen Xu, Shuicheng Yan, and Alexander G. Hauptmann. How related exemplars help complex event detection inweb videos? In *ICCV*, 2013.

[94] Yi Yang, Feiping Nie, Dong Xu, Jiebo Luo, Yueting Zhuang, and Yunhe Pan. A multimedia retrieval framework based on semi-supervised ranking and relevance feedback. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(4):723–742, 2012.

[95] Yi Yang, Heng Tao Shen, Zhigang Ma, Zi Huang, and Xiaofang Zhou. $l_{2,1}$-norm regularized discriminative feature selection for unsupervised learning. In *IJCAI*, pages 1589–1594, 2011.

[96] Yi Yang, Dong Xu, Feiping Nie, Jiebo Luo, and Yueting Zhuang. Ranking with local regression and global alignment for cross media retrieval. In *ACM Multimedia*, pages 175–184, 2009.

[97] Yi Yang, Dong Xu, Feiping Nie, Shuicheng Yan, and Yueting Zhuang. Image clustering using local discriminant models and global integration. *IEEE Transactions on Image Processing*, 19(10):2761–2773, 2010.

[98] Yi Yang, Yueting Zhuang, Fei Wu, and Yunhe Pan. Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval. *IEEE Transactions on Multimedia*, 10(3):437–446, 2008.

[99] Yi Yang, Yueting Zhuang, Dong Xu, Yunhe Pan, Dacheng Tao, and Stephen J. Maybank. Retrieval based interactive cartoon synthesis via unsupervised bi-distance metric learning. In *ACM Multimedia*, pages 311–320, 2009.

[100] Ming yu Chen and Alexander G. Hauptmann. Mosift: Recognizing human actions in surveillance videos. *Technical Report CMU-CS-09-161, Carnegie Mellon University*, 2009.

[101] Gloria Zen and Elisa Ricci. Earth mover's prototypes: A convex learning approach for discovering activity patterns in dynamic scenes. In *CVPR*, pages 3225–3232, 2011.

[102] Zheng-Jun Zha, Meng Wang, Yan-Tao Zheng, Yi Yang, Richang Hong, and Tat-Seng Chua. Interactive video indexing with statistical active learning. *IEEE Transactions on Multimedia*, 14(1):17–27, 2012.

[103] Changshui Zhang, Feiping Nie, and Shiming Xiang. A general kernelization framework for learning algorithms based on kernel pca. *Neurocomputing*, 73(4-6):959–967, 2010.

[104] Jidong Zhao, Ke Lu, and Xiaofei He. Locality sensitive semi-supervised feature selection. *Neurocomputing*, 71(10-12):1842–1849, 2008.

[105] Zheng Zhao and Huan Liu. Semi-supervised feature selection via spectral analysis. In *SDM*, 2007.

[106] Zheng Zhao and Huan Liu. Spectral feature selection for supervised and unsupervised learning. In *ICML*, pages 1151–1157, 2007.

[107] Zheng Zhao, Lei Wang, and Huan Liu. Efficient spectral feature selection with minimum redundancy. In *AAAI*, 2010.

[108] Zhen zhong Lan, Lei Bao, Shoou-I Yu, Wei Liu, and Alexander G. Hauptmann. Double fusion for multimedia event detection. In *MMM*, pages 173–185, 2012.

[109] Xiaojin Zhu. Semi-supervised learning literature survey. *Technical Report 1530, University of Wisconsin, Madison*, 2007.

[110] Xiaojin Zhu, Zoubin Ghahramani, and John D. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, pages 912–919, 2003.

[111] Yueting Zhuang, Yi Yang, and Fei Wu. Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval. *IEEE Transactions on Multimedia*, 10(2):221–229, 2008.