MULTI-MODAL LEARNING WITH GENERALIZABLE NONLINEAR DIMENSIONALITY REDUCTION

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

SEMİH KAYA

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
ELECTRICAL AND ELECTRONICS ENGINEERING

MAY 2019

Approval of the thesis:

MULTI-MODAL LEARNING WITH GENERALIZABLE NONLINEAR DIMENSIONALITY REDUCTION

submitted by **SEMİH KAYA** in partial fulfillment of the requirements for the degree of **Master of Science in Electrical and Electronics Engineering Department, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. İlkay Ulusoy
Head of Department, **Electrical and Electronics Engineering** _____

Assist. Prof. Dr. Elif Vural
Supervisor, **Electrical and Electronics Engineering, METU** _____

Examining Committee Members:

Prof. Dr. Abdullah Aydın Alatan
Electrical and Electronics Eng., METU _____

Assist. Prof. Dr. Elif Vural
Electrical and Electronics Eng., METU _____

Prof. Dr. İlkay Ulusoy
Electrical and Electronics Eng., METU _____

Assist. Prof. Dr. Mustafa Mert Ankaralı
Electrical and Electronics Eng., METU _____

Assoc. Prof. Dr. Seniha Esen Yüksel
Electrical and Electronics Eng., Hacettepe University _____

Date: 14.05.2019

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Semih Kaya

Signature :

ABSTRACT

MULTI-MODAL LEARNING WITH GENERALIZABLE NONLINEAR DIMENSIONALITY REDUCTION

Kaya, Semih
M.S., Department of Electrical and Electronics Engineering
Supervisor: Assist. Prof. Dr. Elif Vural

May 2019, 64 pages

Thanks to significant advancements in information technologies, people can acquire various types of data from the universe. This data may include multiple features in different domains. Widespread machine learning methods benefit from distinctive features of data to reach desired outputs. Numerous studies demonstrate that machine learning algorithms that make use of multi-modal representations of data have more potential than methods with single modal structure. This potential comes from the mutual agreement of modalities and the existence of additional information. In this thesis, we introduce a multi-modal supervised learning algorithm to represent the data in lower dimensions. We intend to increase within-class similarity and between-class discrimination for intra- and inter-modal exemplars by a generalizable nonlinear interpolator, which satisfies Lipschitz continuity. In order to measure the performance of the proposed supervised learning algorithm, we have conducted several multi-modal face recognition and image-text retrieval experiments on frequently used multi-modal data sets in the literature and achieved quite satisfactory classification and retrieval accuracy in comparison with existing multi-modal learning approaches. These exper-

imental findings suggest that the incorporation of the generalizability of the embedding to the whole ambient space and unseen test data in the learning objective yields promising performance gains in multi-modal representation learning.

Keywords: Cross-modal learning, multi-view learning, cross-modal retrieval, nonlinear embedding, RBF interpolators.

ÖZ

GENELLENEBİLİR DOĞRUSAL OLMAYAN BOYUT DÜŞÜRME İLE ÇOKLU MODALİTE ÖĞRENME

Kaya, Semih
Yüksek Lisans, Elektrik ve Elektronik Mühendisliği Bölümü
Tez Yöneticisi: Dr. Öğr. Üyesi. Elif Vural

Mayıs 2019 , 64 sayfa

Bilgi teknolojilerindeki önemli ilerlemeler sayesinde insanlar evrenden çeşitli türde veriler elde edebilmektedir. Bu veriler farklı alanlarda farklı öznitelikler barındırabilmektedir. Yaygın yapay öğrenme yöntemleri istenilen çıktılara ulaşmak için verinin farklı özniteliklerinden faydalanmaktadır. Çok sayıda çalışma, çoklu modalite gösterimleri kullanan yapay öğrenme algoritmalarının tekli modalite algoritmalarına göre daha yüksek potansiyele sahip olduğunu göstermiştir. Bu potansiyel modalitelerin uyumluluğu ve birbirlerine göre içerdikleri ilave bilgiden gelmektedir. Bu tezde, veriyi daha düşük boyutlarda gösterebilmek için gözetimli bir çoklu modalite öğrenme algoritması geliştirilmiştir. Genellenebilir, doğrusal olmayan, Lipschitz devamlılığını sağlayan bir interpolasyon ile modalite içi ve modaliteler arası örnekler için aynı sınıf içerisindeki benzerliğin ve sınıflar arası ayrışımın arttırılması hedeflenmiştir. Önerilen gözetimli öğrenme algoritmasının performansını ölçmek için literatürde sıkça kullanılan çoklu modalite veri kümeleri üzerinde çeşitli çoklu modalite yüz tanıma ve görüntü-metin erişim deneyleri gerçekleştirilmiş, önerilen yöntem ile var olan çoklu modalite öğrenme yaklaşımlarına kıyasla oldukça tatmin edici sınıflandırma ve erişim

doğruluklarına ulaşılmıştır. Bu deney bulguları, öğrenmede kullanılan amaç fonksiyonuna gömülümün bütün çevresel uzaya ve önceden görülmemiş test verilerine genellenebilirliğinin dahil edilmesinin çoklu modalite gösterim öğreniminde geleceği parlak performans kazanımları sağladığına işaret etmektedir.

Anahtar Kelimeler: Çapraz modalite öğrenme, çoklu görü öğrenme, çapraz modalite erişim, doğrusal olmayan gömülüm, RBF interpolasyonları.

This thesis is dedicated to my supportive family, my cheerful fellows and my one and only wife

ACKNOWLEDGMENTS

This study would not have been conducted without precious supervision of Assist. Prof. Elif Vural. She mentored me from the beginnig to the end of my research in order to understand the cross-modal learning world, construct my hypothesis and conduct trustful experiments.

Nobody has been happier than my family for this study. I am appreciated with their supports. I especially would like to thank my inspring, lovely wife, İrem.

The last but not the least, I am also grateful to my "Mellon"s [1]. They always want to entertain me in spite of the difficulties of life. High-school friends, university collegues, workmates and others... So glad I have you!

[1] An Elvish word implies "friend"

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

BLM	Bilinear Model
CCA	Canonical Correlation Analysis
EM	Expectation Maximization
GMA	Generalized Multiview Analysis
GMMFA	Generalized Multiview Marginal Fisher Analysis
GMLDA	Generalized Multiview Linear Dirichlet Allocation
GMPCA	Generalized Multiview Principal Component Analysis
JFSSL	Joint Feature Selection and Subspace Learning
KNN	K Nearest Neighborhood
LDA	Linear Dirichlet Allocation
LiDAR	Ligth Detection and Ranging
MFDA	Multiview Fisher Discriminant Analysis
PCA	Principal Component Analysis
PLS	Partial Least Square
RBF	Radial Basis Function
SAR	Synthetic Aperture Radar
SVM	Support Vector Machine

NOMENCLATURE

diag(.)	Diagonal matrix which is constructed from the input
Tr(.)	Trace of a matrix

CHAPTER 1

INTRODUCTION

1.1 Motivation

Machine learning is a technique that enables computing systems to automatically generate and develop models from available data. A computing system does not need any explicit programming or human assistance for producing a model [5]. In the past decades, data acquisition has become easier for a computing system with the help of growing sensor technology. As a result of this, companies have started to make considerable investments and construct big data centres in order to deal with huge amounts of data. At this point, machine learning becomes a powerful tool for the smart data analysis. It helps data analysers to have solutions that solicit less money, less time and less energy cost for their specific problems [6].

Machine learning takes fundamental part in various applications in different areas. Security systems frequently combine face, speech, iris or fingerprint recognition solutions to build more secure access control mechanisms [6]. The main purpose of a security system is to discriminate people or objects using image, text, audio or video data. For instance, a smart security camera system can decide whether a known terrorist tries to enter a building stealthily if it is well-trained through images of this terrorist. Additionally, popular web search, social media, streaming, shopping, mailing applications on the Internet widely use machine learning. Google tries to generate most relevant search results for users through previously searched contents. YouTube suggests videos that are highly correlated with previously liked videos by a user. Spotify has similar features for audio streams. Advertisement industry also uses machine learning to meet people's needs by using their Internet activities. Detection of spam

e-mails from their contents is a crucial task for an e-mail application [6]. This is also highly related to personal or corporate security. Finance sector should be grateful to machine learning because of the fact that portfolio and risk management applications, insurance systems and smart trading systems in stock market contain advanced machine learning algorithms inside [7]. In addition to security systems, Internet applications and finance, which widely apply machine learning, medical applications also benefit from it. According to a research conducted by Stanford University, a kind of skin cancer can be detected earlier using machine learning, in comparison with traditional methods [8]. This study helps skin cancer patients to recover at the beginning stages of the disease. Furthermore, human genome studies are thankful to machine learning.

Although machine learning is a powerful tool for different applications, the following issues needed to be considered in the literature:

- In order to deal with big training data, relevant and irrelevant features need to be separated. This is also known as **"feature selection"**.
- Describing sample spaces with as fewest features as possible is important for machine learning algorithms. It is emphasized as **"dimensionality reduction"**.
- Machine learning algorithms may require to protect training data characteristics when they use **"linear/nonlinear projection"** techniques. Linear projection is easy to apply but nonlinear projection may provide extra capabilities to the machine learning algorithm even if its computational load is higher.
- Real world training sets may not contain the desired output information. If the desired output information is known, it is called as **"supervised learning"**. Otherwise, it is called as **"unsupervised learning"**. If some training samples do not have any label information, then it creates **"semi-supervised learning"** problems.
- Data is often available in distinct modalities such as image, text, audio or video representations. For this situation, single-modality-based machine learning algorithms may not be adequate to generate convenient classification or retrieval results. Intermodal relations can be indicated better through the implementation

of **"cross-modal learning"** algorithms.

This thesis aims to explore the cross-modal learning problem and propose a method to solve an optimization problem in order to obtain a nonlinear embedding of samples. For these reasons, a brief overview of multi-modal learning would be beneficial to understand why multi-modal learning methods are useful in data analysis.

As indicated earlier, the great progress of technology in the last decades has provided multi-modal data from multiple sources. For instance, face, fingerprint, signature or iris information can help to identify a person [2]. Moreover, any content of a web page can be illustrated via textual description, high-resolution images or videos [9]. Another example of multi-modal data can be given as the future transportation service suggestion in [1]. Figure 1.1 demonstrates available data sources for a suggested transportation system.

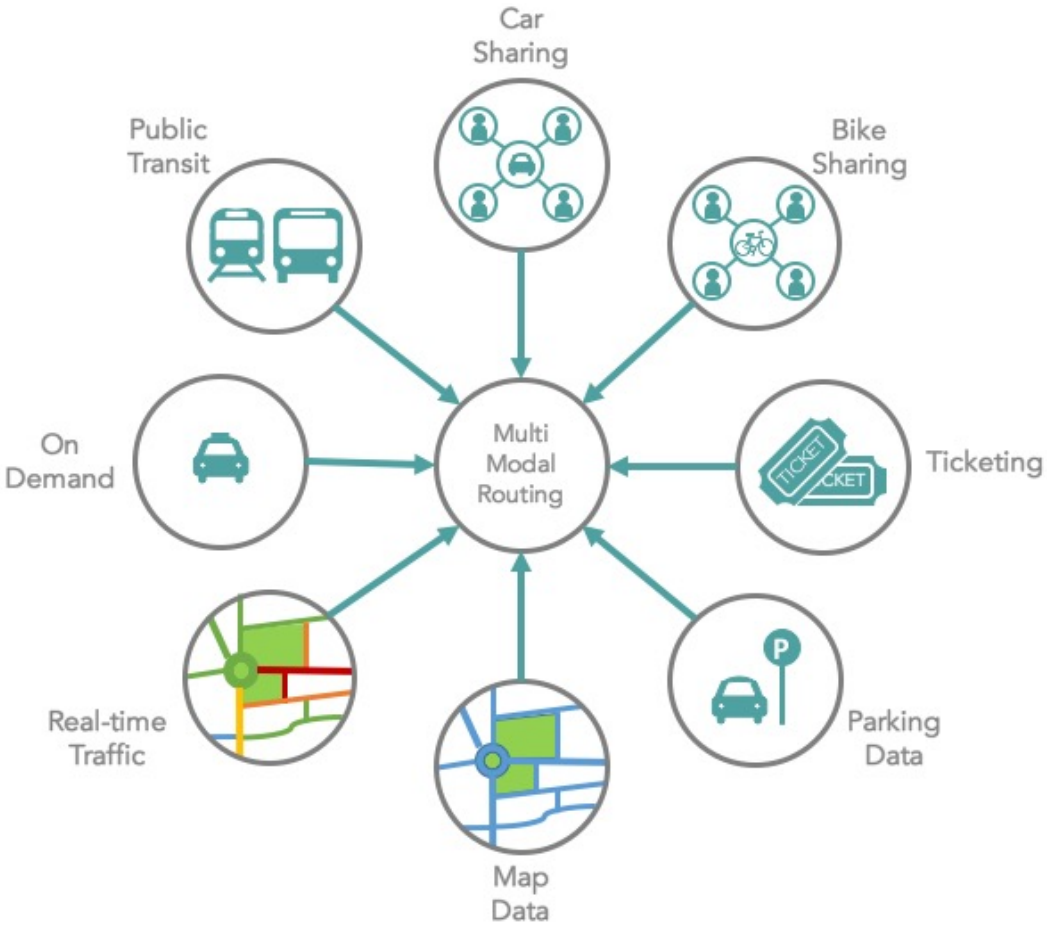


Figure 1.1: Multi-modal data needs of a future transportation system [1]

Remote sensing and Earth observation systems also use multimodality. Light Detection and Ranging (LiDAR) and Synthetic Aperture Radar (SAR) technologies provide topographic information such as elevation, 3-D structures of observed objects and surface properties. LiDAR samples are especially used for measuring distance to objects and generated through narrow pulsed laser. SAR images are constructed from the illumination and the backscattering of electromagnetic waves. More accurate topological information can be gathered via convenient fusion of LiDAR and SAR technologies [10].

In order to have successful weather forecasting, meteorological data is monitored with using rain gauges, radars, satellite-borne remote sensing devices. Hydrology, agriculture, and aeronautical services directly take advantage of rain, snow, fog and temperature information [11]. These suggest that multi-modal data can be quite useful for some areas.

All given examples demonstrate that there usually exist multi-modal representations of a sample. It provides a multi-modal learning algorithm to have various additional information that a single-modal learning algorithm cannot have. Despite the benefits of multi-modal data representations, there are also complicated challenges to be faced with. A multi-modal learning survey in [12] emphasizes 5 considerable phenomena that can be summarized as follows:

- Finding an efficient representation for multi-modal data without damaging its structure is a tough issue. For example, image modality can be represented as a signal, but text modality cannot. A text contains semantic structures to be taken into account.

- Translating one modality to another modality is also an issue that needs to be considered. An image can be described with more than one way.

- Obtaining direct intermodal relations leads to alignment problems for multi-modal data.

- Using joint information between modalities may not be simple. Incomplete data or noise may degrade the performance of joint information.

- Transferring knowledge from one modality to another modality, called as co-learning, is a challenging task due to missing annotations which may help the training of a computational model in the other modality.

It can be inferred that missing samples may result in unexpected multi-modal data representations. Therefore, a robust multi-modal learning algorithm requires an interpolator to get rid of the effects of missing samples. Another point to consider is that algorithms in the multi-modal learning literature endeavour to increase within-class similarity and between-class separation using training samples only. It causes missing samples to lead to more perturbation. Nevertheless, all sample space can be covered approximately with the help of a robust interpolator.

Vural and Guillemot [13] indicate that a radial basis function (RBF) may increase classification accuracy with the help of Lipschitz continuity. Their experimental results show that generalization of nonlinear supervised manifold algorithms may preserve intra class structure and ease inter class separability not only for training data but also for test data. Then, in [14], Örnek and Vural implement a nonlinear supervised dimension reduction algorithm through regular interpolators in a single modal domain. According to their results, the joint optimization of training embeddings and interpolator parameters via a smooth RBF interpolator produces better classification results than well-known single modal algorithms.

The main purpose of this study is to learn efficient representations for the analysis of multi-modal data samples. We propose a method that learns nonlinear projections of multi-view data into a common domain, which preserves within-class similarities while boosting between-class distances. First of all, an RBF interpolation mechanism is proposed to maximize intra class similarity and inter class separation. Then, interpolator parameters and training data embeddings to lower dimensions are obtained via an iterative optimization process for each modality. In the test phase, misclassification rates and retrieval performances are compared among several multi-modal algorithms on popular multi-modal datasets. Experimental results reveal that widely used linear multi-modal learning methods are not as successful as the proposed multi-modal nonlinear smooth embedding to training method in classification and retrieval applications.

The distinctive features of the proposed multi-modal learning algorithm in comparison with the current methods in the literature can be explained as follows:

- Nonlinear embedding
- Generalization to all sample space
- Incorporation of the interpolator regularity in the learning objective

The proposed method can be easily applicable for small multi-modal data sets. For some fields, the acquisition of the data is difficult or even not possible such as in medical and military applications. The proposed method might be a powerful tool for representing data in such domains.

1.2 Thesis Outline

The organization of the thesis is as follows: A few notable cross-modal learning techniques in the literature are explained briefly and related examples are provided in Chapter 2. Then, the definition of the multi-modal nonlinear supervised learning problem and the proposed solution to this problem are examined in detail in Chapter 3. In Chapter 4, various learning experiments conducted on different real-world datasets are presented in order to measure and evaluate the algorithm performance. At the end, a summary of the thesis study is given and possible recommendations are discussed to improve this study in Chapter 5.

CHAPTER 2

RELATED WORK

This chapter gives an overview of the trending approaches in the multi-modal learning literature. Section 2.1 explains the fundamental principles of multi-modal learning frameworks and gives a taxonomy of the current multi-modal learning approaches. Section 2.2 discusses co-training algorithms and their essential properties. After that Section 2.3 presents several multiple kernel learning methods and Section 2.4 demonstrates various subspace learning-based approaches in the literature. Section 2.5 describes deep multi-modal learning methods. Then, Section 2.6 briefly overviews the computation of nonlinear smooth embeddings in a single modality. This study has made a significant contribution to our work. Finally, Section 2.7 discusses the limitations of the previously mentioned algorithms and motivates the proposed multi-modal learning algorithm.

2.1 Fundamentals of Multi-modal Learning

Multi-modal learning algorithms have additional training samples for the same input data compared to single-modal learning algorithms. This may help the learning technique to achieve better performance in problems such as classification and regression if extra information is used properly. Therefore, inter modality relationships should be indicated efficiently. At this point, the “consensus” and “complementariness” concepts can be considered as the backbone of multi-modal learning algorithms. The consensus phenomenon mainly serves the maximization of cross modal data concurrency [2]. It decreases the disagreement between modalities and it provides improvement in the accuracy of learning. Furthermore, in some cases, some modalities may

include additional useful information. This can be expressed with the complementariness principle, which a single modal learning algorithm cannot have [2]. The consensus and complementariness properties of multi-modal data will be analysed further in the next sections of the chapter when various multi-modal learning algorithms are discussed. The multi-modal learning approaches in the literature can be classified as follows [2]:

- Co-training
- Multiple Kernel Learning
- Subspace Learning-based Approaches
- Deep Learning Methods

2.2 Co-training

Co-training allows each modality to be trained separately with correlated learners. Although the parameters of distinct modalities are learnt individually, validation data can be used in a back-propagation mechanism that results in smaller disagreement between modalities. It is an iterative learning process that provides consistent predictors to the model. Earlier co-training algorithms notice that under the sufficiency, compatibility and conditional independence assumptions, training data can be grouped successfully for a semi-supervised learning setting according to statements in [2]. Figure 2.1 illustrates a co-training mechanism for two modalities:

A study in [15] improved the co-training algorithm using Expectation Maximization (EM) technique to assign probabilistic labels to unlabelled samples. Each modality classifier iteratively uses the probabilities of class labels in Co-EM training. Therefore, Co-EM may achieve better training performance even if the conditional independence assumption is not met in a multi-modal dataset. Bayesian classifiers, which are commonly used in Co-EM frameworks, reduce classification errors.

A probabilistic model for Support Vector Machine (SVM) was constructed in [16] in order to get the powerful sides of the Co-EM algorithm. This leads to better classification results for co-training algorithms in many classification problems.

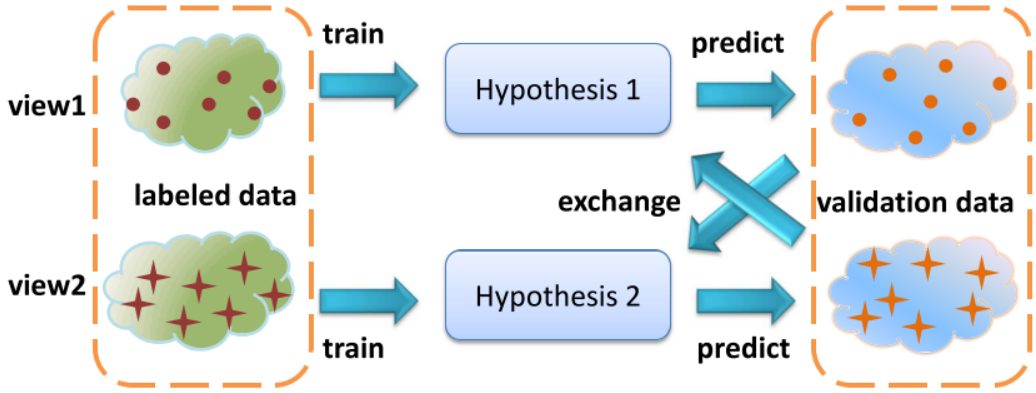


Figure 2.1: Co-training structure [2]

Co-regularization is another implementation of co-training. Assume two prediction functions f^1 and f^2 are defined on two hypothesis spaces H^1 and H^2 respectively. For given labelled samples $(x_i, y_i)_L$ and unlabelled samples $(x_i)_U$, co-regularization aims to solve the following optimization problem:

$$(f_*^1, f_*^2) = \min_{\substack{f^1 \in H^1 \\ f^2 \in H^2}} \gamma^1 \left\| f^1 \right\|_{H^1}^2 + \gamma^2 \left\| f^2 \right\|_{H^2}^2 + \mu \sum_{i \in U} [f^1(x_i) - f^2(x_i)]^2 + \sum_{i \in L} V(y_i, f(x_i)). \tag{2.1}$$

where $V(.)$ denotes loss function for labelled data predictions. Optimization problem states that norms of the prediction functions for each modality needs to be small. Co-regularization is provided with the third term in the optimization formula, which indicates mutual agreement on unlabeled samples. Using individual prediction functions for each modality, common predictor can be obtained as follows:

$$f_*(x) = \frac{1}{2}\left(f_*^1(x) + f_*^2(x) \right) \tag{2.2}$$

In the literature, there also exist co-regression algorithms which can be implemented in co-training way. A regression algorithm that uses two kNN regressors is presented to learn appropriate labels for unlabelled samples in [17]. Moreover, a study in [18] tries to minimize the following function for co-regression problems:

$$Q(f) = \sum_{v=1}^{M} \left[\sum_{x \in X_v} V\big(y(x), f_v(x)\big) + \nu \left\| f_v(.) \right\|^2 \right] + \lambda \sum_{u,v=1}^{M} \sum_{z \in Z} V\big(f_u(z), f_v(z)\big) \tag{2.3}$$

where u and v denote indexes of modalities from 1 to M, $f(.)$ states a prediction function, $V(.)$ indicates a loss function for labelled data predictions, z refers to unlabelled samples, x refers to all samples and $y(x)$ are given class labels.The first term in the equation gives the loss between available labelled samples and the generated predictors. The second term is used to decrease the norms of the predictors. The last term imposes mutual agreement between prediction functions.

A multi-modal clustering approach is presented in [19] through k-means clustering. First of all, k-means algorithm is applied to one modality. Then, class partition information of each modality is transferred to other modalities iteratively. After the loss function is minimized, multi-modal clustering is terminated.

Co-training technique is also used in graph based methods such as in [20]. Gaussian process is applied to Bayesian undirected graph representation of all modalities.

Other interesting co-training algorithms in the multi-modal data sets are proposed in [21], [22], and [23].

2.3 Multiple Kernel Learning

A single kernel function may not be adequate to attain the desired learning performance. In order to deal with this issue, multiple kernel combinations are proposed. Linear or nonlinear kernel combinations are quite popular techniques in multiple kernel learning [2]. Figure 2.2 illustrates the kernel combination process.

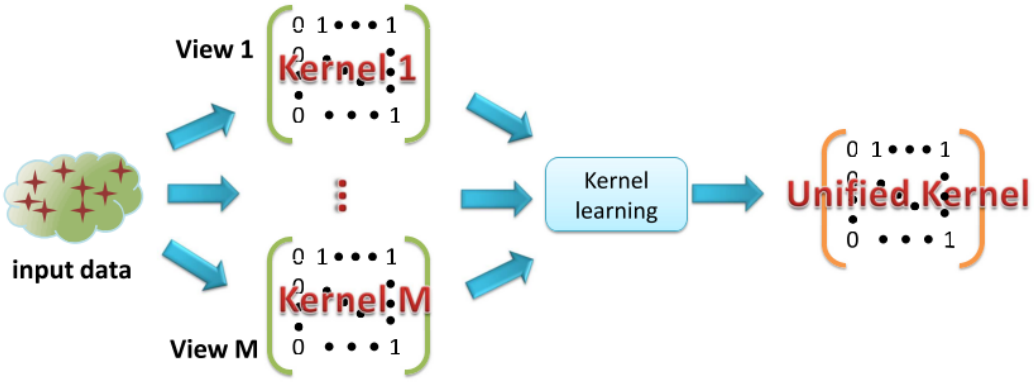


Figure 2.2: Multiple kernel learning structure [2]

For samples x_i, x_j, a unified kernel $K(x_i, x_j)$, kernel functions $K_k(x_i, x_j)$ and kernel weights d_k with k from modality 1 to modality M, multi-modal kernel combinations can be produced with various methods [2]. Linear combination methods can be expressed as follows:

$$\textbf{Direct summation: } K(x_i, x_j) = \sum_{k=1}^{M} K_k(x_i, x_J) \tag{2.4}$$

$$\textbf{Weighted summation: } K(x_i, x_j) = \sum_{k=1}^{M} d_k K_k(x_i, x_J) \tag{2.5}$$

$$\textbf{Restricted: } K(x_i, x_j) = \sum_{k=1}^{M} d_k K_k(x_i, x_J), \text{ where } K \geq 0, tr(K) \leq c \tag{2.6}$$

$$\textbf{Locally combined: } K(x_i, x_j) = \sum_{k=1}^{M} d_k(x_i) K_k(x_i, x_J) d_k(x_j) \tag{2.7}$$

There are also nonlinear combination methods for multi-modal kernel functions according to [2]:

$$\textbf{Exponential: } K(x_i, x_j) = exp\big(- \sum_{k=1}^{M} d_k x_i^T A_k x_j \big) \tag{2.8}$$

$$\textbf{Power: } K(x_i, x_j) = \big(d_0 + \sum_{k=1}^{M} d_k x_i^T A_k x_j \big)^n \tag{2.9}$$

where A_k denotes the affinity matrix of each modality and it can be constructed with several ways to determine similarities.

Moreover, several multiple kernel learning approaches are presented in [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], and [34].

2.4 Subspace Learning

Subspace learning methods are based on finding suitable linear projections or transformations that align samples from different modalities. Figure 2.3 shows a common subspace representation for given data samples.

The well-known unsupervised subspace learning algorithm CCA (Canonical Correlation Analysis) maximizes the correlation between modalities [2]. For a given data set

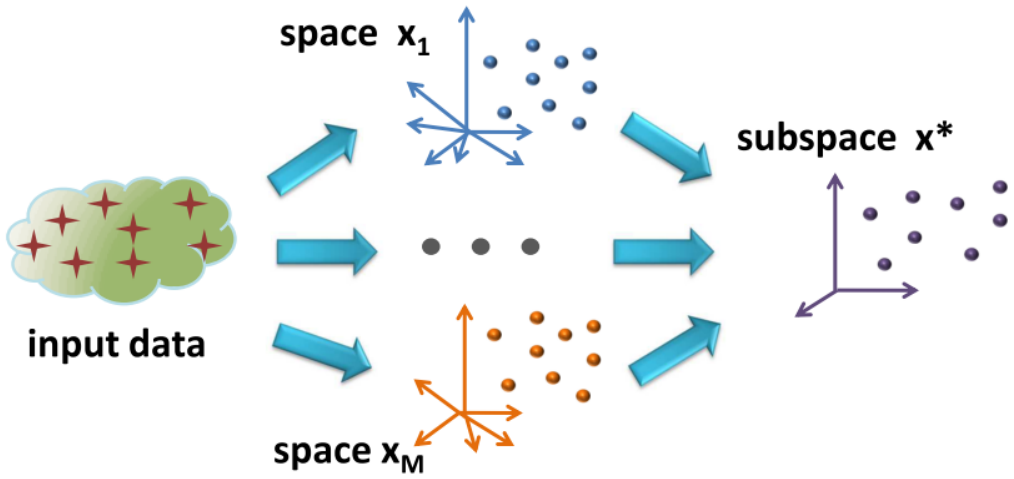


Figure 2.3: Subspace learning structure [2]

with two modalities $X = [x_1, ..., x_N]$ and $Y = [y_1, ..., y_N]$, the correlation coefficient (ρ) between two modalities can be defined as follows in [35]:

$$\rho = \frac{cov(w_x^T X, w_y^T Y)}{\sqrt{var(w_x^T X) var(w_y^T Y)}} = \frac{w_x^T C_{xy} w_y}{\sqrt{(w_x^T C_{xx} w_x)(w_y^T C_{yy} w_y)}} \tag{2.10}$$

The main purpose of CCA to maximize ρ, but the scales of the projection directions w_x, w_y do not affect ρ. Therefore, using (2.10), the optimization of CCA is reformulated by [35] as follows:

$$\max_{w_x, w_y} \rho = \max_{w_x, w_y} w_x^T C_{xy} w_y \text{ such that } w_x^T C_{xx} w_x = 1, w_y^T C_{yy} w_y = 1 \tag{2.11}$$

By applying Lagrange multipliers, the following equation can be obtained:

$$L(w_x, w_y, \lambda_x, \lambda_y) = w_x^T C_{xy} w_y - \frac{\lambda_x}{2}(w_x^T C_{xx} w_x - 1) - \frac{\lambda_y}{2}(w_y^T C_{yy} w_y - 1) \tag{2.12}$$

Computing the derivatives of (2.12) with respect to w_x and w_y produces the following outcomes:

$$C_{xy} w_y - \lambda_x C_{xx} w_x = 0 \tag{2.13}$$

$$C_{yx} w_x - \lambda_y C_{yy} w_y = 0 \tag{2.14}$$

Multiplying (2.13) with w_x^T and (2.14) with w_y^T from the left hand side results in the following expression:

$$\lambda_y w_y^T C_{yy} w_y - \lambda_x w_x^T C_{xx} w_x = 0 \longrightarrow \lambda_y - \lambda_x = 0 \longrightarrow \lambda_x = \lambda_y \tag{2.15}$$

w_y can be calculated for an invertible C_{yy} by indicating $\lambda_x = \lambda_y = \lambda$:

$$w_y = \frac{1}{\lambda} C_{yy}^{-1} C_{yx} w_x \tag{2.16}$$

Through (2.13) and w_y, which is given in (2.16), the following equation is obtained:

$$C_{xy} C_{yy}^{-1} C_{yx} w_x = \lambda^2 C_{xx} w_x \tag{2.17}$$

The solution of (2.17) is the eigenvector that corresponds the largest eigenvalue. Thus, w_x needs to be obtained at first and then be normalized. After that w_y can be calculated easily via (2.16). w_y can be normalized with the constraints in (2.11). Finally, the normalized projection vectors, w_x and w_y can be used to align the two modalities.

Alternative versions of CCA such as cluster CCA [36], multilabel CCA [37] and three view CCA [38] have been proposed to improve the performance for different data sets and various tasks; but all of them contain linear projections. Thereby, given CCA based algorithms may produce good results in multi-modal data sets that are suitable for linear transformations. There also exists a nonlinear extension of CCA, which is called as kernel CCA [2]. However, kernel CCA representations may face flexibility issues.

There are also other popular unsupervised learning algorithms like Bilinear Model (BLM) [39] and Partial Least Squares (PLS) [40] for multi-modal data sets. Their drawback is the inability to use the label information in case this is available.. On the other hand, supervised multi-modal learning algorithms utilize class representations so that more accurate results can be obtained. Multiview Fisher Discriminant Analysis (MFDA) is proposed to maximize the consensus of predicted class labels in different modalities [41]. Since it is suitable for data sets with two classes, an advanced version of MFDA is developed with a hierarchical architecture [42]. Moreover, a Latent Dirichlet Allocation (LDA) system is designed to discriminate multi-modal data samples by benefitting from domain information [43].

The study in [44] suggests that an ideal cross-modal learning algorithm should be supervised, generalizable, multi-modal, efficient, kernelizable and domain independent. For this reason, the authors of [44] introduce a generic method called as Generalized Multiview Analysis (GMA) to learn a common subspace from a multi-modal frame-

work. According to [44], the GMA approach starts with a joint optimization of two projection directions as follows:

$$\begin{aligned}[\hat{v}_1, \hat{v}_2] = \arg\max_{v_1, v_2} \; & v_1^T A_1 v_1 + \mu v_2^T A_2 v_2 \\ \text{s.t. } & v_1^T B_1 v_1 = v_2^T B_2 v_2 = 1\end{aligned} \tag{2.18}$$

where A_i and B_i with $i \in \{0, 1\}$ are square symmetric matrices that respectively represent within-class similarities and between-class differentiations. In (2.18), the parameter μ is used to balance the weights of two optimization terms. In [44], simplified version of (2.18) is obtained as follows by combining the constraints via a parameter γ:

$$\begin{aligned}[\hat{v}_1, \hat{v}_2] = \arg\max_{v_1, v_2} \; & v_1^T A_1 v_1 + \mu v_2^T A_2 v_2 \\ \text{s.t. } & v_1^T B_1 v_1 + \gamma v_2^T B_2 v_2 = 1, \text{ where } \gamma = \frac{tr(B_1)}{tr(B_2)}\end{aligned} \tag{2.19}$$

The GMA algorithm claims that the projections of the i^{th} samples in different modalities ought to become as close as possible. Denoting the projections as α and the samples as z, the projections can be calculated as in (2.20):

$$\alpha_1^i = v_1^T z_1^i \text{ and } \alpha_2^i = v_2^T z_2^i \tag{2.20}$$

In order to decrease the distances between multi-modal samples of the same class, the covariance between the samples should be maximized. The following optimization problem can be constructed to formulate this idea:

$$[\hat{v}_1, \hat{v}_2] = \arg\max_{v_1, v_2} v_1^T Z_1 Z_2^T v_2, \text{ where } Z_m = \left[z_m^1 \; z_m^2 \cdots z_m^j\right]. \tag{2.21}$$

Combining the individual problems in (2.19) and (2.21), the overall optimization problem can be obtained as follows:

$$\begin{aligned}[\hat{v}_1, \hat{v}_2] = \arg\max_{v_1, v_2} \; & v_1^T A_1 v_1 + \mu v_2^T A_2 v_2 + 2\beta v_1^T Z_1 Z_2^T v_2 \\ \text{s.t. } & v_1^T B_1 v_1 + \gamma v_2^T B_2 v_2 = 1\end{aligned} \tag{2.22}$$

Using matrix notation, optimization formula, which is given (2.22), can be written as

follows:

$$\begin{bmatrix} \hat{v}_1 \\ \hat{v}_2 \end{bmatrix} = \arg\max_{v_1, v_2} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}^T \begin{bmatrix} A_1 & \beta Z_1 Z_2^T \\ \beta Z_2 Z_1^T & \mu A_2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \\ \text{s.t.} \begin{bmatrix} v_1^T & v_2^T \end{bmatrix} \begin{bmatrix} B_1 & 0 \\ 0 & \gamma B_2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 1 \tag{2.23}$$

A more compact form of (2.23) is the following:

$$\hat{v} = \arg\max_{v} v^T A v \\ \text{s.t. } v^T B v = 1. \tag{2.24}$$

This leads to the generalized eigenvalue problem $\tilde{\boldsymbol{A}}\hat{\boldsymbol{v}} = \boldsymbol{\lambda}\tilde{\boldsymbol{B}}\hat{\boldsymbol{v}}$ with square symmetric matrices $\tilde{A}$ and $\tilde{B}$. The solution of this closed form is the eigenvector which is relevant with the largest eigenvalue of $B^{-1}A$ when B is an invertible matrix. If B is not full rank, then dimension reduction or regularization techniques are needed to be applied before the GMA algorithm.

Various linear projection methods in the literature can be derived from the GMA structure according to [44]. These projections can be computed with the derivations in Table 2.1:

Table 2.1: How to derive well known algorithms through GMA

Algorithm	$\boldsymbol{A_i}$	$\boldsymbol{B_i}$	$\boldsymbol{Z_i}$	$\boldsymbol{W_i}$
GMPCA	$X_i W_i X_i^T$	I	X_i	I_i/N_i
CCA	0	$X_i W_i X_i^T$	X_i	I_i/N_i
BLM	$X_i W_i X_i^T$	I	X_i	I_i/N_i
PLS	0	I	X_i	not used
GMLDA	$X_i W_i X_i^T$	$X_i D_i X_i^T$	X_i or M_i	$[W_i^{kl}]$
GMMFA	$X_i(S_{bi} - W_{bi})X_i^T$	$X_i(S_{wi} - W_{wi})X_i^T$	X_i	not used

where;

I_i is the identity matrix for the modality i,

N_i is the number of samples in the modality i,

$$W_i^{kl} = \begin{cases} 1/N_i^c, & \text{if } X_i^k \text{ and } X_i^l \text{ are in the same class} \\ 0, & \text{otherwise} \end{cases}$$

N_i^c is the number of samples from class c in the modality i,

M_i is the matrix which contains class means in its columns ,

$D_i = I - W_i$ and i denotes the modality index,

The within-class compression matrix is;

$$W_{wi}^{kl} = \begin{cases} 1, & \text{if } X_i^k \text{ and } X_i^l \text{ are in the same class} \\ 0, & \text{otherwise} \end{cases}$$

The between-class separation matrix is;

$$W_{bi}^{kl} = \begin{cases} 1, & \text{if } X_i^k \text{ and } X_i^l \text{ are in different classes} \\ 0, & \text{otherwise} \end{cases}$$

Cross-modal retrieval applications should deal with how to measure the relevance and how to select coupled features. Previously mentioned methods only measure the relevance, while the recent method, which is called as JFSSL (Joint Feature Selection and Subspace Learning), achieves feature selection and common subspace learning simultaneously [3]. The JFSSL algorithm tries to obtain linear transformations for different modalities by choosing relevant and irrelevant features while constructing a multi-modal graph. The optimization problem, which is defined for JFSSL, can be formulated as follows:

$$\min_{U_1,\cdots,U_m} \sum_{p=1}^{M} \left\| X_p^T U_p - Y \right\|_F^2 + \lambda_1 \sum_{p=1}^{M} \|U_p\|_{21} + \lambda_2 \Omega(U_1, \cdots, U_M) \tag{2.25}$$

In (2.25), U_p refers to projection matrices, X_p stands for the labelled data matrices, Y_p denotes the low-dimensional representations of X_p, $\Omega(.)$ is the loss function for the joint graph and p indicates the modality index from 1 to M. λ_1 and λ_2 are parameters

that are used for the regularization of the optimization terms.

In order to minimize the projection errors among different modalities, the first term is added to (2.25). The second optimization term describes the selection of relevant and redundant features through the l_{21} norm. The last term in (2.25) is defined for the multi-modal graph representation, which is generated from intra-modal and inter-modal similarity relationships. For modalities p and q, the inter-modal similarity can be shown with a matrix W^{pq} as follows:

$$W_{ij}^{pq} = \begin{cases} 1, & \text{if } x_i^p \text{ has similar semantics to } x_j^q \\ 0, & \text{otherwise} \end{cases} \tag{2.26}$$

Neighbouring samples in the original space need to be as close as possible after projections. Hence, the following intra-modal similarity matrix is introduced with a kNN graph:

$$W_{ij}^{p} = \begin{cases} \exp\big(- \left\| x_i^p - x_j^p \right\|^2 / 2\sigma^2 \big), & \text{if } x_i^p \text{ and } x_j^p \text{ are linked with kNN} \\ 0, & \text{otherwise} \end{cases} \tag{2.27}$$

Using (2.26) and (2.27), the overall similarity matrix can be written as follows:

$$W = \begin{bmatrix} \beta W^1 & W^{12} & \cdots & W^{1M} \\ W^{21} & \beta W^2 & \cdots & W^{2M} \\ \vdots & \vdots & \ddots & \vdots \\ W^{M1} & W^{M2} & \cdots & \beta W^M \end{bmatrix}, \text{ where } \beta \text{ balances two similarities.} \tag{2.28}$$

A joint graph can be constructed with the overall similarity and projected data samples as follows:

$$\Omega(U_1, \cdots, U_M) = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} W_{ij} \left\| f_i - f_j \right\|^2 = Tr(FLF^T) \tag{2.29}$$

where N indicates the number of samples, L represents the graph Laplacian matrix and can be calculated as $L = D - W$. The diagonal degree matrix, D, contains the sums of the rows in W. The projected data is $F = [U_1^T X_1 \cdots U_M^T X_M]$.

In order to obtain the projection vectors iteratively, the l_{21} norm can be relaxed with an auxiliary vector r_p which has the ith element $r_p^i = \big(2\sqrt{\left\| u_p^i \right\|_2^2 + \epsilon}\big)^{-1}$. Thus, the

overall optimization can be written as follows:

$$\min_{U_1,\cdots,U_m} \sum_{p=1}^{M} \left\| X_p^T U_p - Y \right\|_F^2 + \lambda_1 \sum_{p=1}^{M} Tr\left(U_p^T R_p U_p\right) + \lambda_2 \sum_{p=1}^{M} \sum_{q=1}^{M} Tr\left(U_p^T X_p L_{pq} X_q^T U_q\right), \text{ where } R_p = diag(r_p) \tag{2.30}$$

The term L_{pq} in (2.30) implies the graph Laplacian matrix. It is obtained from the inter-modal similarity matrix, W_{pq}, which is introduced in (2.26).

Figure 2.4 illustrates the joint feature selection and subspace learning mechanism on the multi-modal data set, which includes image and text samples.

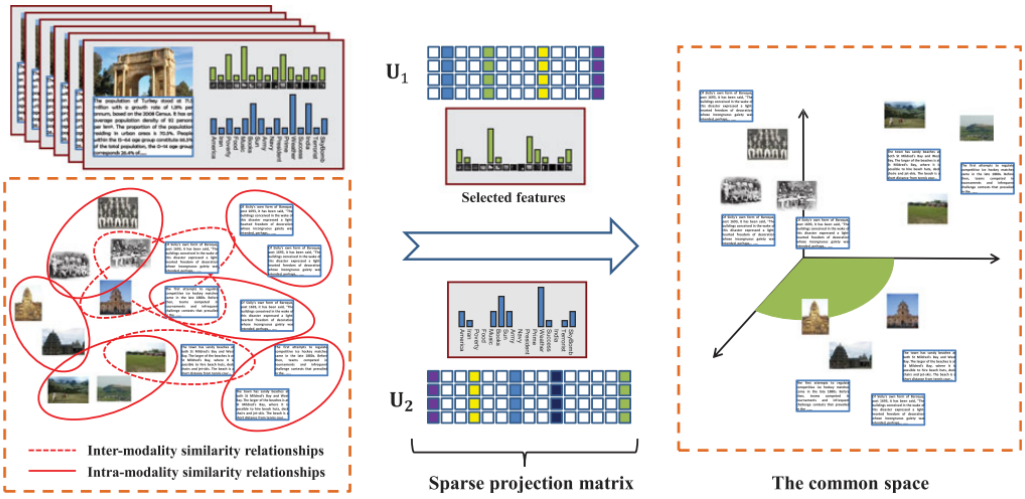


Figure 2.4: General architecture of the JFSSL algorithm [3]

Moreover, another graph based method was presented in [4]. This method endeavours to obtain a low-dimensional smooth embedding of all modalities at the same time. The objective function of the multi-modal spectral embedding can be stated as follows:

$$\arg\min_{Y,\alpha} \sum_{i=1}^{M} \alpha_i^r Tr\left(Y L_n^{(i)} Y^T\right) \text{ s.t. } YY^T = I \text{ and } \sum_{i=1}^{M} \alpha_i = 1, \tag{2.31}$$

where $\alpha_i \geq 0$ for the modality index i from 1 to M.

Y points out the low-dimensional embedding of the multi-modal data X. $L_n^{(i)}$ indicates the normalized graph Laplacian matrix of the ith modality and can be obtained

as $L_n^{(i)} = I - \left(D^{(i)}\right)^{(-1/2)} W^{(i)} \left(D^{(i)}\right)^{(-1/2)}$. The similarity matrix is computed through the Gaussian kernel function.

Figure 2.5 explains the overall structure of multi-modal spectral embedding with graph Laplacians.

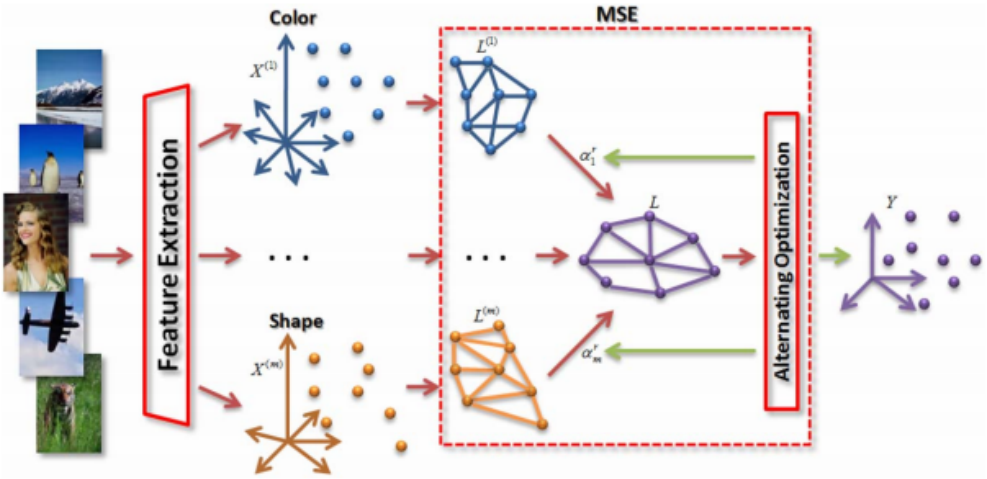


Figure 2.5: Multi-modal spectral embedding overview [4]

In the literature, there exists a multi-modal metric learning algorithm [45], which aims to keep similar pairs close and dissimilar pairs away from each other while maintaining the inner characteristics. Furthermore, a multi-modal matrix factorization technique [46] was proposed to solve the following optimization problem:

$$\min_{\substack{U^{(v)},V^{(v)} \\ V^* \geq 0}} \sum_{v=1}^{M} \left\| X^{(v)} - U^{(v)} \left(V^{(v)}\right)^T \right\|_F^2 + \sum_{v=1}^{M} \lambda_v \left\| V^{(v)} Q^{(v)} - V^* \right\|_F^2 \tag{2.32}$$

where v is the modality index from 1 to M, $Q^{(v)}$ is produced with the constraint of the column sums of $U^{(v)}$, the term λ_v is used as the balancing parameter in the optimization. The main goal of this algorithm is to find the best approximation of multi-modal data samples such that $X^{(v)} \approx U^{(v)} \left(V^{(v)}\right)^T$. The approximation provides the reduced data matrix, V^*.

Additionally, several subspace learning algorithms were presented in [47], [48], [49] and [50] in order to decrease the effects of the noise, which are caused from real world data samples.

2.5 Deep Learning

Remarkable developments in hardware technologies, especially in GPU (Graphical Power Unit), has allowed researchers to build complex computational models that process data in multiple layers [51]. These powerful models produce successful classification, recognition and analysis results on big data sets with high dimensions.

In [52], The Term Frequency - Inverse Document Frequency (TFIDF) feature, the context feature, the low-dimensional graph node representation feature and the timestamp feature are used as inputs to a Multi-Entry Neural Network (MENET) in order to detect geolocation of Twitter users. Furthermore, correlation learning errors and representation learning errors are minimized concurrently thanks to hidden representations of different modalities in [53].

A study in [54] suggests that the proposed convolutional neural network and the natural language model efficiently obtain label information even if the input data is noisy. Another research in [55] claims that in order to make hidden representations aligned for two different modalities, regularized cross-modal convolutional networks can be implemented.

Recently developed deep learning algorithms for multi-modal frameworks can be found in [56], [57], [58], [59] and [60].

2.6 Nonlinear Embeddings with Smooth Interpolators

In [14], a nonlinear supervised embedding technique is proposed to provide the Lipschitz continuity to the interpolator, decrease the distances between neighbouring samples from the same classes, and increase the separation between samples from different classes. The proposed optimization problem is shown in (2.33)

$$\begin{aligned} \min_{Y,\sigma} Tr\Big(Y^T L_w Y\Big) - \mu_1 Tr\Big(Y^T L_b Y\Big) + \mu_2 Tr\Big(Y^T \Psi^{-2} Y\Big) + \mu_3/\sigma^2, \\ \text{s.t. } Y^T Y = I. \end{aligned} \tag{2.33}$$

According to the experimental results in [14], the proposed learning method has been quite successful compared to other single modality learning approaches. Therefore,

we consider this algorithm as a starting point and extend this framework to a multi-modal setting.

2.7 Discussion

The alignment of different modalities via linear projections or transformations as in subspace learning might have limitations in real data sets where different modalities are weakly linked. In particular, when the data from different modalities have significantly dissimilar geometric structures, linear methods may fall short of providing effective joint representations since they mostly conserve the geometry of the individual modalities. Kernel methods provide nonlinear representations that may improve some of these shortcomings; however, the resulting representations might still lack in flexibility in certain scenarios. Deep learning algorithms using cross-modal autoencoders and CNNs provide powerful nonlinear representations achieving impressive performance in retrieval problems [61], [55], [53]. Meanwhile, these methods often need much larger training data sets.

The recent study in [13] focuses on nonlinear dimensionality reduction and proposes generalization bounds on the performance of classification. It is shown that in addition to increasing the separation between different classes and preserving the within-class similarity, another important condition that must be satisfied for successful generalization to new test data is that the interpolation function extending the nonlinear embedding to the whole data space must be sufficiently regular. The regularity of the interpolator is characterized in terms of its Lipschitz continuity in [13]. These results have been successfully applied in the single-modal supervised manifold learning problem in [14].

In this thesis, we build on the theoretical results in [13] and propose a nonlinear multi-modal dimensionality reduction algorithm for cross-modal classification and retrieval, which aims to achieve flexibility and robustness in the learning via the nonlinearity of the representations. We compute a nonlinear embedding of the training samples from different modalities, where we aim to increase the separation between different classes, preserve the within-class geometric structure of each modality, and

also align the same-class samples from different modalities. The nonlinear embedding of the training samples is extended to the whole space via an RBF (Radial Basis Function) interpolator. In line with the theoretical findings of [13], we consider the Lipschitz regularity of the interpolation function in our optimization objective as well. The resulting objective function is minimized with an iterative optimization procedure, where the nonlinear embedding coordinates are learnt jointly with the Lipschitz-continuous interpolator parameters. Experimental results in multi-view face recognition and image-text cross-modal retrieval applications show that the proposed method gives quite satisfactory performance in comparison with state-of-the-art algorithms.

CHAPTER 3

PROPOSED METHOD

3.1 Notation and Theoretical Background

We first set the notation and briefly summarize the theoretical findings underlying our method. Let $X^{(p)} \in \mathbb{R}^{N^{(p)} \times d^{(p)}}$ denote the training data matrix of modality p, each row of which is a data sample $x_i^{(p)}$. Here $N^{(p)}$ is the number of samples in modality p, and $d^{(p)}$ is the dimension of the samples in modality p. The vectors $x_i^{(1)}, \ldots, x_i^{(p)}$ are considered to represent observations of the same data sample x_i under different modalities. Uppercase letters (e.g. X) and lowercase letters (e.g. x) respectively indicate matrices and vectors. The notation tr(A) stands for the trace of a matrix A, and A_{ij} indicates its entry in the i-th row and j-th column. $C(x_i^{(p)})$ refers to the class label of the sample $x_i^{(p)}$.

Given the training samples $X^{(p)}$ from modalities $p = 1, \ldots, V$, we would like to compute embeddings $Y^{(p)} \in \mathbb{R}^{N^{(p)} \times m}$ of the training samples, such that each training sample $x_i^{(p)} \in \mathbb{R}^{d^{(p)}}$ is mapped to a vector $y_i^{(p)} \in \mathbb{R}^m$ in a common space of dimension m. Our main purpose is to find an embedding that can be successfully generalized to initially unavailable test samples of unknown class for classification or retrieval purposes. We propose to generalize the embedding of the training samples to the whole data space through interpolation functions $f^{(p)} : \mathbb{R}^{d^{(p)}} \to \mathbb{R}^m$, so that each training sample is mapped to its embedding as $f^{(p)}(x_i^{(p)}) = y_i^{(p)}$.

The study in [13] proposes a generalization bound for supervised nonlinear dimensionality reduction in a single-modal setup. Let us first recall the definition of a Lipschitz-continuous function.

Definition 1. *A function $f : \mathbb{R}^d \to \mathbb{R}^m$ is Lipschitz continuous with constant $L > 0$ if*

for any $u, v \in \mathbb{R}^d$

$$\|f(u) - f(v)\| \leqslant L \|u - v\|.$$

Assume a function f is differentiable, satisfies the Lipschitz continuity and defined as $f : \mathbb{R} \to \mathbb{R}$. Then, Lipschitz condition can be expressed as follows according to [62]:

$$\begin{aligned} \|f(u) - f(v)\| \leqslant L \|u - v\| &\Longrightarrow \left| \frac{f(u) - f(v)}{u - v} \right| \leqslant L \\ &\Longrightarrow \left| \frac{f(u+h) - f(u)}{h} \right| \leqslant L \text{ where } v = u + h \end{aligned}$$

If $h \to 0$, then $f'(u) \leqslant L$. It means that the Lipschitz constant bounds the derivative of the function. However, this limit cannot be established if the function is not differentiable. Even if the function is not differentiable, the Lipschitz continuity theorem indicates that the function, which satisfies the Lipschitz continuity, cannot highly fluctuate.

The study in [13] considers a single-modal setting where the training sample $x_i \in \mathbb{R}^d$ is mapped to its embedding $y_i \in \mathbb{R}^m$ in a supervised way. A test sample x is then classified by first mapping it to $\mathbb{R}^m$ with an interpolation function $f : \mathbb{R}^d \to \mathbb{R}^m$, and then finding the estimate $\hat{C}(x)$ of its class label via nearest-neighbour classification in $\mathbb{R}^m$. The main result in [13] is summarized as follows:

Theorem 1. *Let* $X = \{x_i\}_{i=1}^N \subset \mathbb{R}^d$ *be a set of training samples and* $Y = \{y_i\}_{i=1}^N$ *be an embedding of* X *in* $\mathbb{R}^m$*. Let* $\gamma > 0$ *and* A_δ *be parameters such that*

$$\begin{aligned} &\|y_i - y_j\| < A_\delta, \text{ if } \|x_i - x_j\| \leqslant 2\delta \text{ and } C(x_i) = C(x_j), \\ &\|y_i - y_j\| > \gamma, \text{ if } C(x_i) \neq C(x_j). \end{aligned}$$

For given $\epsilon > 0$ *and* $\delta > 0$*, let the Lipschitz-continuous interpolator* f *with Lipschitz constant* L *satisfy*

$$L\delta + \sqrt{m}\epsilon + A_\delta \leqslant \frac{\gamma}{2}. \tag{3.1}$$

Then the probability of correctly classifying a test sample x *from class* c *is lower bounded as*

$$P\left(\hat{C}(x) = c\right) \geqslant 1 - e^{-O(N)} - 2me^{-O\left(\frac{N\epsilon^2}{L^2\delta^2}\right)}. \tag{3.2}$$

This result intuitively suggests the following: For successful classification, a nonlinear embedding should have sufficiently small distance (A_δ) between nearby samples from the same class, and sufficiently large separation (γ) between different classes. Meanwhile, the Lipschitz constant L of the interpolator should also be sufficiently small so that the condition in (3.1) can be satisfied. Under these conditions, the probability of correct classification exponentially approaches 1 as the number of samples increases. Although Theorem 1 addresses a single-modal setting, the same principles apply to multi-modal problems as well, and we thus consider its findings in our learning objective.

3.2 Problem Formulation

We can now formulate our multi-modal learning problem in the light of Theorem 1, where we have the following goals:

Lipschitz regularity of the interpolator. We extend the embeddings of training samples to the data space with RBF interpolation functions of the form $f^{(p)}(x) = [f_1^{(p)}(x) \dots f_m^{(p)}(x)]$ for each modality $p = 1, \dots, V$, where the k-th component of $f^{(p)}(x)$ is

$$f_k^{(p)}(x) = \sum_{i=1}^{N^{(p)}} C_{ik}^{(p)} \, \phi^{(p)}(\|x - x_i^{(p)}\|). \tag{3.3}$$

Here $\phi^{(p)}(r) = e^{-r^2/(\sigma^{(p)})^2}$ is a Gaussian RBF kernel with scale parameter $\sigma^{(p)}$ and $C_{ik}^{(p)}$ are the interpolator coefficients. A Lipschitz constant for Gaussian RBF interpolators has been proposed in [14], which implies that $f^{(p)}(x)$ is Lipschitz-continuous with constant

$$L^{(p)} = \sqrt{2} e^{-\frac{1}{2}} \sqrt{N^{(p)}} (\sigma^{(p)})^{-1} \left\| C^{(p)} \right\|_F \tag{3.4}$$

where $C^{(p)}$ is the coefficient matrix with entries given by $C_{ik}^{(p)}$, and $\|\cdot\|$ denotes the Frobenius norm. The interpolator coefficients can be easily obtained by fitting the embeddings $Y^{(p)}$ to the training data $X^{(p)}$ as follows:

$$C^{(p)} = (\Psi^{(p)})^{-1} Y^{(p)} \tag{3.5}$$

where $\Psi^{(p)}$ is the matrix consisting of the values of the RBF kernels through

$$\Psi_{ij}^{(p)} = \phi^{(p)}(\|x_i^{(p)} - x_j^{(p)}\|). \tag{3.6}$$

Hence, in order to ensure that the interpolators of all modalities have small Lipschitz constants, for each modality $p = 1, \ldots, V$, we propose to minimize

$$\left\|C^{(p)}\right\|_F^2 = \|(\Psi^{(p)})^{-1}Y^{(p)}\|_F^2 = Tr\Big(Y^{(p)^T}(\Psi^{(p)})^{-2}\, Y^{(p)}\Big) \tag{3.7}$$

in addition to the minimization of the kernel scale term $(\sigma^{(p)})^{-1}$.

Within-class compactness and between-class separation. The total weighted distance between the embeddings of samples from the same class is commonly formulated as

$$\sum_{i,j=1}^{N^{(p)}} (W_w^{(p)})_{ij}\, \|y_i^{(p)} - y_j^{(p)}\|^2 = Tr\Big(Y^{(p)^T} L_w^{(p)} Y^{(p)}\Big) \tag{3.8}$$

in the manifold learning literature. Here $W_w^{(p)}$ is a weight matrix with entries representing the similarity between the data samples as follows:

$$(W_w^{(p)})_{ij} = \begin{cases} \exp\Big(-\frac{\|x_i^{(p)} - x_j^{(p)}\|^2}{(\theta^{(p)})^2}\Big), & \text{if } x_i^{(p)} \text{ and } x_j^{(p)} \text{ are from the same class,} \\ 0, & \text{otherwise} \end{cases} \tag{3.9}$$

where $\theta^{(p)}$ is a scale parameter. Defining the diagonal degree matrix $D_w^{(p)}$ with i-th diagonal entry given by $\sum_j (W_w^{(p)})_{ij}$, the matrix $L_w^{(p)}$ denotes the within-class Laplacian given by:

$$L_w^{(p)} = D_w^{(p)} - W_w^{(p)} \tag{3.10}$$

The term in (3.8) hence imposes nearby samples $x_i^{(p)}$, $x_j^{(p)}$ from the same class to be mapped to nearby coordinates. The graph Laplacian term is frequently used in the manifold learning with unsupervised ([63]) and supervised ([64], [65], [66], [67] ,and [68]) manner.

Similarly, in order to enhance the separation between the samples from different classes, for each modality $p = 1, \ldots, V$, we maximize

$$\sum_{i,j=1}^{N^{(p)}} (W_b^{(p)})_{ij}\, \|y_i^{(p)} - y_j^{(p)}\|^2 = Tr\Big(Y^{(p)^T} L_b^{(p)} Y^{(p)}\Big) \tag{3.11}$$

where $L_b^{(p)} = D_b^{(p)} - W_b^{(p)}$ is the between-class Laplacian matrix obtained from the weight matrix $W_b^{(p)}$. The only nonzero entries of $W_b^{(p)}$ are given by $(W_b^{(p)})_{ij} = 1$ when $x_i^{(p)}$ and $x_j^{(p)}$ are from different classes, and $D_b^{(p)}$ is the diagonal between-class degree matrix with $(D_b^{(p)})_{ii} = \sum_j (W_b^{(p)})_{ij}$.

Alignment of different modalities. In learning multi-modal representations, an important purpose is to align different modalities in a suitable way. In computing nonlinear embeddings, we aim to map similar samples from different modalities p, q to nearby points by minimizing:

$$\sum_{i=1}^{N^{(p)}} \sum_{j=1}^{N^{(q)}} \left\| y_i^{(p)} - y_j^{(q)} \right\|_2^2 (W_w^{(pq)})_{ij} = Tr\Big({Y^{(p)}}^T L_w^{(pq)} Y^{(p)} \Big) \tag{3.12}$$

while the separation between samples from different classes from modalities p, q are increased by maximizing:

$$\sum_{i=1}^{N^{(p)}} \sum_{j=1}^{N^{(q)}} \left\| y_i^{(p)} - y_j^{(q)} \right\|_2^2 (W_b^{(pq)})_{ij} = Tr\Big({Y^{(p)}}^T L_b^{(pq)} Y^{(p)} \Big) \tag{3.13}$$

Here the matrix $W_w^{(pq)}$ represents the similarity relation between cross-modal samples and the matrix $W_b^{(pq)}$ indicates the separation between modalities.

The matrix $W_w^{(pq)}$ can be calculated as follows:

$$(W_w^{(pq)})_{ij} = \begin{cases} \exp\Big(- \left\| x_i^{(p)} - x_j^{(p)} \right\|^2 / 2\sigma^2 \Big), & \text{if } x_i^{(p)} \text{ and } x_j^{(q)} \text{ from the same class} \\ 0, & \text{otherwise} \end{cases} \tag{3.14}$$

where $x_j^{(p)}$ is the cross-modal pair of $x_j^{(q)}$. Indeed, a data sample can be described as $x_j^{(p)}$ in the modality p and represented as $x_j^{(q)}$ in the modality q. For example, the fingerprint image and the voice of a person can be a cross-modal pair. We cannot measure the distances directly between image samples and audio samples because of the fact that they are different features or signals. Therefore, we use the cross-modal pairs and class relationships to compute the similarity between modalities.

The matrix $W_b^{(pq)}$ can be found as follows:

$$(W_b^{(pq)})_{ij} = \begin{cases} 1, & \text{if } x_i^{(p)} \text{ and } x_j^{(q)} \text{ are from different classes,} \\ 0, & \text{otherwise} \end{cases} \tag{3.15}$$

The inter-modal Laplacian matrices $L_w^{(pq)}$ and $L_w^{(pq)}$ can be obtained from the inter-modal weight matrices $W_w^{(pq)}$ and $W_b^{(pq)}$ as follows:

$$L_w^{(pq)} = D_w^{(pq)} - W_w^{(pq)} \text{ and } L_b^{(pq)} = D_b^{(pq)} - W_b^{(pq)} \tag{3.16}$$

where $D_w^{(pq)}$ is the diagonal inter-modal within-class degree matrix with $(D_w^{(pq)})_{ii} = \sum_j (W_w^{(pq)})_{ij}$ and $D_b^{(pq)}$ is the diagonal inter-modal between-class degree matrix with $(D_b^{(pq)})_{ii} = \sum_j (W_b^{(pq)})_{ij}$.

Overall problem. All these objectives can be formulated in the following overall optimization problem

$$\begin{aligned}
\underset{\{Y^{(p)}\},\, \{\sigma^{(p)}\}}{\text{minimize}} \sum_{p=1}^{V} \Bigg\{ & \operatorname{tr}\Big({Y^{(p)}}^T L_w^{(p)} Y^{(p)} \Big) - \mu_1 \operatorname{tr}\Big({Y^{(p)}}^T L_b^{(p)} Y^{(p)} \Big) \\
& + \mu_2 \operatorname{tr}\Big({Y^{(p)}}^T (\Psi^{(p)})^{-2}\, Y^{(p)} \Big) + \mu_3 \left(\sigma^{(p)}\right)^{-2} \Bigg\} \\
+ \sum_{p=1}^{V} \sum_{q \neq p} & \Big\{ \mu_4 Tr\Big({Y^{(p)}}^T L_w^{(pq)} Y^{(p)} \Big) - \mu_5 Tr\Big({Y^{(p)}}^T L_b^{(pq)} Y^{(p)} \Big) \Big\}
\end{aligned} \tag{3.17}$$

subject to ${Y^{(p)}}^T Y^{(p)} = I$, where $\mu_1, \ldots, \mu_5$ are positive weight parameters, I is the identity matrix, and the optimization constraint ${Y^{(p)}}^T Y^{(p)} = I$ is for the normalization of the learnt coordinates.

3.3 Solution of the Optimization Problem

We first rewrite the problem in (3.17) in a more compact form. Let

$$\tilde{Y} = \Big[\; y_1^{(1)} y_2^{(1)} \cdots y_{N_1}^{(1)} \;\vdots\; y_1^{(2)} y_2^{(2)} \cdots y_{N_2}^{(2)} \;\vdots\; \cdots y_{N_p}^{(p)} \cdots \;\vdots\; y_1^{(V)} y_2^{(V)} \cdots y_{N_V}^{(V)} \; \Big]^T$$

denote the matrix containing the embeddings from all modalities. Let us also define

$$\tilde{\Psi} = \begin{bmatrix} \left(\Psi^{(1)}\right)^{-2} & 0 & \cdots & 0 \\ 0 & \left(\Psi^{(2)}\right)^{-2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \left(\Psi^{(V)}\right)^{-2} \end{bmatrix}$$

$$\tilde{L}_w = \begin{bmatrix} {L_w}^{(1)} & 0 & \cdots & 0 \\ 0 & {L_w}^{(2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & {L_w}^{(V)} \end{bmatrix}, \tilde{L}_b = \begin{bmatrix} {L_b}^{(1)} & 0 & \cdots & 0 \\ 0 & {L_b}^{(2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & {L_b}^{(V)} \end{bmatrix}$$

Next, let $\tilde{W}_{cw}$ and $\tilde{W}_{cb}$ denote the cross-modal within-class and between-class weight matrices obtained, respectively, by tiling the matrices $W_w^{(pq)}$ and $W_b^{(pq)}$ in their (p, q)-th block.

$$\tilde{W}_{cw} = \left[\begin{array}{c|c|c|c} 0 & W_w^{12} & W_w^{13} & \cdots W_w^{1V} \\ \hline W_w^{21} & 0 & W_w^{23} & \cdots W_w^{2V} \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline W_w^{V1} & W_w^{V2} & \cdots & 0 \end{array}\right], \tilde{W}_{cb} = \left[\begin{array}{c|c|c|c} 0 & W_b^{12} & W_b^{13} & \cdots W_b^{1V} \\ \hline W_b^{21} & 0 & W_b^{23} & \cdots W_b^{2V} \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline W_b^{V1} & W_b^{V2} & \cdots & 0 \end{array}\right]$$

We can then define the corresponding Laplacian matrices $\tilde{L}_{cw} = \tilde{D}_{cw} - \tilde{W}_{cw}$ and $\tilde{L}_{cb} = \tilde{D}_{cb} - \tilde{W}_{cb}$, where $\tilde{D}_{cw}$ and $\tilde{D}_{cb}$ are the corresponding diagonal degree matrices obtained by summing up the entries of $\tilde{W}_{cw}$ and $\tilde{W}_{cb}$ in each row. Letting

$$A = \tilde{L}_w - \mu_1 \tilde{L}_b + \mu_2 \tilde{\Psi} + \mu_4 \tilde{L}_{cw} - \mu_5 \tilde{L}_{cb}$$

the problem in (3.17) can be rewritten as

$$\underset{\tilde{Y},\, \{\sigma^{(p)}\}}{\text{minimize}}\, \text{tr}(\tilde{Y}^T A \tilde{Y}) + \mu_3 \sum_{p=1}^{V} (\sigma^{(p)})^{-2}, \text{ subject to } \tilde{Y}^T \tilde{Y} = I. \tag{3.18}$$

The above problem is not jointly convex in $\tilde{Y}$ and $\{\sigma^{(p)}\}$, hence it is not easy to find its global optimum. We minimize the objective function with an iterative alternating optimization scheme, where we first optimize $\tilde{Y}$ by fixing $\{\sigma^{(p)}\}$, and then optimize $\{\sigma^{(p)}\}$ by fixing $\tilde{Y}$ in each iteration as follows:

Optimization of $\tilde{Y}$: When $\{\sigma^{(p)}\}$ are fixed, the optimization problem in (3.18) becomes

$$\underset{\tilde{Y}}{\text{minimize}}\ \text{tr}(\tilde{Y}^T A\tilde{Y}) \text{ subject to } \tilde{Y}^T\tilde{Y} = I. \tag{3.19}$$

The solution of this problem is given by the m eigenvectors of the matrix A corresponding to its smallest m eigenvalues.

Optimization of $\{\sigma^{(p)}\}$: Fixing $\tilde{Y}$, the problem (3.18) becomes

$$\underset{\{\sigma^{(p)}\}}{\text{minimize}}\ \mu_2\text{tr}\big(\tilde{Y}^T\tilde{\Psi}\tilde{Y}\big) + \mu_3\sum_{p=1}^{V}\Big(\frac{1}{\sigma^{(p)}}\Big)^2. \tag{3.20}$$

Note that the first term in the objective depends on the kernel scale parameters $\{\sigma^{(p)}\}$ through the entries of the kernel matrix $\tilde{\Psi}$. Due to the block diagonal structure of $\{\tilde{\Psi}\}$ and the separability of the second term, the objective (3.20) can be decomposed into V objectives, each one of which is a function of $\sigma^{(p)}$, for $p = 1, \ldots, V$. We minimize these objective functions one by one, by optimizing one scale parameter $\sigma^{(p)}$ at a time through exhaustive search.

The outline of the proposed algorithm can thus be summarized as follows:

Algorithm **M**ulti-modal **N**onlinear **S**upervised **E**mbedding (MNSE)

1: **Input:**
   Training data matrix $X^{(p)}$
   Training data labels
2: **Initialization:**
   Obtain the graph Laplacian matrices $\tilde{L}_w, \tilde{L}_b, \tilde{L}_{cw}$, and $\tilde{L}_{cb}$,
   Assign weight parameters $\{\mu_1, \mu_2, \cdots, \mu_5\}$, and initial kernel scales $\sigma^{(p)}$
3: **repeat**
4: Compute the nonlinear embeddings $Y^{(p)}$ through (3.19) by fixing $\sigma^{(p)}$
5: Compute the kernel scale parameters $\sigma^{(p)}$ through (3.20) by fixing $Y^{(p)}$
6: **until** the maximum number of iterations or the convergence of the objective
7: **Output:**
   Kernel coefficients matrix $C^{(p)} = \big(\Psi^{(p)}\big)^{-1}Y^{(p)}$
   Kernel scale parameters $\sigma^{(p)}$ and projected training data $Y^{(p)}$

Nonlinear embeddings are in the form $y = f(x)$. The goal of the proposed algorithm

is to obtain the optimum embedding function $f(.)$. The proposed algorithm directly optimizes the embeddings y through the optimum embedding function. Therefore, the proposed method is nonlinear. On the other hand, linear embeddings are in the form $y = P^T x$. The main aim of a linear embedding is the optimization of the projection matrix P instead of the embeddings y.

3.4 Convergence of the Algorithm

Definition 2. *A matrix M is positive semi-definite, if it is symmetric ($n \times n$) and satisfies the following property:*

$$v^T M v \geq 0, \forall v \in \mathbb{R}^n$$

If the weight parameters μ_1 and μ_5 are chosen sufficiently small, the matrix A becomes positive semi-definite by Theorem 2. In this case the objective function is guaranteed to converge since it is nonnegative, and both updates on $\tilde{Y}$ and $\{\sigma^{(p)}\}$ reduce it.

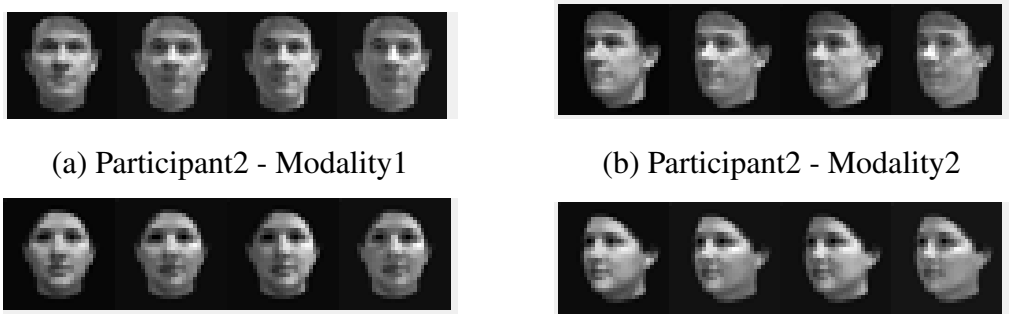
CHAPTER 4

EXPERIMENTAL RESULTS

4.1 Data Sets Used In the Experiments

We tried our proposed solution on several frequently used multi-modal data sets and compared classification and retrieval results with some state-of-the-art methods in the multi-modal learning literature. These data sets can be explained as follows:

- **MITCBCL face images data set:** This data set includes 3240 face images, which were captured and published by MIT University CBCL Community [69]. The dataset contains face images of 10 participants captured under 36 illumination conditions and 9 different pose angles. We have conducted the classification experiments on images with frontal and profile poses, which are considered to represent two different modalities. Some sample images of two different participants in both modalities are shown in Figure 4.1.



(a) Participant2 - Modality1 (b) Participant2 - Modality2

(c) Participant5 - Modality1 (d) Participant5 - Modality2

Figure 4.1: Sample face images from the MITCBCL data set

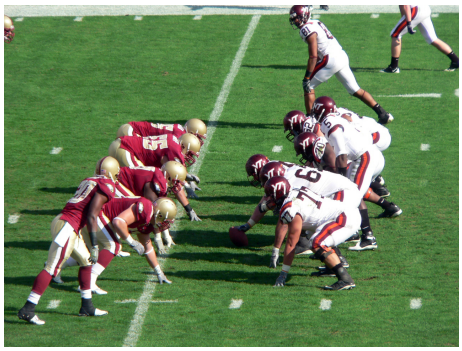
- **Wikipedia image-text pair data set:** The retrieval experiments are done on

the Wikipedia image-text data set [70]. The data set contains 2866 image-text pairs describing the contents of the articles, which are categorized into 10 classes. 128-dimensional SIFT histograms are used in the image modality, and 10-dimensional text features obtained with a latent Dirichlet allocation model are used for the text modality [71], [72]. Several samples in the Wikipedia data set are given in Figures 4.2 and 4.3.

(a) Category "war"

(b) Category "geography"

(c) Category "art"



(d) Category "sport"

Figure 4.2: Sample Pascal VOC images

Boston College won the ceremonial pre-game coin toss to determine first possession and elected to kick off to begin the game, ensuring that the Eagles would receive the ball to begin the second half. Virginia Tech kick returner Dyrell Roberts fielded the ball at the Tech six-yard line and returned it to the Tech 33-yard line before the first play of the game, a four-yard pass from quarterback Tyrod Taylor to wide receiver Danny Coale. Despite that initial gain, the Hokies were unable to advance the ball further and punted after going three and out. Boston College's Rich Gunnell fair caught the kick at the Eagles' 16-yard line, where the Boston College offense ran its first play of the game. Running back Montel Harris ran for a one-yard gain, but Boston College was unable to gain the needed ten yards for a first down, just as Virginia Tech had failed to do in the prior series. After going three and out, the Eagles punted back to Virginia Tech and the Hokies' offense returned to the field at the Tech 39-yard line with 11:30 remaining in the first quarter.

Figure 4.3: Sample tag for image (d) in Figure 4.2

- **Pascal VOC 2007 data set:** It is a challenging image-text data set that is constructed from the various feature types such as the Gist vectors of the images and the number of words used in the texts [73]. There exists 5011 training samples and 4952 test samples in overall set. Nevertheless, we used the images that contain only one object in the experiments. Thus, all data set was reduced to 2808 training samples and 2841 test samples for each modality.

  The data set includes 20 different object classes. A few Pascal VOC 2007 data samples are shown in Figure 4.4.

(a) **Tags:** person(x5), horse(x5), tree

(b) **Tags:** person, car

(c) **Tags:** person, baby carriage

(d) **Tags:** table, chair(x6), rug, laptop, door

Figure 4.4: Sample Pascal VOC 2007 images and corresponding tags

4.2 Image Classification Experiments

The image classification experiments are conducted on the MIT CBCL face images data set. 360 frontal images are used as Modality 1 samples and 360 profile images are used as Modality 2 [1] samples. In the training phase, firstly, K-nearest neighbourhood graphs are constructed for each modality and the graph Laplacians are found. Then,

[1] "Modality" and "view" have equivalent meanings in the thesis

the embeddings of the data samples in the modalities are computed with the proposed MNSE algorithm. Obtained kernel functions are used to project the training and the test samples to another domain, which has lower dimension than the original domain. The NN classifiers are learned through the projected training samples and the class labels. The trained classifiers and the projected test samples are used to estimate the class labels of the test samples. Finally, the estimated class labels are compared to the true class labels of the test samples to obtain the misclassification error rates.

4.2.1 The effect of the algorithm parameters on the classification performance

The following parameters may have an impact on the algorithm performance:

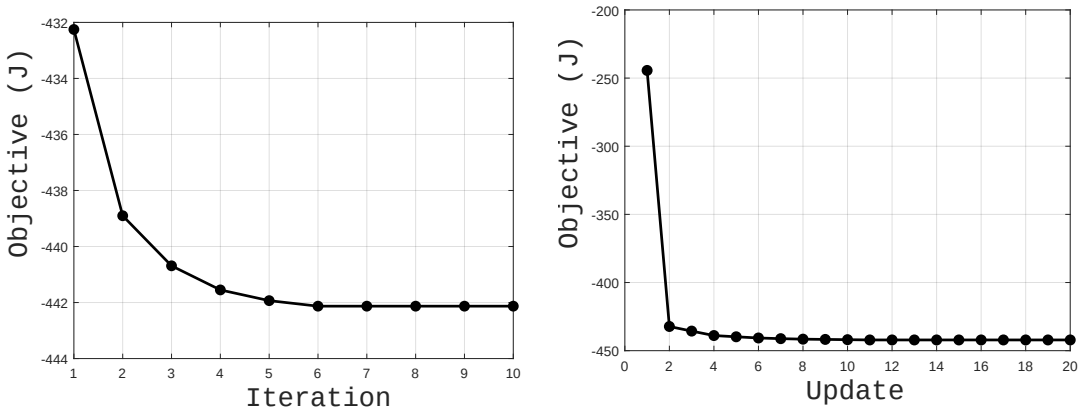
- Number of iterations
- Weight parameters ($\mu_1, \mu_2, ..., \mu_5$)
- Embedding dimension

In order to analyse the algorithm performance, 100 training and 260 test samples were used for each modality. Firstly, we studied the evolution of the misclassification error throughout the iterations of the optimization algorithm. The variation of the optimization objective function throughout the iterations can be seen in Figure 4.5. The term "update" indicates update of the embeddings (Y) and update of the kernel scale parameter (σ) so that there exists two updates in each algorithm iteration.

Figure 4.5 indicates that the overall objective function is a nonincreasing function for each algorithm iteration. This confirms that our iterative algorithm works efficiently to solve the established optimization problem. The updates on both the embeddings (Y) and the kernel scale parameter (σ) ensure that the overall cost decrease or remain constant.

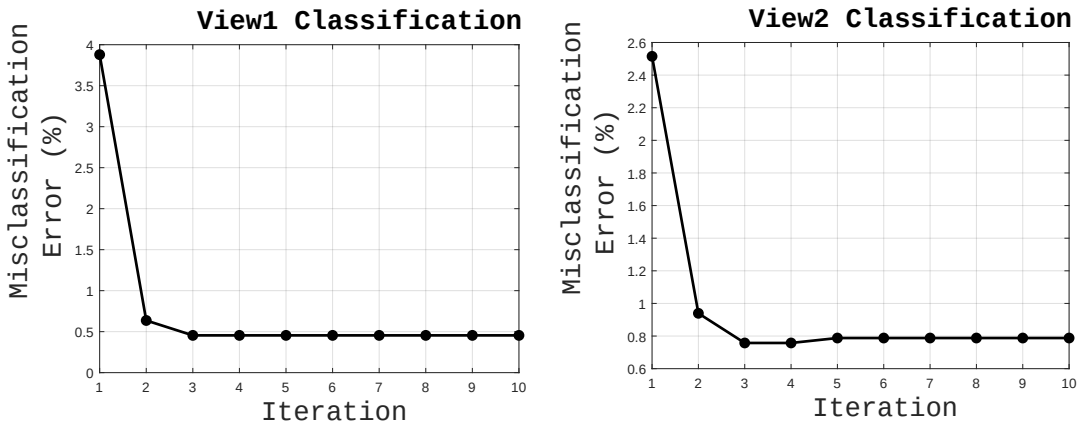
Misclassification errors of the modalities are illustrated in Figure 4.6 for each algorithm iteration.

Figure 4.6 shows that the MNSE algorithm rapidly converges to its optimum point for each modality. This result is consistent with the results in Figure 4.5, because of



(a) Objective vs algorithm iteration (b) Objective vs each update of the algorithm

Figure 4.5: Objective vs algorithm iteration for the MIT CBCL face data set



(a) Modality 1 (b) Modality 2

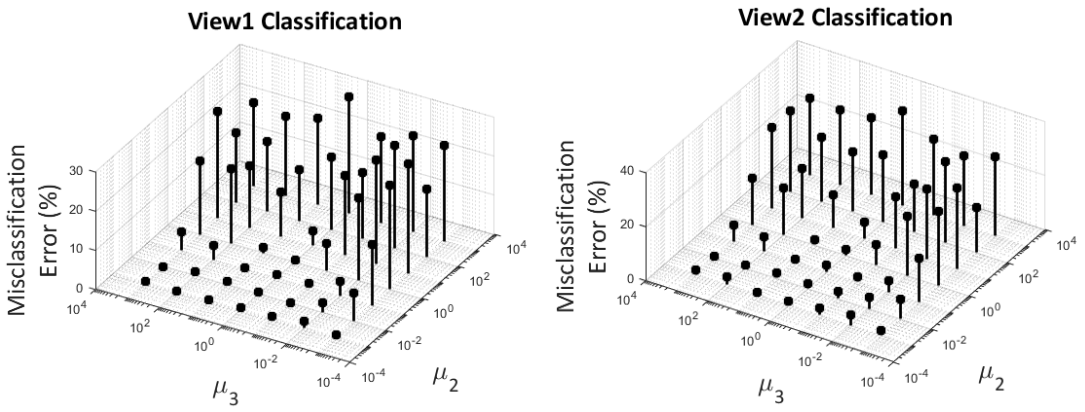
Figure 4.6: Iteration vs misclassification error for the MIT CBCL face data set

the fact that the decrease in the overall cost function also leads to a decrease in the misclassification error. Thus, it can be inferred that the objective function is indeed well representative of the classification performance.

Next, the effects of the weight parameters are studied in Figures 4.7 and 4.8. Since it is quite difficult to display the effects of the weight parameters on one plot at the same time, pairwise effects are analysed independently in 3-D graphs. Moreover, values of the weight parameters are assigned from the set $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10^{1}, 10^{2}, 10^{3}\}$.

Figure 4.7 shows the misclassification errors for different values of the parameters μ_2 and μ_3 when other algorithm parameters are constant. Values of the other weight

parameters μ_1, μ_4, and μ_5 are taken as 10^2, 1, and 10^2 respectively for this experiment.



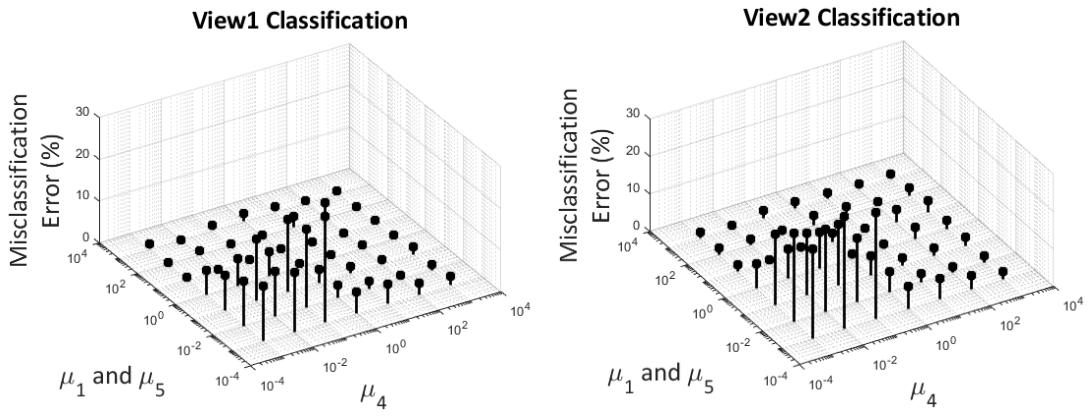
(a) Modality 1 (b) Modality 2

Figure 4.7: Weight parameters (μ_2, μ_3) vs misclassification error for the MIT CBCL face data set

Figure 4.7 indicates that the weight parameter for the kernel function norm constraint (μ_2) should be low, but the weight parameter for the kernel scale parameter constraint (μ_3) should be close to 1. This can be explained with the significant differences in the orders of magnitudes of these terms. An appropriate assignment of these weight parameters basically aims to balance the orders of the magnitudes.

Figure 4.8 demonstrates the misclassification errors for different values of the parameters μ_4 and μ_1-μ_5 when other algorithm parameters are constant. Weight parameters for intra and inter modalities between class separations (μ_1-μ_5) are chosen as exactly the same, because between class separation matrices are constructed in a same way. Remaining weight parameters μ_2, and μ_3 are taken as 10^{-2} and 1 respectively for this experiment.

Figure 4.8 indicates that the weight parameter (μ_4) for the inter-modal within-class similarity of modalities should be in the interval $[1, 10^1]$ and the weight parameters (μ_1, μ_5) for the between-class discrimination of modalities should be high. This result suggests that between-class discrimination terms should be dominant over within-class similarity terms for this data set. This may be caused by the fact that within-class scattering is high for the data set. Therefore, our algorithm tends to perform better



(a) Modality 1 (b) Modality 2

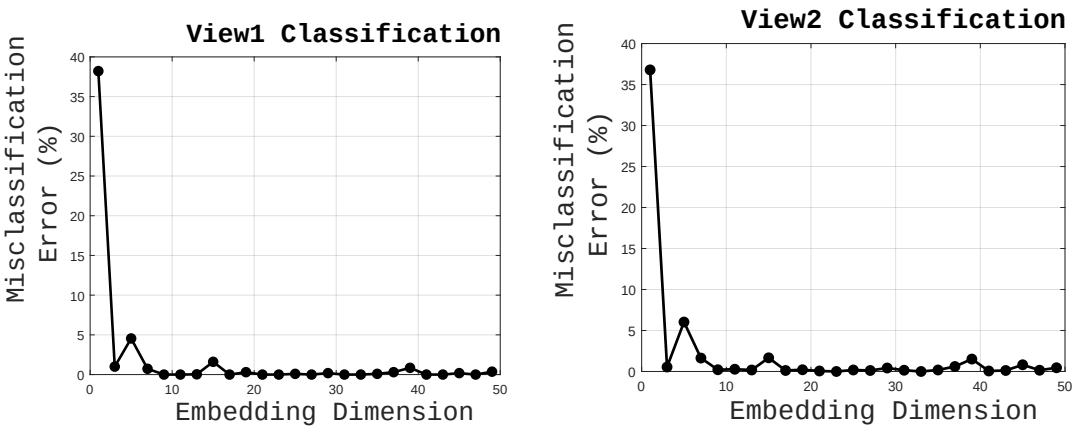
Figure 4.8: Weight parameters (μ_1, μ_4, μ_5) vs misclassification error for the MIT CBCL face data set

when the distances between class means are increased. Another important outcome of this experiment is that weights of the intra-modal and the inter-modal within-class similarities can be taken as equal in the optimization.

Reducing the dimensions of the feature vectors by compromising on the minimum loss of information is one of the important goals of the proposed algorithm. Thereby, choosing a suitable dimension value as an algorithm input should be taken into account carefully. Figure 4.9 shows the relationship between the embedding dimension and the misclassification error.

From Figure 4.9, it can be understood that the MNSE algorithm works well at lower dimensions for the MIT CBCL face data set. The smallest dimension value that yields a reasonable misclassification error in Figure 4.9 can be observed as 9. Achieving high classification accuracy at low dimensions is also helpful as it decreases the computational load. For these reasons, the embedding dimension is chosen as 9 during the experiments.

For the MIT CBCL face images data set, there exist 10 classes and the optimum embedding dimension is obtained as 9. This suggests that the optimum embedding dimension may be expected to be close to the number of classes for the data set.

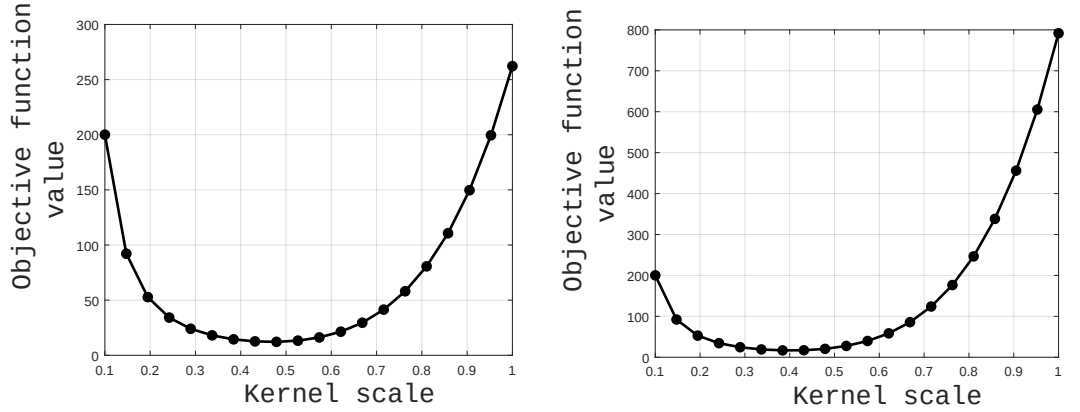


(a) Modality 1 (b) Modality 2

Figure 4.9: Embedding dimension vs misclassification error for the MIT CBCL face data set

4.2.2 The MNSE algorithm behaviour on the MIT CBCL face images data set

During the algorithm iterations, alternating optimization of the kernel scale parameter is made in order to determine the proper kernel scale parameters for each modality. Figure 4.10 demonstrates the relationship between the kernel scale parameter and the objective function when embeddings are fixed.



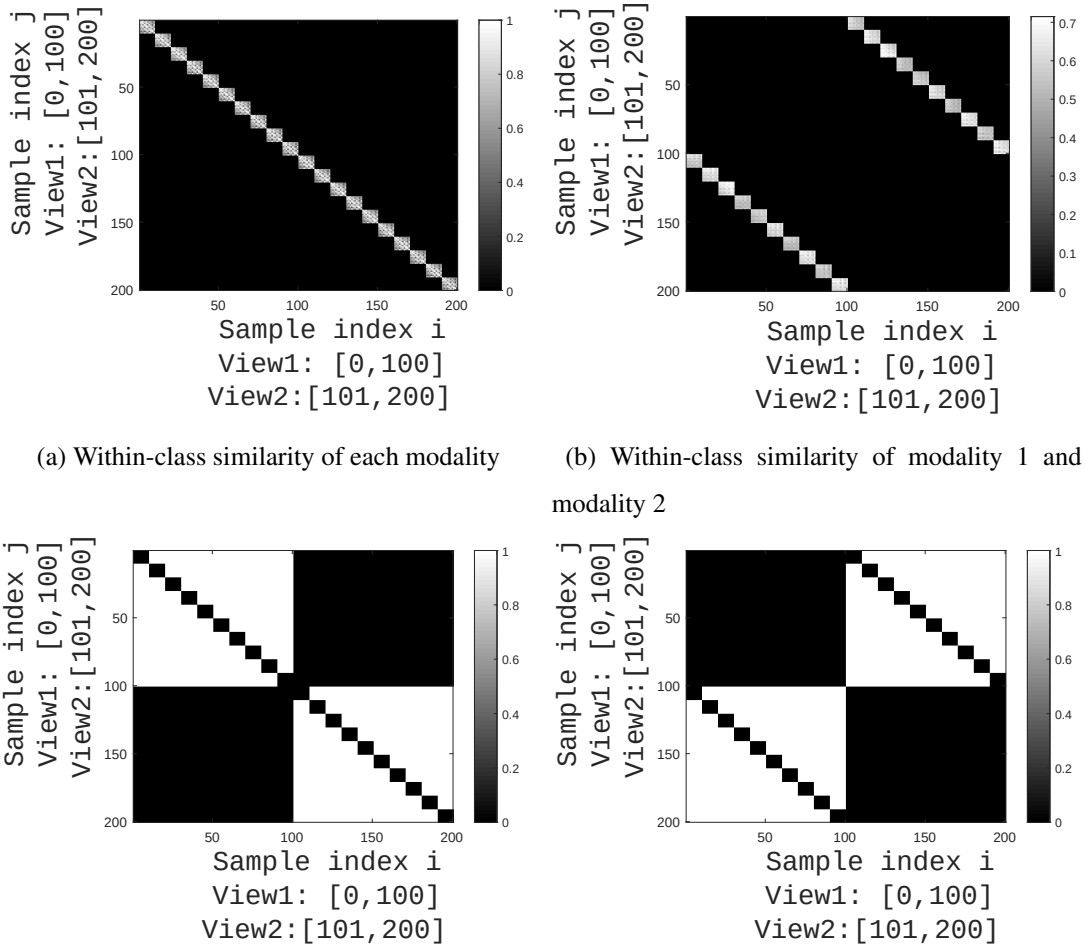
(a) Modality 1 (b) Modality 2

Figure 4.10: Kernel scale parameters and corresponding objective functions for the MIT CBCL face data set

It can be understood from Figure 4.10 that a suitable value for the kernel scale pa-

rameter σ can be found through a basic search in an interval. This is the result from the fact that any increase in σ decreases the cost of the kernel scale parameter in the overall objective, but it increases the cost of the kernel function norm and vice versa.

The appearances of the within-class similarity and the between-class dissimilarity matrices for the reduced MIT CBCL face training data set are shown in Figure 4.11.



(a) Within-class similarity of each modality

(b) Within-class similarity of modality 1 and modality 2

(c) Between class separation of each modality

(d) Between class separation of modality 1 and modality 2

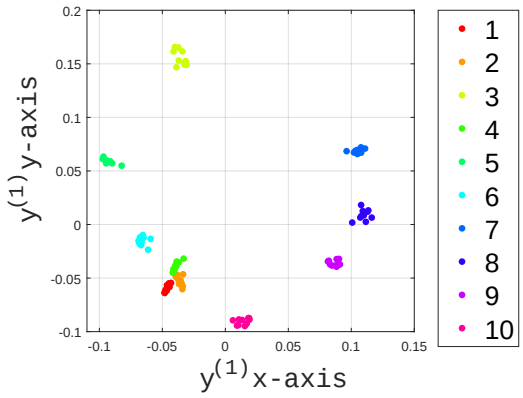
Figure 4.11: Intra and inter modality class relationships of the MIT CBCL face images

Figure 4.11 presents the block-diagonal affinity matrices constructed from class-ordered training samples. Intra modality similarity relationships for each modality are computed from a Gaussian distance metric so that if a suitable scale parameter is chosen,

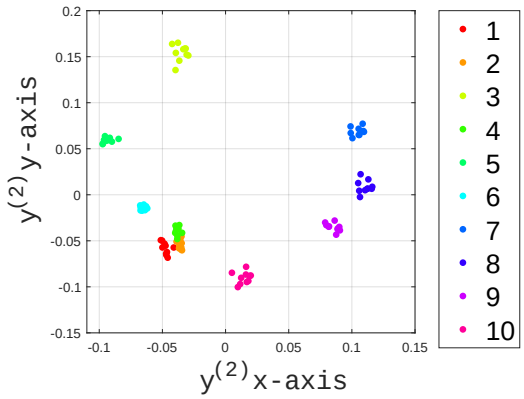
the white and gray areas are distributed evenly in the similarity matrix. This situation can be observed in Figure 4.11a. In order to represent the inter modality similarity relationships, one-to-one correspondences between modality 1 and modality 2 are used. Figure 4.11b illustrates the top right block diagonals showing the within-class similarities between modalities 1 and 2, which are constructed from the within-class similarity relations in modality 2. Similarly, the bottom left block diagonal matrices show the within-class similarities between modalities 1 and 2, constructed using the within-class similarity relations in modality 1.

Similar observations can be made for the between-class separation relationships of modalities. Instead of using a Gaussian distance metric, a constant value, i.e. 1, is directly assigned for the samples which are not in the same classes. Thus, the weight matrix in Figure 4.11c is obtained, which contains only black and white areas. Like in the within-class similarity relationships, block diagonal matrices are observed when training samples are ordered with respect to classes. The same method as in the inter modality within-class similarity is applied to build the inter modality between-class relationships. The resulting inter modality between-class dissimilarity matrices can be seen in Figure 4.11d. The embeddings of the training and test samples are demonstrated in Figure 4.12.

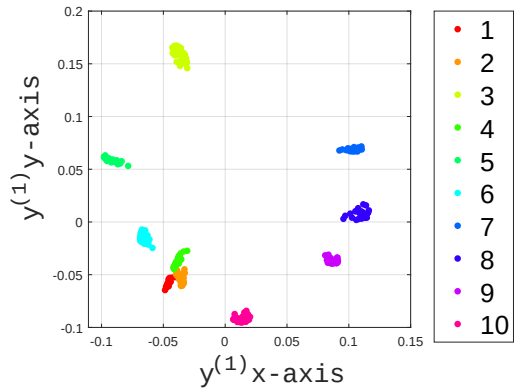
Furthermore, Figure 4.13 and Figure 4.14 illustrate the embedding results of the other algorithms.



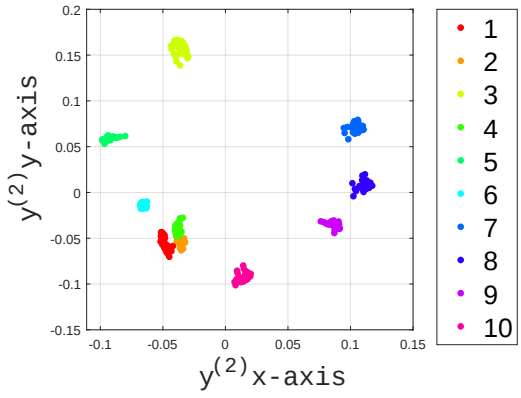
(a) Embeddings of the modality 1 training samples coloured with respect to the class labels



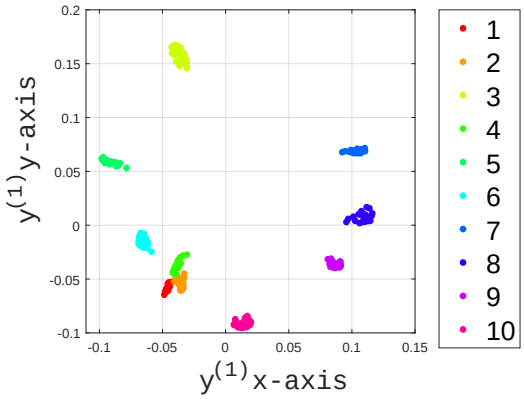
(b) Embeddings of the modality 2 training samples coloured with respect to the class labels



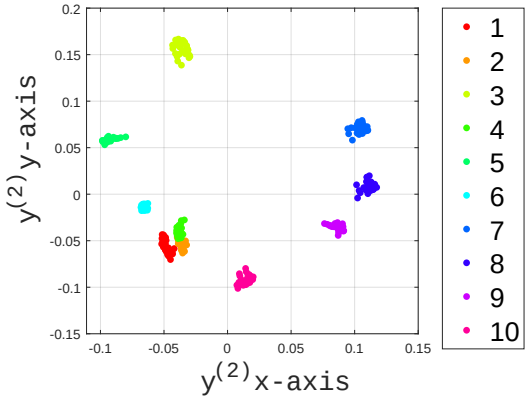
(c) Embeddings of the modality 1 test samples coloured with respect to the true class labels



(d) Embeddings of the modality 2 test samples coloured with respect to the true class labels

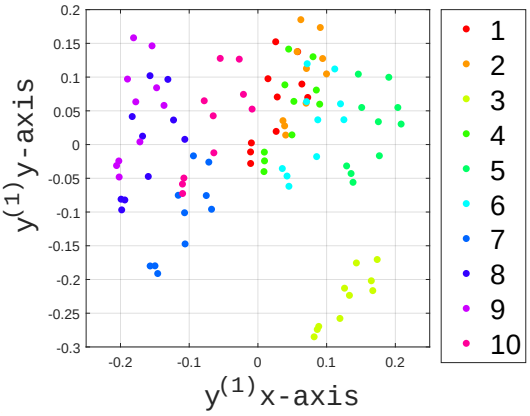


(e) Embeddings of the modality 1 test samples coloured with respect to the estimated class labels

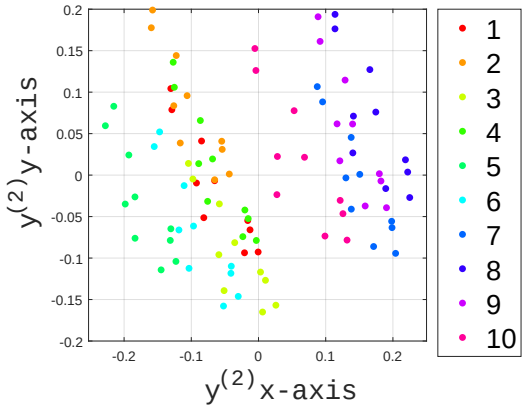


(f) Embeddings of the modality 2 test samples coloured with respect to the estimated class labels

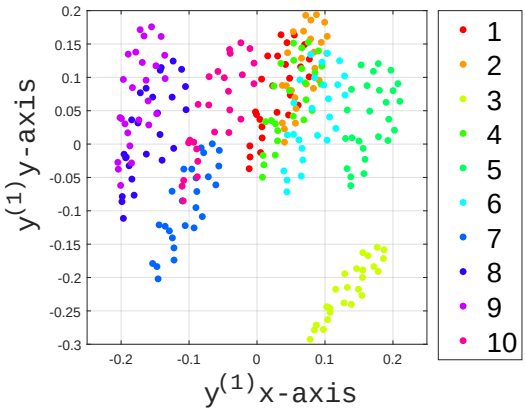
Figure 4.12: Embeddings of the MIT CBCL face images with the proposed method. Each color indicates a different class label in 1-10.



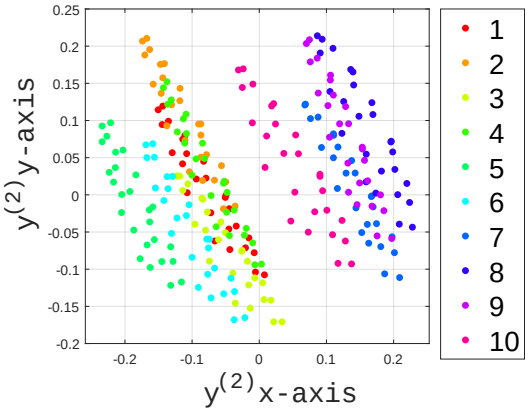
(a) PCA embeddings of the modality 1 training samples coloured with respect to the class labels



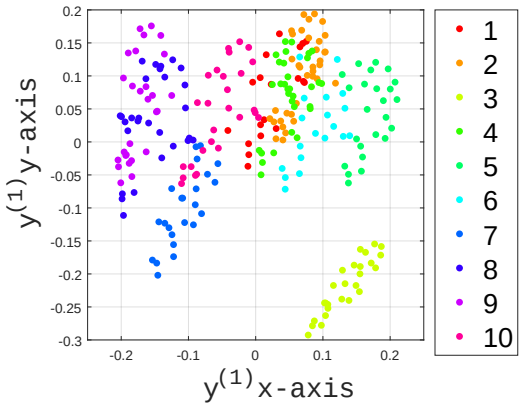
(b) PCA embeddings of the modality 2 training samples coloured with respect to the class labels



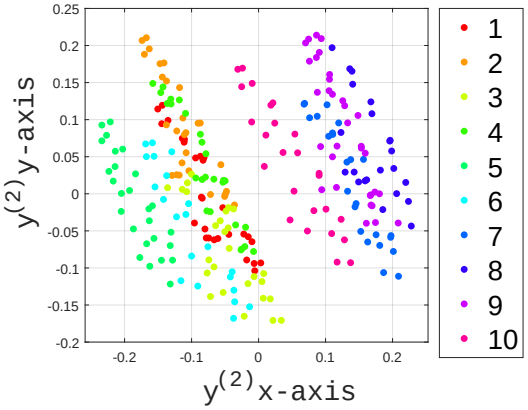
(c) PCA embeddings of the modality 1 test samples coloured with respect to the true class labels



(d) PCA embeddings of the modality 2 test samples coloured with respect to the true class labels

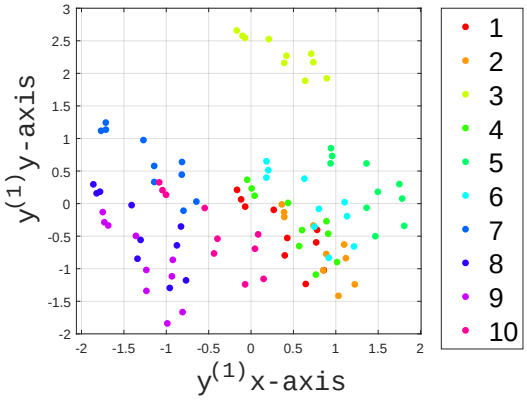


(e) PCA embeddings of the modality 1 test samples coloured with respect to the estimated class labels

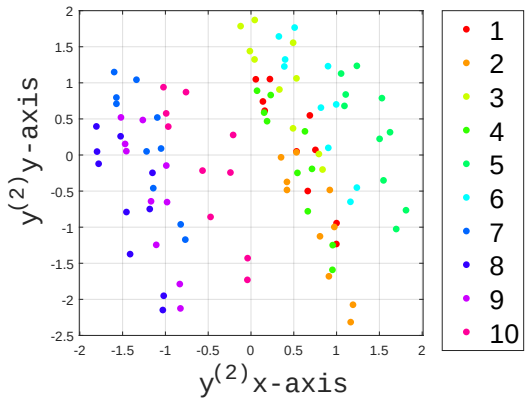


(f) PCA embeddings of the modality 2 test samples coloured with respect to the estimated class labels

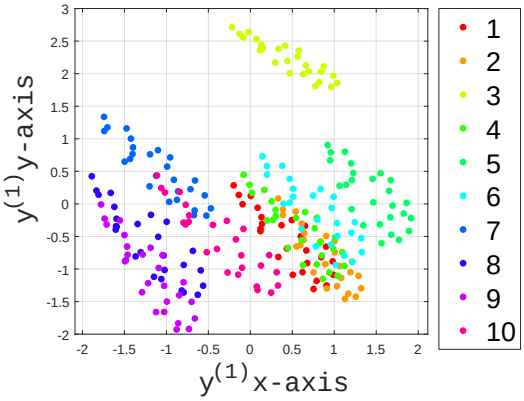
Figure 4.13: PCA embeddings of the MIT CBCL face images. Each color indicates a different class label in 1-10.



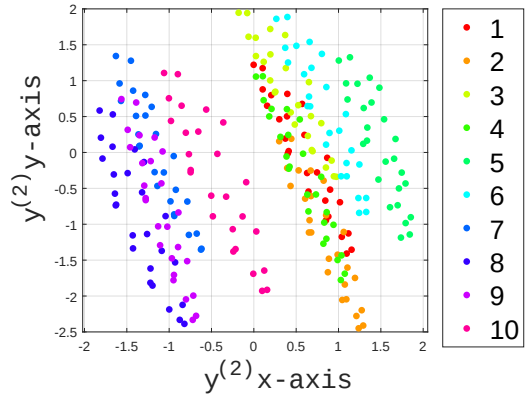
(a) PCA+CCA embeddings of the modality 1 training samples coloured with respect to the class labels



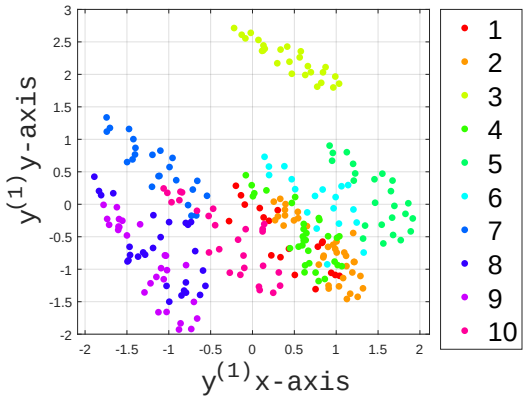
(b) PCA+CCA embeddings of the modality 2 training samples coloured with respect to the class labels



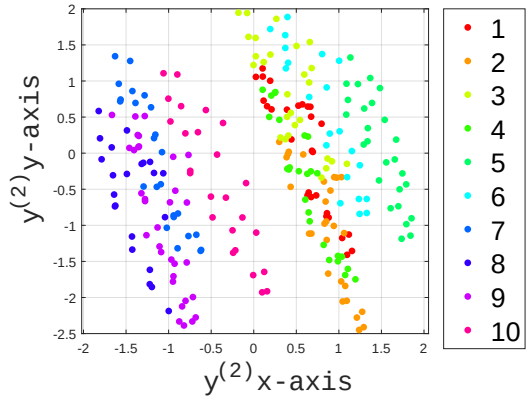
(c) PCA+CCA embeddings of the modality 1 test samples coloured with respect to the true class labels



(d) PCA+CCA embeddings of the modality 2 test samples coloured with respect to the true class labels



(e) PCA+CCA embeddings of the modality 1 test samples coloured with respect to the estimated class labels



(f) PCA+CCA embeddings of the modality 2 test samples coloured with respect to the estimated class labels

Figure 4.14: PCA+CCA embeddings of the MIT CBCL face images. Each color indicates a different class label in 1-10.

One can clearly observe that the spontaneous grouping of the embeddings of the test samples largely agrees with the class labels of the samples, and the estimated class labels are very close to true class labels in Figure 4.12. This can be an important sign for the success of the MNSE algorithm.

Figure 4.13 indicates that the efficiency of the dimensionality reduction with the PCA algorithm is not as successful as the MNSE algorithm. Similar observations can be made for the PCA+CCA algorithm by Figure 4.14. It can be concluded that the nonlinear embedding with MNSE leads to better between-class separation than PCA and the linear embedding with CCA.

4.2.3 Comparison of the proposed method with other algorithms

In this section, the proposed MNSE algorithm is compared to the multi-modal representation learning algorithms CCA, GMLDA [44], JFSSL [3], as well as the baseline single-modal methods PCA, NSSE [14], and NN classification in the original domain. Moreover, the proposed MNSE algorithm is also compared to the specialized version of the MNSE algorithm, which does not have the Lipschitz regularity term. The data set is separated randomly into training and test sets at different ratios. The multi-modal CCA and GMLDA algorithms are applied after a dimensionality reduction step with PCA, which provides more accurate results. For the multi-modal methods, projections from different modalities into a common space are learnt with the training data. In the test stage, a scenario is considered where a test image is available in only one modality. Test images are projected with the learnt embeddings and are classified with NN classification using the projections of the training samples of their own modality. The single-modal methods are applied independently in each modality. Table 4.1 shows the misclassification rates (in percentage) of test images for different training sizes, using their representations in Modalities 1 and 2. The results are averaged over 10 random repetitions of the experiment.

The results in Table 4.1 show that the proposed MNSE method outperforms all single-modal methods and CCA in all setups. The comparison between MNSE and the single-modal NSSE method is particularly interesting. Both methods compute non-linear smooth projection functions and perform the final NN classification with the

Table 4.1: Misclassification rates (%) of compared methods. Top and bottom rows show the errors obtained with Modalities 1 and 2.

Algorithm	**Training size**				
	5.6%	**8.3%**	**11.1%**	**13.9%**	**27.8%**
NN	22.12	19	10.69	2.97	0.77
	19.68	17.64	6.94	1.71	0
PCA	3.68	0.06	0.34	0.10	0
	4.29	0.54	0.06	0	0
NSSE	1.94	0.03	0.03	0	0
	4.56	1	0.09	0.03	0
CCA	3.67	0.06	0.34	0.10	0
	4.29	0.55	0.06	0	0
GMLDA	0	0	0	0	0
	0.30	0.06	0.03	0	0
JFSSL	0	0	0	0	0
	0.12	0	0	0	0
MNSE	0.15	0	0	0	0
	1.35	0.27	0.03	0	0
MNSE without Lipschitz	3.74	0.06	0.03	0	0
	1.94	0.39	0.03	0	0

embeddings of training samples from only one modality. Hence, the fact that MNSE outperforms NSSE confirms that it successfully exploits the information from both modalities during the computation of the projection function, in contrast to the single-modal NSSE. One can also observe that the linear JFSSL and GMLDA methods yield similar classification performance to MNSE and can outperform MNSE in some cases. Computing linear projections of the two modalities into a common space, JFSSL and GMLDA perform particularly well in this synthetic and regularly structured face data set, as the images viewing the same participants from different angles are quite convenient to align via linear transformations.

Another point to consider from Table 4.1 is that representation based approaches are more successful than classification in the original domain. Furthermore, the unsuper-

vised learning method CCA is not as accurate as supervised methods, as it does not employ the information of the class labels when learning projections.

It can also be observed from Table 4.1 that the MNSE algorithm without the Lipschitz regularity condition, i.e. fixed kernel scale parameters, has higher misclassification error rate than the MNSE algorithm with the Lipschitz term. It shows the positive effects of including the Lipschitz regularity condition in the learning objective, which provides the generalization of interpolator to the whole sample space.

Computation time analysis is made for the compared methods as follows: 8 experiments with 10 repetitions are conducted through the different number of training and test samples. Training set sizes are chosen as 20, 30, 40, 50, 100, 150, 200 and 250. In the data set, there exist 360 samples for each modality. In order to obtain average run times for each algorithm, the duration between the loading of all data and the production of the classification results is divided to the number of repetitions, which is 10. A computer, which has 16 GB RAM, 512 GB SSD and Intel Xeon Processor model E3-1240 v6 with 4 cores, 8M Cache, 3.70 GHz base frequency, is used to conduct the experiments. Table 4.2 demonstrates the average run times of each experiment.

Table 4.2: Computation times of the compared methods

Algorithm	**Average run time (seconds)**
Original domain	1.7746
PCA	4.3159
NSSE	8.0059
CCA	4.2083
GMLDA	3.5710
JSSL	7.4058
MNSE	21.3442

According to Table 4.2, the computation time of the MNSE algorithm is the highest one. It probably results from the solution of the nonlinear optimization problem and the incorporation of the inter-modal relations of the data. The computation time of the NSSE algorithm is shorter than the MNSE algorithm, because of the fact that the

MNSE algorithm uses the information of the additional modality. The PCA, CCA, GMLDA and JFSSL algorithms have a smaller run time than the MNSE algorithm, since it is easier to solve the optimization problems in these methods.

4.3 Image-Text Retrieval Experiments

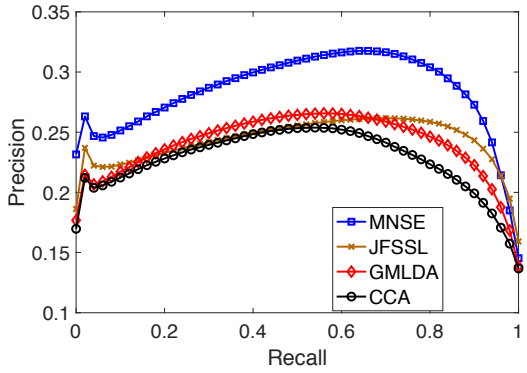
The retrieval experiments are done on the Wikipedia image-text data set [70] and the Pascal VOC 2007 image-text data set [73]. Firstly, proper projection functions are learnt using the training set with the compared methods. Using these projections, all samples are moved to the new space of embedding. Then, the retrieval task is performed on the test set, by searching the relevant matches of an image query in the text database (based on the nearest neighbours in the common space), and vice versa. The precision and recall rates are computed by considering a retrieved item relevant if it is from the same category as the query. (4.1) demonstrates how precision and recall metrics are calculated in the experiments.

$$\begin{aligned} \textbf{Precision} &= \frac{\#\text{ of relevant samples retrieved}}{\#\text{ of all retrieved samples}} \\ \textbf{Recall} &= \frac{\#\text{ of relevant samples retrieved}}{\#\text{ of all relevant samples}} \end{aligned} \tag{4.1}$$

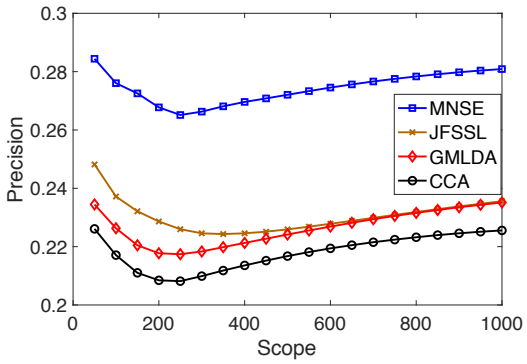
After that the Average Precision (AP) score of each query sample is calculated by averaging precision values over all relevant retrieved samples [3]. Finally, the Mean Average Precision (MAP) scores of the methods, computed by averaging all average precision values over all query samples [3]. Hence, if the MAP score of an algorithm is higher than that of the others, it can be concluded that this algorithm works more precisely in a retrieval experiment.

4.3.1 Retrieval Experiments on the Wikipedia Data Set

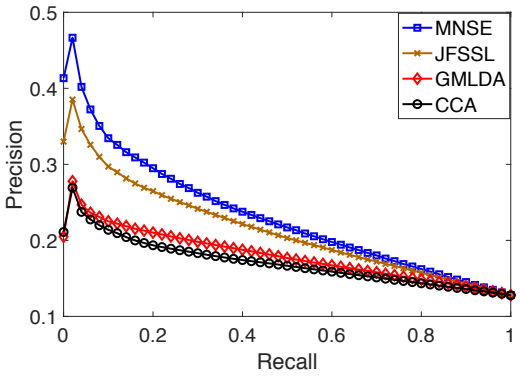
For the Wikipedia data set, obtained precision-recall and precision-scope curves are illustrated in Figure 4.15 and calculated MAP scores are shown in Table 4.3:



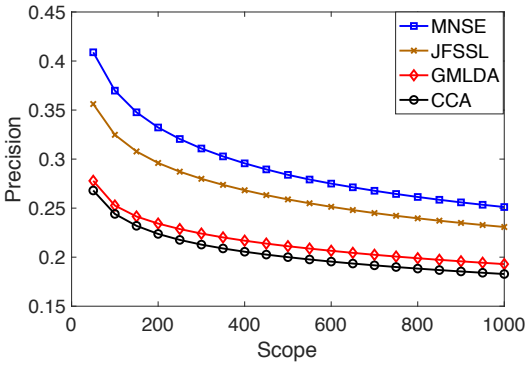
(a) Image query



(b) Image query



(c) Text query



(d) Text query

Figure 4.15: Retrieval results for the Wikipedia data set

Table 4.3: MAP scores for the Wikipedia data set

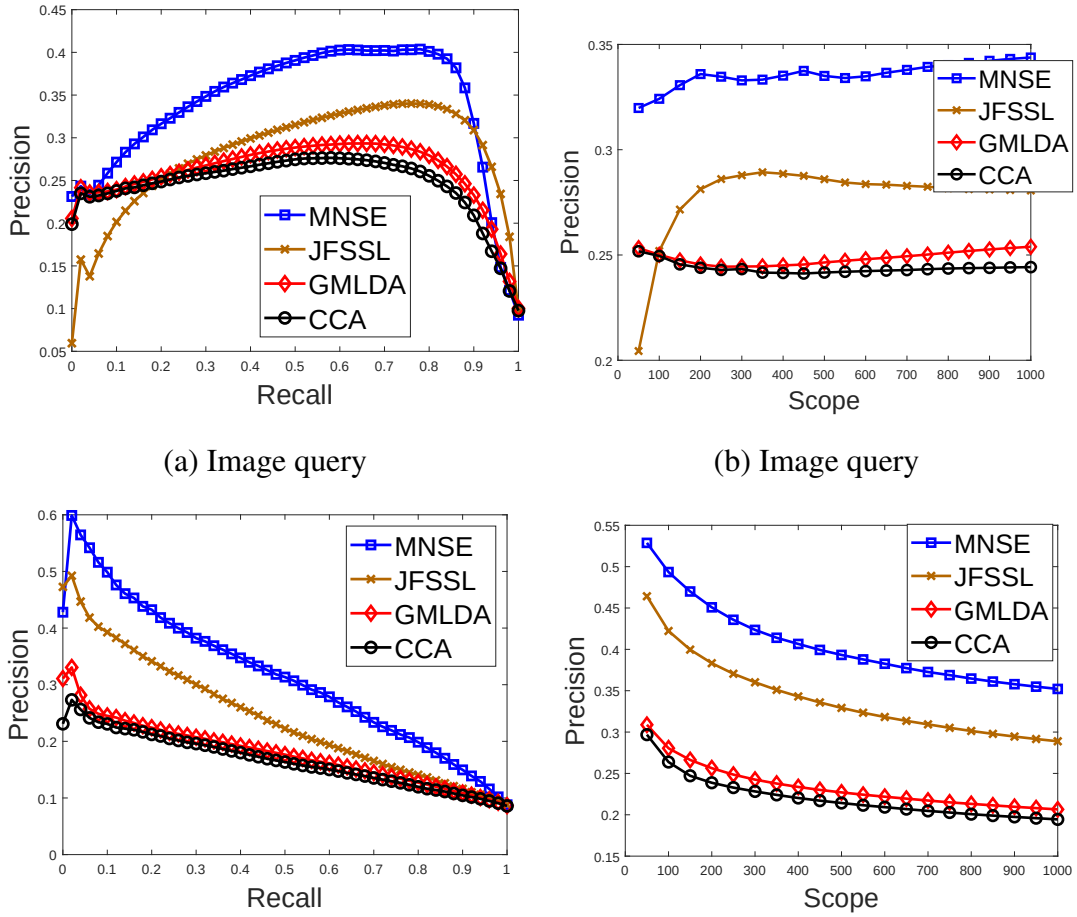
Algorithm	**CCA**	**GMLDA**	**JFSSL**	**MNSE**
Image Query	0.2280	0.2407	0.2440	0.2847
Text Query	0.1720	0.1815	0.2143	0.2321

4.3.2 Retrieval Experiments on the Pascal VOC 2007 Data Set

The precision-recall, precision-scope curves and the MAP scores for the Pascal VOC 2007 data set are presented in Figure 4.16 and Table 4.4:

4.3.3 Results of the Retrieval Experiments

The results in Figures 4.15, 4.16 and Tables 4.3, 4.4 show that the proposed MNSE



(a) Image query

(b) Image query

(c) Text query

(d) Text query

Figure 4.16: Retrieval results for the Pascal VOC 2007 data set

Table 4.4: MAP scores for the Pascal VOC 2007 data set

Algorithm	**CCA**	**GMLDA**	**JFSSL**	**MNSE**
Image Query	0.2470	0.2609	0.2814	0.3390
Text Query	0.1674	0.1791	0.2418	0.3159

method outperforms all other multi-modal methods in all retrieval experiments. In contrast to the face data set used in the previous experiment, the Wiki and the Pascal VOC 2007 data sets have more diverse and irregular contents, and the two modalities bear less resemblance. This makes the multi-modal representation learning task more challenging, where the flexibility of the proposed nonlinear embedding approach brings clear advantages over the linear methods in comparison. These results seem to support the theory that considering the Lipschitz-regularity of interpolators

in the learning has a positive effect on the generalization performance.

CHAPTER 5

CONCLUSION

Recent progress on the information technologies have provided people to collect several types of data from different sources. Multi-modal data sets have been constructed through the combination of various features, which allow a machine learning algorithm to create powerful models to analyse and discriminate data. The latest studies discover that single modal machine learning methods are likely to become less successful than multi-modal learning techniques. The success of multi-modal machine learning is highly related to the accordance between modalities and the complementary properties of the feature spaces in different modalities.

Co-training, multiple kernel learning, subspace learning and deep learning are four essential branches of the multi-modal machine learning approaches. However, they may encounter various problems such as lack of joint optimization, non-flexibility on real world data and limitations of linear transformations on data sets with nonlinear and intricate geometries. In order to deal with these issues, we have proposed in this thesis a supervised nonlinear projection based multi-modal learning algorithm building on the promising results of [13] and [14]. Our algorithm relies on the interpolator regularity, class similarities and discriminations of intra and inter modality relationships.

We tested our projection-based algorithm on several well-known multi-modal data sets in classification and retrieval tasks. We compared the performances of some important multi-modal learning approaches in the literature according to frequently used metrics.

In the first experiment, we used 720 face images of 10 participants provided by MIT

CBCL community. Firstly, we have studied how the algorithm parameters affect the classification performance. We found that our algorithm reaches its best results after a few iterations. Thus, we have observed that our algorithm can be terminated after a small number of iterations in order to decrease the computational load. Furthermore, we analysed the effects of the weight parameters used in our objective function. Our findings demonstrated that the weight parameters for the intra and the inter modalities between-class dissimilarities should be high. On the other hand, the weight parameter for the kernel function norm constraint needs to be low in order to balance the effects of the optimization terms. Additionally, the weight parameters for the kernel scale parameter constraint and the inter-modality class similarity need to be close to 1. Secondly, we measured the performance of classification at different embedding dimensions and observed that the algorithm was more successful when the embedding dimension was low. For this reason, we preferred to keep the embedding dimension small in order to reduce the computational complexity. Lastly, we compared the misclassification errors for various numbers of training samples. Our algorithm was observed to be among the methods providing the highest classification accuracy.

In the second experiment, we tested the methods on the Wikipedia data set, which includes 2866 image-text pairs. We used the SIFT features of the images and the Latent Dirichlet Allocation model parameters of the texts as two different modalities. We used the MAP metric to evaluate the algorithm performances. We conducted retrieval experiments on this data set by dividing it into 1300 training and 1566 test samples. The experimental results suggested that our algorithm was more successful than the CCA, GMLDA and JFSSL algorithms.

As a last experiment, we studied on a challenging PASCAL VOC data set that is formed with different feature vectors such as the SIFT, GIST vectors of the images and the word frequencies of the texts. We tested the retrieval performance of the methods and acquired better MAP scores with the proposed method than the other baseline methods. This is owed to the flexibility of the nonlinear projections learnt with the proposed algorithm, which can be generalized to the whole data space. Linear projection techniques are more likely to fail on this challenging data set.

All experimental findings clearly indicate that our newly proposed solution can be a

good alternative for classification or retrieval tasks on real world multi-modal data sets, since it aims to increase within-class connection and between-class separation at the same time while benefiting from the Lipschitz continuity of the interpolator generalizing the embedding to the whole data space. Although the optimization problem consists of five major terms and a constraint, it can be easily solved with an iterative algorithm. Moreover, the objective function converges within a few iterations and a small embedding dimension is sufficient to achieve high performance. Thereby, the computational load is relatively low and the probability of success is high for bigger data sets.

One of our future directions is to test our algorithm on bigger data sets. Additionally, the extension to incomplete data sets can be another challenge that we might address. Nonexistent correspondences of samples in different modalities may affect the algorithm performance. Another point to study is how our method extends to more than two modalities. The optimization formula is appropriately established to be fitted with three or more modalities cases. These cases may be good test scenarios for the capabilities of our algorithm. All these issues can be considered in the possible future research efforts related to this thesis study.

REFERENCES

[1] J. Herrlin, "The future of transport? shared services built on data." Online, jan 2017. Available: https://theodi.org/article/the-future-of-transport-shared-services-built-on-data/.

[2] C. Xu, D. Tao, and C. Xu, "A survey on multi-view learning," *CoRR*, vol. abs/1304.5634, 2013.

[3] K. Wang, R. He, L. Wang, W. Wang, and T. Tan, "Joint feature selection and subspace learning for cross-modal retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2010–2023, 2016.

[4] T. Xia, D. Tao, T. Mei, and Y. Zhang, "Multiview spectral embedding," *IEEE Trans. Systems, Man, and Cybernetics, Part B*, vol. 40, no. 6, pp. 1438–1446, 2010.

[5] A. L. Samuel, "Some studies in machine learning using the game of checkers," *IBM Journal of Research and Development*, vol. 3, pp. 210–229, 1959.

[6] A. Smola, *Introduction to Machine Learning*. Cambridge University Press, 2008.

[7] D. Faggella, "Machine learning in finance – present and future applications," jan 2019. Available: https://emerj.com/ai-sector-overviews/machine-learning-in-finance/.

[8] T. Kubota, "Deep learning algorithm does as well as dermatologists in identifying skin cancer." Online, jan 2017. Available: https://news.stanford.edu/2017/01/25/artificial-intelligence-used-identify-skin-cancer/.

[9] K. Wang, Q. Yin, W. Wang, S. Wu, and L. Wang, "A comprehensive survey on cross-modal retrieval," *CoRR*, vol. abs/1607.06215, 2016.

[10] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: An overview of methods, challenges, and prospects," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1449–1477, 2015.

[11] H. Messer, A. Zinevich, and P. Alpert, "Environmental monitoring by wireless communication networks," *Science*, vol. 312, no. 5774, pp. 713–713, 2006.

[12] T. Baltrusaitis, C. Ahuja, and L. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, 2019.

[13] E. Vural and C. Guillemot, "A study of the classification of low-dimensional data with supervised manifold learning," *Journal of Machine Learning Research*, vol. 18, pp. 157:1–157:55, 2017.

[14] C. Ornek and E. Vural, "Nonlinear supervised dimensionality reduction via smooth regular embeddings," *Pattern Recognition*, vol. 87, pp. 55–66, 2019.

[15] K. Nigam and R. Ghani, "Analyzing the effectiveness and applicability of co-training," in *Proceedings of the Ninth International Conference on Information and Knowledge Management*, CIKM '00, (New York, NY, USA), pp. 86–93, ACM, 2000.

[16] U. Brefeld and T. Scheffer, "Co-em support vector learning," in *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04, (New York, NY, USA), pp. 16–, ACM, 2004.

[17] Z.-H. Zhou and M. Li, "Semi-supervised regression with co-training," in *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, IJCAI'05, (San Francisco, CA, USA), pp. 908–913, Morgan Kaufmann Publishers Inc., 2005.

[18] U. Brefeld, T. Gärtner, T. Scheffer, and S. Wrobel, "Efficient co-regularised least squares regression," in *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, (New York, NY, USA), pp. 137–144, ACM, 2006.

[19] S. Bickel and T. Scheffer, "Multi-view clustering," in *Proceedings of the 4th IEEE International Conference on Data Mining (ICDM 2004), 1-4 November 2004, Brighton, UK*, pp. 19–26, 2004.

[20] S. Yu, B. Krishnapuram, R. Rosales, and R. B. Rao, "Bayesian co-training," *Journal of Machine Learning Research*, vol. 12, pp. 2649–2680, 2011.

[21] N. Poh, J. Kittler, and A. Rattani, "Handling session mismatch by fusion-based co-training: An empirical study using face and speech multimodal biometrics," in *2014 IEEE Symposium on Computational Intelligence in Biometrics and Identity Management, CIBIM 2014, Orlando, FL, USA, December 9-12, 2014*, pp. 81–86, 2014.

[22] R. Hinami, J. Liang, S. Satoh, and A. G. Hauptmann, "Multimodal co-training for selecting good examples from webly labeled video," *CoRR*, vol. abs/1804.06057, 2018.

[23] R. Yan and M. R. Naphade, "Multi-modal video concept extraction using co-training," in *Proceedings of the 2005 IEEE International Conference on Multimedia and Expo, ICME 2005, July 6-9, 2005, Amsterdam, The Netherlands*, pp. 514–517, 2005.

[24] K. P. Bennett, M. Momma, and M. J. Embrechts, "Mark: A boosting algorithm for heterogeneous kernel models," in *Proc. KDD-2002: Knowledge Discovery and Data Mining*, pp. 24–31, 2002.

[25] G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M. Jordan, "Learning the kernel matrix with semidefinite programming," *Journal of Machine Learning Research*, vol. 5, pp. 27–72, 2004.

[26] F. R. Bach and G. Lanckriet, "Multiple kernel learning, conic duality, and the SMO algorithm," in *Proc. 21st International Conference on Machine Learning*, 2004.

[27] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf, "Large scale multiple kernel learning," *Journal of Machine Learning Research*, vol. 7, pp. 1531–1565, 2006.

[28] Z. Xu, R. Jin, H. Yang, I. King, and M. Lyu, "Simple and efficient multiple kernel learning by group Lasso," in *Proc. 27th International Conference on Machine Learning*, pp. 1175–1182, 2010.

[29] T. Lange and J. M. Buhmann, "Fusion of similarity data in clustering," in *Advances in Neural Information Processing Systems 18 [Neural Information Processing Systems, NIPS 2005, December 5-8, 2005, Vancouver, British Columbia, Canada]*, pp. 723–730, 2005.

[30] H. Valizadegan and R. Jin, "Generalized maximum margin clustering and unsupervised kernel learning," in *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*, pp. 1417–1424, 2006.

[31] H. Zeng and Y. Cheung, "Kernel learning for local learning based clustering," in *Artificial Neural Networks - ICANN 2009, 19th International Conference, Limassol, Cyprus, September 14-17, 2009, Proceedings, Part I*, pp. 10–19, 2009.

[32] H. Zeng and Y. Cheung, "Feature selection and kernel learning for local learning-based clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1532–1547, 2011.

[33] B. Zhao, J. T. Kwok, and C. Zhang, "Multiple kernel clustering," in *Proceedings of the SIAM International Conference on Data Mining, SDM 2009, April 30 - May 2, 2009, Sparks, Nevada, USA*, pp. 638–649, 2009.

[34] G. Tzortzis and A. Likas, "Kernel-based weighted multi-view clustering," in *12th IEEE International Conference on Data Mining, ICDM 2012, Brussels, Belgium, December 10-13, 2012*, pp. 675–684, 2012.

[35] S. Sun, "A survey of multi-view machine learning," *Neural Computing and Applications*, vol. 23, no. 7-8, pp. 2031–2038, 2013.

[36] N. Rasiwasia, D. Mahajan, V. Mahadevan, and G. Aggarwal, "Cluster canonical correlation analysis," in *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, AISTATS 2014, Reykjavik, Iceland, April 22-25, 2014*, pp. 823–831, 2014.

[37] V. Ranjan, N. Rasiwasia, and C. V. Jawahar, "Multi-label cross-modal retrieval," in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 4094–4102, 2015.

[38] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, "A multi-view embedding space for modeling internet images, tags, and their semantics," *International Journal of Computer Vision*, vol. 106, no. 2, pp. 210–233, 2014.

[39] J. B. Tenenbaum and W. T. Freeman, "Separating style and content with bilinear models," *Neural Computation*, vol. 12, no. 6, pp. 1247–1283, 2000.

[40] R. Rosipal and N. Krämer, "Overview and recent advances in partial least squares," in *Subspace, Latent Structure and Feature Selection, Statistical and Optimization, Perspectives Workshop, SLSFS 2005, Bohinj, Slovenia, February 23-25, 2005, Revised Selected Papers*, pp. 34–51, 2005.

[41] T. Diethe, D. R. Hardoon, and J. Shawe-Taylor, "Constructing nonlinear discriminants from multiple data views," in *Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part I*, pp. 328–343, 2010.

[42] Q. Chen and S. Sun, "Hierarchical multi-view fisher discriminant analysis," in *Neural Information Processing, 16th International Conference, ICONIP 2009, Bangkok, Thailand, December 1-5, 2009, Proceedings, Part II*, pp. 289–298, 2009.

[43] A. Sharma, A. Dubey, P. Tripathi, and V. Kumar, "Pose invariant virtual classifiers from single training image using novel hybrid-eigenfaces," *Neurocomputing*, vol. 73, no. 10-12, pp. 1868–1880, 2010.

[44] A. Sharma, A. Kumar, H. Daumé, and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2160–2167, 2012.

[45] J. Hu, J. Lu, and Y. Tan, "Sharable and individual multi-view metric learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 9, pp. 2281–2288, 2018.

[46] D. Hidru and A. Goldenberg, "EquiNMF: Graph regularized multiview nonnegative matrix factorization," *arXiv Preprint*, 2014.

[47] M. Shao, D. Kit, and Y. Fu, "Generalized transfer subspace learning through low-rank constraint," *International Journal of Computer Vision*, vol. 109, no. 1-2, pp. 74–93, 2014.

[48] Z. Ding, M. Shao, and Y. Fu, "Latent low-rank transfer subspace learning for missing modality recognition," in *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada.*, pp. 1192–1198, 2014.

[49] S. Li and Y. Fu, "Robust subspace discovery through supervised low-rank constraints," in *Proceedings of the 2014 SIAM International Conference on Data Mining, Philadelphia, Pennsylvania, USA, April 24-26, 2014*, pp. 163–171, 2014.

[50] Z. Ding and Y. Fu, "Low-rank common subspace for multi-view learning," in *2014 IEEE International Conference on Data Mining, ICDM 2014, Shenzhen, China, December 14-17, 2014*, pp. 110–119, 2014.

[51] Y. LeCun, Y. Bengio, and G. E. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[52] T. H. Do, D. M. Nguyen, E. Tsiligianni, B. Cornelis, and N. Deligiannis, "Multiview deep learning for predicting twitter users' location," *CoRR*, vol. abs/1712.08091, 2017.

[53] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in *Proc. ACM International Conference on Multimedia*, pp. 7–16, 2014.

[54] W. Wang, X. Yang, B. C. Ooi, D. Zhang, and Y. Zhuang, "Effective deep learning-based multi-modal retrieval," *VLDB J.*, vol. 25, no. 1, pp. 79–101, 2016.

[55] L. Castrejón, Y. Aytar, C. Vondrick, H. Pirsiavash, and A. Torralba, "Learning aligned cross-modal representations from weakly aligned data," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2940–2949, 2016.

[56] Y. Niu, Z. Lu, J. Wen, T. Xiang, and S. Chang, "Multi-modal multi-scale deep learning for large-scale image annotation," *IEEE Trans. Image Processing*, vol. 28, no. 4, pp. 1720–1731, 2019.

[57] N. Bouteldja, D. Merhof, J. Ehrhardt, and M. P. Heinrich, "Deep multi-modal encoder-decoder networks for shape constrained segmentation and joint representation learning," in *Bildverarbeitung für die Medizin 2019 - Algorithmen - Systeme - Anwendungen. Proceedings des Workshops vom 17. bis 19. März 2019 in Lübeck*, pp. 23–28, 2019.

[58] S. Kamada and T. Ichimura, "Fast training of adaptive structural learning method of deep learning for multi modal data," *IJCIStudies*, vol. 7, no. 3/4, pp. 169–191, 2018.

[59] J. Garg, S. V. Peri, H. Tolani, and N. C. Krishnan, "Deep cross modal learning for caricature verification and identification (cavinet)," in *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018*, pp. 1101–1109, 2018.

[60] X. Feng, *Multi-modal and deep learning for robust speech recognition*. PhD thesis, Massachusetts Institute of Technology, Cambridge, USA, 2017.

[61] Y. Wei, Y. Zhao, C. Lu, S. Wei, L. Liu, Z. Zhu, and S. Yan, "Cross-modal retrieval with CNN visual features: A new baseline," *IEEE Trans. Cybernetics*, vol. 47, no. 2, pp. 449–460, 2017.

[62] R. van Hassel, "Ade (g1156) spring 2006 handout 3: Lipschitz condition and lipschitz continuity." Online, 2016.

[63] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[64] D. Cai, X. He, K. Zhou, J. Han, and H. Bao, "Locality sensitive discriminant analysis," in *IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007*, pp. 708–713, 2007.

[65] H. Chen, H. Chang, and T. Liu, "Local discriminant embedding and its variants," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 20-26 June 2005, San Diego, CA, USA*, pp. 846–853, 2005.

[66] Q. Gao, J. Ma, H. Zhang, X. Gao, and Y. Liu, "Stable orthogonal local discriminant embedding for linear dimensionality reduction," *IEEE Trans. Image Processing*, vol. 22, no. 7, pp. 2521–2531, 2013.

[67] B. Raducanu and F. Dornaika, "A supervised non-linear dimensionality reduction approach for manifold learning," *Pattern Recognition*, vol. 45, no. 6, pp. 2432–2444, 2012.

[68] W. K. Wong and H. T. Zhao, "Supervised optimal locality preserving projection," *Pattern Recognition*, vol. 45, no. 1, pp. 186–197, 2012.

[69] "MIT-CBCL face recognition database." Available: http://cbcl.mit.edu/software-datasets/heisele/facerecognition-database.html.

[70] N. Rasiwasia, J. C. Pereira, E. Coviello, G. Doyle, G. Lanckriet, R. Levy, and N. Vasconcelos, "A New Approach to Cross-Modal Multimedia Retrieval," in *ACM International Conference on Multimedia*, pp. 251–260, 2010.

[71] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[72] K. Wang, Q. Yin, W. Wang, S. Wu, and L. Wang, "A comprehensive survey on cross-modal retrieval," *arXiv Preprint*, 2016.

[73] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results." http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.