Discovering a Domain Knowledge Representation for Image Grouping: Multimodal Data Modeling, Fusion, and Interactive Learning

by

Xuan Guo, B.S.

DISSERTATION

A Dissertation Submitted in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy
in the B. Thomas Golisano College of
Computing and Information Sciences

Rochester Institute of Technology

Rochester, NY

August 2017

Discovering a Domain Knowledge Representation for Image Grouping: Multimodal Data Modeling, Fusion, and Interactive Learning

by
Xuan Guo

Committee Approval:

We, the undersigned committee members, certify that we have advised and/or supervised the candidate on the work described in this dissertation. We further certify that we have reviewed the dissertation manuscript and approve it in partial fulfillment of the requirements of the degree of Doctor of Philosophy in Computing and Information Sciences.

Prof. Anne R. Haake — Date
Dissertation Co-advisor

Prof. Qi Yu — Date
Dissertation Co-advisor

Prof. Cecilia Ovesdotter Alm — Date
Dissertation Committee

Prof. Rui Li — Date
Dissertation Committee

Prof. Pengcheng Shi — Date
Dissertation Committee

Prof. Daniel Phillips — Date
Dissertation Defense Chairperson

Certified by:

Prof. Pengcheng Shi — Date
Director, Computing and Information Sciences

Abstract

Discovering a Domain Knowledge Representation for Image Grouping: Multimodal Data Modeling, Fusion, and Interactive Learning

Publication No. _____

Xuan Guo, Ph.D.
Rochester Institute of Technology, 2017

Supervisors: Dr. Anne R. Haake
Dr. Qi Yu

In visually-oriented specialized medical domains such as dermatology and radiology, physicians explore interesting image cases from medical image repositories for comparative case studies to aid clinical diagnoses, educate medical trainees, and support medical research. However, general image classification and retrieval approaches fail in grouping medical images from the physicians' viewpoint. This is because fully-automated learning techniques cannot yet bridge the gap between image features and domain-specific content for the absence of expert knowledge. Understanding how experts get information from medical images is therefore an important research topic.

As a prior study, we conducted data elicitation experiments, where physicians were instructed to inspect each medical image towards a diagnosis while describing image content to a student seated nearby. Experts' eye movements and their verbal descriptions of the image content were recorded to capture various aspects of expert image understanding. This dissertation aims at an intuitive approach to extracting expert knowledge, which is to find patterns in expert data elicited from image-based diagnoses. These patterns are useful to understand both the characteristics of the medical images and the experts' cognitive reasoning processes.

The transformation from the viewed raw image features to interpretation as domain-specific concepts requires experts' domain knowledge and cognitive reasoning. This dissertation also approximates this transformation using a matrix factorization-based framework, which helps project multiple expert-derived data modalities to high-level abstractions.

To combine additional expert interventions with computational processing capabilities, an interactive machine learning paradigm is developed to treat experts as an integral part of the learning process. Specifically, experts refine medical image groups presented by the learned model locally, to incrementally re-learn the model globally. This paradigm avoids the onerous expert annotations for model training, while aligning the learned model with experts' sense-making.

Acknowledgments

I would like to express the deepest appreciation to my advisors, Profs. Anne R. Haake and Qi Yu, for the support, encouragement, inspiration and guidance they provided me throughout my PhD study and life.

I am grateful to my dissertation committee: Profs. Cecilia Ovesdotter Alm, Rui Li, and Pengcheng Shi for their insightful comments.

My appreciation goes to my colleagues and good friends Jingjia Xu, Wenbo Wang, Tong Liu, Mohamed Elshrif, Jwala Dhamala, Ruslan Dautov, Katy Tarrit, Preethi Vaidyanathan, and Dong Wang for their priceless presence and support.

Lastly, I would like to express my gratitude to my dear parents for their support, love, presence and patience.

The work presented in this dissertation is supported by grants from the National Science Foundation, and the National Institute of Health.

Table of Contents

List of Tables

List of Figures

List of Algorithms

Chapter 1

Introduction

1.1 Background

In visually-oriented specialized medical domains such as dermatology and radiology, physicians need to study and compare medical image cases to aid clinical diagnoses, support medical research, and educate medical trainees. This can be aided by computational systems that organize medical images according to the physicians' understanding of the image content. This dissertation proposes to understand the medical images from the perspective of experts' domain knowledge.

Since medical knowledge tends to be tacit and difficult to convey and obtain, medical training usually takes years of internships and specialized residencies. Medical training will benefit from approaches that can properly represent and visualize expert knowledge. This dissertation develops various representation learning strategies to extract expert knowledge from multimodal datasets collected from physicians when they inspect medical images. The datasets were constructed by the Human-Centric Multi-Modal Modeling (HCM^3) Lab through an interdisciplinary approach. More specifically, two data elicitation experiments were conducted by using a repository of dermato-

logical images (courtesy of Logical Images, Inc.) as visual stimuli. The images in Experiment I represent a wide range of dermatology diagnoses [6], whereas those in Experiment II focus on more examples of a few diagnoses. Dermatology was chosen as a testbed, as it is a visually-based medical specialty that requires specific, complex perceptual expertise. Following a modified master-apprentice approach [7], each participating physician was asked to describe the visual content of each image aloud, as if teaching a student who was seated nearby [8]. The physicians' speech and eye movements were recorded during the experiments. These experiments were approved by Rochester Institute of Technology's Institutional Review Board, and all participants provided informed consent before participating in the experiments.

The approach to gathering expert data during image inspection effectively traces how experts use their knowledge for image-based problem solving. This is because this approach involves a more natural task for physicians to perform in contrast to asking them to conform to predefined image annotation labels and rules. Figure 1.1 presents one of the images, and the experimental setup. Figure 1.2 contains one corresponding eye tracking data instance and diagnostic narrative transcribed from physicians' spoken descriptions. Similar to this example, all the spoken narratives were comprehensively transcribed with sequences of words labeled by time stamps using the speech analysis tool Praat [9]. Overall, the datasets comprise 1670 transcribed spoken narratives and the same number of eye tracking trials.

Additionally, 58 spoken narratives are labeled by physician-defined and

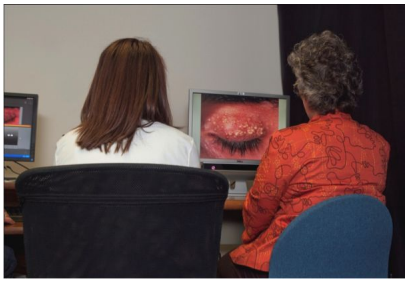
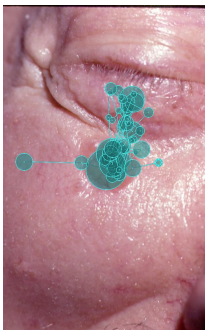


Figure 1.1: *Left*: One medical image case used in the study (diagnosed as *basal cell carcinoma*; image courtesy of Dr. Cara Calvelli), *Right*: image inspection, audio recording, and eye tracking.

-identified diagnostic thought units (TUs). These TUs cover the terminology to standardize the description of skin lesions, including lesion arrangement, distribution, texture, color, primary lesion type, and diagnosis [10]. For a follow-up study of the reasoning processes of the collected diagnostic verbal narratives, we recruited three dermatologists to evaluate the narratives from the 16 participating physicians in Experiment I in terms of their diagnostic correctness. A correctness score was assigned to each narrative, which balances the correctness of described primary lesion type, differential diagnosis, and final diagnosis. We refer to them as Type II thoughts to distinguish them (the indirect findings) from the rest TUs that are direct findings. The identified TUs are shown in Table 1.1, and one example annotation is in Figure 1.2-(b). In addition to the semantic relevance, there is a temporal correspondence between the data modalities, as both modalities of expert data were synchronously collected.



(a) A collected gaze data instance.

(SIL) um this is a (SIL) pearly lobulated (SIL) pink (SIL) papule with telangiectasias multi-lobulated papule with telangiectasias (SIL) on the upper cheek (SIL) of (SIL) an elderly individual (SIL) say nodular basal cell carcinoma it's a nodular basal cell carcinoma (SIL) uh (SIL) there is some surrounding sun damage (SIL) and surrounding telangiectasias

PRI ; SEC ; LOC ; DEM ; DX or DIF

(b) A dermatologist's diagnostic narrative for the image. *(SIL)* represents silent pause.

Figure 1.2: The example eye gaze data instance (a) and a diagnostic narrative annotated by thought unit labels (b) that correspond to the same image.

Given the specialized domain, in which expert knowledge is required to solve difficult problems (i.e., making diagnoses), the collected multimodal expert data are ideal observations for this dissertation research to learn representation of knowledge from human data. Besides, these datasets also reflect the difficulty of research in the field—numerous medical images are not an-

Table 1.1: Two types of thought units.

	Thought Unit Labels (Abbr.)	**Instances**
Type I	Patient DEMographics (DEM)	elderly, caucasian, woman
	Body LOCation (LOC)	arm, upper lip, knuckles
	Lesion CONfiguration (CON)	linear, annular, grouped
	SECondary finding (SEC)	crust, ulcer, erythematous
	Lesion DIStribution (DIS)	solitary, bilateral, extensive
Type II	PRImary lesion type (PRI)	papule, plaque, patch
	DIFferential diagnosis (DIF)	X, Y or Z
	Final Diagnosis (DX)	this is X
	RECommendations (REC)	P should Q

notated, meaning that we have to learn from a small annotated dataset for knowledge that can be generally utilized. One important finding during data collection is that the dermatology images can be difficult to inspect even for experts, because these photographic images often contain complex visual features which may or may not be relevant to making an accurate diagnosis. Moreover, in some cases the information in the image is not sufficient to support a final diagnosis where the physicians requested biopsy as additional evidence. These findings show the tight connection of the proposed research to the real-world clinical settings and educational image uses.

1.2 Problem Definition

Eye movements provide insights into experts' interests of the key visual features and perceptually important regions in the images. Existing work in the HCM^3 Lab shows the correlation between image feature distribution and eye fixation arrangement [11]. **These indicate the promising direction to understand image content from expert-derived data**.

Previous studies in HCM^3 also include developing computational modeling to discover hidden visual behavioral patterns from eye movement data [12, 13, 14]. Several distinctive types of patterns (i.e., Signature Patterns) were discovered by a model and verified by a domain expert [15, 16]. **This indicates that it is possible to extract expert knowledge from the collected behavioral data**.

A system of annotation for conceptual semantic units of thought was

developed with domain experts in order to examine participating physicians' diagnostic narration structure [8, 17]. A correspondence was found between the learned eye movement patterns and the conceptual units of thought by time-aligning the patterns with annotated narratives [12]. **This finding suggests another research direction taken in this dissertation; that is, to mathematically fuse multimodal expert data for comprehensive image understanding**.

We are learning from a small dataset (compared with the large volume of medical images in the field), and the collected multimodal expert data contain some data instances where the diagnoses are inaccurate. This makes the outcome (e.g., the learned data representation, and the estimated parameters) sensitive to random variations in data (i.e., overfitting). To include prior knowledge in the representation learning, **this dissertation also aims at developing a framework to receive additional inputs of domain knowledge through expert interactions**.

Given the elicited multimodal expert data and the promising research directions above, the objectives of this dissertation are—(1) To build models with the diagnostic verbal narratives for discovering expert-produced behavioral and cognitive patterns. (2) To develop a framework that integrates the features in multiple modalities for a unified data representation, which explains the observations in these modalities and their correspondences. (3) To implement a machine learning system that allows expert manipulation of images and uses such interactive inputs to improve the resulting model. These are es-

sentially inverse problems, where I discover experts' uses of domain knowledge and their diagnostic reasoning processes from the collected eye movements and transcribed verbal narratives.

1.3 Dissertation Contributions

- The visual stimuli (i.e., the image content) and the expert cognitive processing of the stimuli are interwoven. The collected expert data contain the variances intrinsically from both the images and the experts. This dissertation discovers interpretable behavioral patterns from expert data and discloses both the characters of the medical images and those of the participating experts.

- The multimodal data were collected during in-scenario experiments and hence the data reflect different aspects of the same cognitive processes—i.e., image-based diagnostic reasoning. In spite of the obvious semantic relevance across data modalities, the relationship between the modalities is hidden and needs modeled. This dissertation develops and studies a machine learning framework that fuses multimodal expert data, so as to recover the underlying conceptual and cognitive elements.

- The sparsity of the multimodal observations introduces an additional challenge for accurate estimation of model parameters. The studies in this dissertation incorporate knowledge resources as constraints in machine learning models to tackle the *small data* issue—using external

medical knowledge resources (for representation learning) and additional expert inputs (for interactive machine learning).

1.4 Dissertation Organization

The following chapters in this dissertation are organized as follows (see Figure 1.3): Chapter 2 first introduces the background and related studies, including the domain knowledge and visual perception in medical diagnosis, and the idea of human computation through multimodal expert data. Chapter 2 also reviews various fundamental approaches that are used and referred to in later chapters. Chapter 3 models the collected diagnostic verbal narratives for medically meaningful topics. Section 3.2 presents the narrative processing with the Unified Medical Language System (UMLS) [18] to extract medical terms that are used for modeling and analysis in the rest of the dissertation. Section 3.3 presents narrative clustering based on a topic modeling approach. Section 3.4 develops two lexical metrics that are useful to explore the attributes of physician groups and their diagnostic relevance based on the verbal narratives. Section 3.5 develops a hierarchical dynamic model to recognize the narration patterns that match physicians' diagnostic reasoning stages and are useful to predict diagnostic correctness. Chapter 4 reports on the mixture components identified in physicians' eye movement data that match the medical abnormalities in the images. This finding supports two other studies in the chapter to develop a human-centered information retrieval (HCIR) system that relies on eye tracking and verbal inputs (Section 4.3) and to use

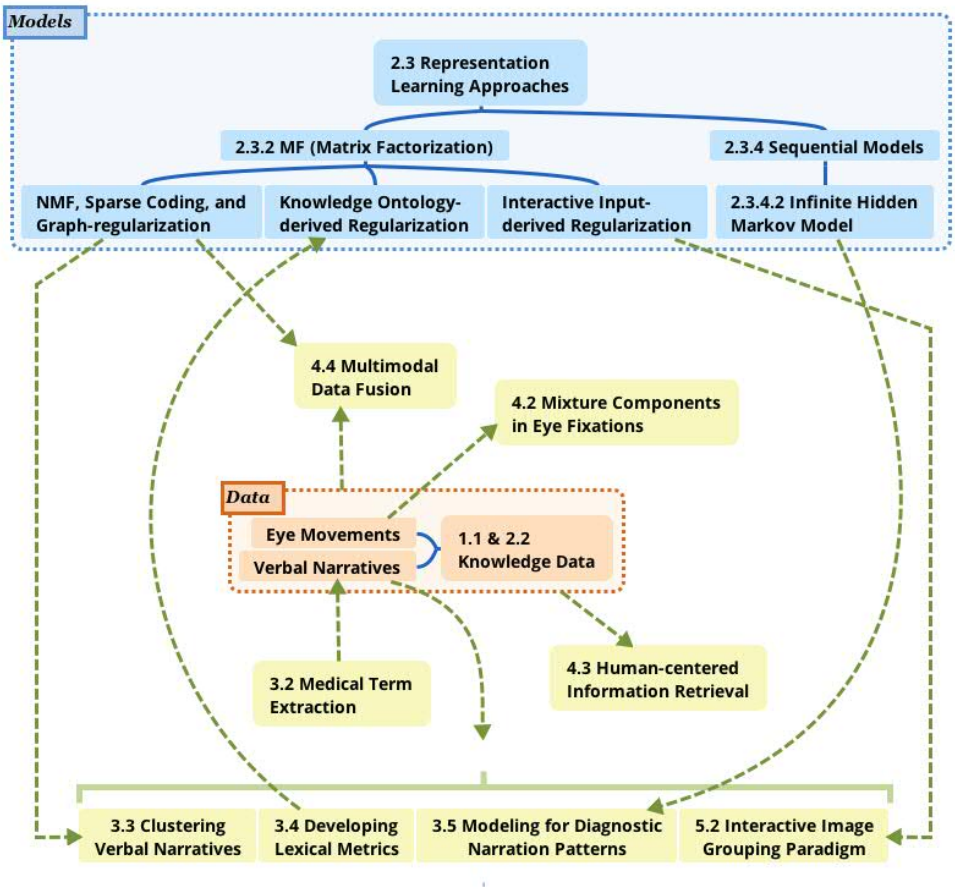


Figure 1.3: Connections between chapters in this dissertation.

multimodal data for a set of unified latent variables (Section 4.4). Chapter 5 describes an interactive machine learning paradigm developed for the case of medical image grouping, where additional expert inputs can be used as constraints to guide the machine learning processes. Chapter 6 summarizes this dissertation and proposes future work in the area.

Chapter 2

Related Work

Since many medical images are inherently complex and noisy due to both photographic inconsistency and different presentations of even the same medical condition, grouping relevant medical images into semantically related and meaningful groups has been a long-standing challenge. This chapter first lays the theoretical foundation for research in the field of medical diagnosis in Section 2.1, followed by describing the external language tools and multiple human sensors from the *human computation* perspective (Section 2.2). Section 2.3 then reviews the technical building blocks that can be used with the tools and data to study expert knowledge and understand medical images. Chapters 3, 4, and 5 use these building blocks for studies of diagnostic reasoning and medical image understanding, and the outcomes are useful to support medical image grouping.

2.1 Image-based Diagnostic Reasoning

Since the content in medical images is beyond colors, textures, and shapes, fully automated computer vision algorithms that only use low-level image features fail to capture the domain-specific semantics. To parse the complex

semantics and represent a medical image by a feature vector, we need to understand the image content, especially the key content based on which a physician can make a diagnosis about the image. However, the task of medical image understanding remains challenging, because it requires domain knowledge and perceptual expertise.

In the medical field, diagnostic reasoning processes can be explained by two cognitive systems in the *dual-process theory*, namely *intuitive* and *analytical* [19]. However, current research on medical diagnosis relies on research interviews or clinical chart records [20], reports of clinicians [21], and physicians' response time [22], and hence overlooks much detailed information.

To contribute to both fields above, we jointly study the content in medical images and physicians' diagnostic decision-making. Specifically, we collected multimodal data from physicians when they were engaged in diagnostic reasoning based on medical images (Section 1.1). To disentangle the underlying factors that relate to expert interpretation of image content, In this dissertation we model the collect data in Sections 3.3, 3.5, and 4.4. The model outcomes help understand both physicians' diagnostic reasoning processes and their interpretations of image content by domain knowledge.

2.2 Conceptual Knowledge, Perceptual Expertise, and Human Computation

The general computer vision algorithms that automatically detect semantic elements from salient image features cannot capture the domain-specific se-

mantics in medical images. Parsing and using the semantics in medical images requires external knowledge [23]. One resource of domain knowledge is the UMLS, which provides access to medical concepts and concept relations in the medical domains. The domain experts constitute another knowledge resource.

2.2.1 Unified Medical Language System

Ontology resources such as WordNet [24], VerbNet [25] and ImageNet [26] have been used in general domains for extracting semantics in texts and images. The UMLS can serve such a purpose in medical domains. It is a knowledge source for medical terminology research and information retrieval [27], constituting the largest existing semantic network of medical terms and their lexical relations [18, 28]. As an ontology of medical concepts it has been used to process clinical records, to relate or disambiguate medical terminologies, and to serve as a knowledge base for health care systems [18]. It has also been used to assist feature engineering to tackle the intrinsic problem of data sparsity in clinical texts [29, 30].

To represent expert knowledge from linguistic data, we have preprocessed the data by programming with MetaMap, which is a knowledge-intensive tool that automatically annotates biomedical text tokens by UMLS Metathesaurus concepts [31]. Our program first filters out the non-medical terms in the verbal narratives, and then detects and reconstructs medical multiwords (phrases) by joining adjacent medical words. See Table 3.1 for an example. In

this manner, each spoken narrative is segmented as a sequence of words and multiword expressions, which can be used for further analysis and modeling.

2.2.2 Natural language and domain knowledge

Language is the primary conduit to express meaning. Our diagnostic verbal image descriptions, collected during medical image inspections, form a non-trivial corpus with different levels of language features. For example, speech features can be used to study experts' certainty in decision-making [32, 33, 34]; natural language can be processed and statistically analyzed at the lexical level to reveal humans' decision styles [35]; additionally, sense relationship can be extracted using an ontology [36, 37]. This effort focuses on sense-based representation as a reflection of human knowledge.

Studies in both computer vision and natural language processing use the correspondence between image content and human annotations to reveal semantics [38]. In natural scene image corpora, the annotations summarize the meanings of visual content [39]. In return, the visual content in images provides semantic context to disambiguate a lexical item (e.g., a *crane* as a bird vs. that in a construction field). For example, Wu et al., defined the Flickr distance to measure the relationship between semantic concepts (objects, scenes) in the visual domain [40]. For each concept, a collection of images are obtained from Flickr, and their visual characteristics are captured. Using information in images helps capture the visual relationship between concepts. Their study customized the meaning of concepts based on a large set of web images, and

the resulting concept network is more statistically coherent to humans' current knowledge. There are more examples integrating verbal metadata with image features [39]. A case of using language for characterizing the meaning of images is shown in Li et al.'s study [41]. Using language tags of image regions, an ontology of related concepts are introduced to achieve hierarchical annotation.

Natural language data provide multiple challenges: (1) Language is, by nature, sparse [42]. In most linguistic data sets, the vast majority of lexical items tend to occur rarely, and speakers can express similar meaning in a variety of ways, both syntactically and lexically. (2) Semantic ambiguity occurs in language data. The understanding and interpretation of ambiguous language data depends on domain knowledge. (3) The difference in narration styles among users results in variability that obscures common strategies of diagnostic reasoning. (4) Language data are influenced by the mode in which they were produced. Naturally occurring speech data differ substantially from standard written text data.

Physicians in specialties such as radiology and dermatology have developed visual perceptual expertise. They better recognize domain-specific patterns than unsupervised algorithms that lack guidance from medical knowledge [13, 43]. However, it is time consuming and impractical for physicians to manually annotate medical images since these images (1) can be stored in large-scale image databases with a large and rapidly growing number of digital images, and (2) may reside within individual medical practices or small, distributed, non-standardized databases.

Researchers have made efforts to incorporate domain knowledge in image clustering [44]. However, truly understanding physicians' use of knowledge (especially during image-based diagnosis) remains a challenging task, because visual diagnostic reasoning is a complex interaction of domain knowledge, perceptual expertise, reasoning processes [45], and idiosyncratic visual information in the image case being inspected by physicians.

We elicit expert knowledge by collecting data from physicians as they engage in medical image inspection. This approach involves a more natural task for experts to perform in contrast to asking them to conform to predefined image annotation labels and rules.

We exploit human experts' knowledge to facilitate medical image grouping by applying a methodology that is more objective and automated than current research [22, 46, 47]. The intuition is that the meaning of a medical image is expected to be mirrored by the spoken narrative of a physician when s/he describes the image during a diagnostic process. In this way, we incorporate physicians' domain knowledge, obtained from years of systematic study and clinical training, to achieve more effective medical image grouping.

2.2.3 Eye tracking and visual perception

People are exposed to plenty of visual information everyday, but relatively little of that information is processed due to our reliance on prior experience. Humans shift the point of regard to regions requiring high resolution to gather the sensory information from the world that is required for visual perception.

As a reflection of human responses to visual information, eye movements reveal the interaction between image features and human visual attention. It highlights the important visual information perceived by human observers.

Eye movements can be described as a combination of *fixations* and *saccades*. Fixations occur when the eyes remain at a particular spatial location in a visual stimulus, typically over a minimum duration of 100-200 milliseconds [48]. To re-orient the eye to other locations of interest, the eye makes rapid, ballistic movements known as saccades. Many eye tracking features/methods have been defined based on statistical analysis of fixations and saccades in raw data, including comparison of scanpaths [49], saliency maps [50], and more complicated scanning patterns [12]. Although eye movements cannot completely explain visual cognitive processes, many studies of individuals' eye movements, as they perform tasks, have established relationships between visual attention and cognition [51].

Eye movements are influenced by both bottom-up visual processing and top-down search tasks and knowledge, so eye tracking opens a window to explore human visual information gathering guided by knowledge and intentions [52]. In particular, human eye movements are a combination of visual input and several cognitive systems, including short-term memory for previously attended information in the current fixated position, stored long-term visual, spatial and semantic information about other similar visual cases (knowledge), and the goals and plans of the viewer (task) [53]. Fixations on particular parts of an image are guided by semantic informativeness of those regions rather than

the structural information inherent in the image. These attributes of human eye movements result in applications of eye tracking in various studies. For example, Kunze et al. inferred participants' language reading level from their reading speed and fixation duration [54]. Human gaze has been utilized to extract regions of interest (ROIs) of an image to perform attention-based image retrieval [55]. Visual attention has also been used as feedback in web searches [56], for predicting salient regions of web pages [57], and to indicate different levels of domain expertise by disclosing how experts vs. non-experts behave visually when they are searching for task-relevant clues [58].

In vision-based complex problem-solving scenarios, such as image use, human experts' tacit knowledge is key to understanding and can be applied toward enhancing computer vision algorithms. Eye tracking helps capture such knowledge, because visual strategies are executed at a level below conscious awareness and eye tracking is able to provide information that is not available through methods such as introspection. Eye tracking a group of experts allows us to study experts' subconscious image viewing behaviors in common by objectively measuring their eye movements.

2.3 Representation Learning Approaches

The data sparseness problem generally exists in natural language from various domains. For example, there are a large number of distinct terms in our dataset, because various naming preferences are used by physicians. Representation learning is a sub-field in machine learning to address this issue by

learning the transformations of raw data to compact and meaningful representations. It has been successful in various domains. For example, topics can be learned from the terms in documents [59], and levels of objects (e.g., edges, parts of faces, and faces) can be learned from pixels in face images [60]. Representation learning can also be applied to learn meaningful representations in audio signals and haptic data [61]. These learned new features are usually in a more compact space and hence reduce the computational burden of classification and prediction that follows.

Another use of representation learning is to visualize high-dimensional data in a user panel as a low-dimensional embedding through techniques such as multidimensional scaling (MDS) [62] and t-distributed stochastic neighborhood embedding (t-SNE) [63]. This allows users to inspect, understand and refine a large volume of data in a simple feature representation through interactions.

In order to systematically arrive at a preferred representation, representation learning essentially applies mathematical operations to the original feature space. Constraints are usually designed in the objectives to keep only those features that co-vary the most with respect to outcomes of interest. A discussion of constraints can be found in Section 2.3.2. Although representation learning can be extended as a framework to include the *human in the loop* for more preferred learning results (see Chapter 5), here we still consider it as a subset of unsupervised learning approaches for ease of explanations.

2.3.1 Classical models

In order to explain the general mechanisms of representation learning, we first describe a popular data clustering technique (i.e., *K-means*) and two unsupervised feature transformations (i.e., *principal component analysis* [PCA] and *independent component analysis* [ICA]). They were developed to learn a new feature representation for different purposes.

K-means is a procedure to partition data set $\boldsymbol{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\} \subset \mathbb{R}^{m \times n}$ into clusters $C_1, \cdots, C_k$ based on a specified number k, where each data point serves as a prototype of the cluster it belongs to.

$$\sum_{i=1}^{k} \sum_{j=1}^{n} z_{ij} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2 \tag{2.1}$$

where the indicator variable $z_{ij} = \begin{cases} 1, & \text{if } \mathbf{x}_j \in C_i \\ 0, & \text{otherwise.} \end{cases}$

K-means is implemented through iterative refinement—in each iteration, assigning each datum to the cluster with the nearest mean (E-step) and updating the new cluster centroids by minimizing sum of distance to all data in cluster (M-step; minimize Eq. (2.1) w.r.t., $\boldsymbol{\mu}$). The original feature space (cardinality m) is consequently projected to cluster labels (cardinality 1).

Principal Component Analysis is a procedure that transforms the original space of possibly correlated features into a space of linearly uncorrelated features called principal components [64]. This transformation is defined in such a way that the first principal component has the largest possible vari-

ance (i.e., it accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to (i.e., uncorrelated with) the preceding components. To retain the useful information (i.e., most of the variability in the data) and remove noise, the number of components is optimized resulting in usually only the first few components in the new space being kept. PCA generates a new representation in which the new features are not correlated. It is often used as data whitening to compress data and speed up the following learning process [65], or for visualization purposes [66].

PCA essentially centers the data matrix $\boldsymbol{X}$ first and then uses singular value decomposition (SVD) to decompose $\boldsymbol{X}$ into a diagonal matrix $\boldsymbol{\Sigma}$ of the same dimension as $\boldsymbol{X}$ and with nonnegative diagonal elements in decreasing order, and unitary matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ such that,

$$\boldsymbol{X}_{n\times m} \approx \boldsymbol{U}_{n\times n}\boldsymbol{\Sigma}_{n\times n}\boldsymbol{V}^{\top}_{n\times m} \tag{2.2}$$

where the rows of $\boldsymbol{V}^{\top}$ are eigenvectors of the covariance matrix $\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{X}^{\top}\boldsymbol{X}$. The matrix $\boldsymbol{\Sigma}$ is diagonal, with each element $\sigma_{ii} = \sqrt{\lambda_i}$ (the i^{th} eigenvalue). Rows of $\boldsymbol{U}$ are coefficients for basis vectors in $\boldsymbol{V}$.

The *manifold* hypothesis believes that the probability mass concentrates near regions that have a much smaller dimensionality than the original space where the data lives [67]. In other words, the data points are actually samples from a low-dimensional manifold that is embedded in a high-dimensional feature space. Under this hypothesis, PCA attempts to uncover

the underlying factors through linear transformation in order to find a low-dimensional representation of the data.

Independent Component Analysis is a computational method to separate a multivariate signal into independent subcomponents. ICA is a special case of blind source separation [68]. A common example application is the *cocktail party problem* of listening in on one person's speech in a noisy room. ICA generates a new representation in which the new features are independent signal sources. Each feature in the original space is a combination of these new independent features. ICA can be used to disentangle noise from the target signal [69].

The un-correlation used in PCA is characterized by,

$$E[xy] = E[x]E[y] \tag{2.3}$$

whereas the independence is given by,

$$E[f(x)g(y)] = E[f(x)]E[g(y)] \tag{2.4}$$

The uncorrelation only measures linear relationship, whereas the independence is stronger to measure the existence of any relationship.

2.3.2 Matrix factorization-based models

Matrix factorization (MF) seeks to learn a low rank approximation from an input data matrix using two factors [59]. Suppose we have a data matrix $\boldsymbol{X} \in \mathbb{R}^{n \times m}$ (m data instances described in an n-dimensional feature space).

The goal of matrix factorization is to generate a more compact and precise representation of the input matrix $\boldsymbol{X}$ by approximating it via the multiplication of two factor matrices $\boldsymbol{H} \in \mathbb{R}^{n \times k}$ and $\boldsymbol{C} \in \mathbb{R}^{k \times m}$.

$$\boldsymbol{X} \approx \boldsymbol{HC} \tag{2.5}$$

$$\min_{\boldsymbol{H},\boldsymbol{C}} \|\boldsymbol{X} - \boldsymbol{HC}\|_F^2 \tag{2.6}$$

where the Frobenius norm $\|\cdot\|_F^2$ is used to measure the error between the original input $\boldsymbol{X}$ and its low rank approximation $\boldsymbol{HC}$.

The matrix $\boldsymbol{H}$ (often referred to as *basis matrix*) can be viewed as a *dictionary*, because it reveals the transformation from the original feature space to latent variables that form a new basis. The matrix $\boldsymbol{C}$ (often referred to as *coefficient matrix*) represents the data instances by combinations of these latent variables. The number of latent variables, k, is usually small to enforce the low rank of both factor matrices.

Probabilistic interpretation of MF: Now assuming that the observed data matrix can be recovered by its low-rank approximation HC with additional Gaussian noise $\boldsymbol{\epsilon}$:

$$\boldsymbol{X} \approx \boldsymbol{HC} \iff \boldsymbol{X} - \boldsymbol{HC} = \boldsymbol{\epsilon} \tag{2.7}$$

where the $\boldsymbol{\epsilon}$ is a matrix to denote the reconstruction error. The entries ϵ_{ij}'s are assumed to be independent and identically distributed (i.i.d.) according to a Gaussian distribution with zero mean and variance σ_r^2 (r stands for

$\underline{\underline{r}}$*econstruction error*). Due to the independent assumption the joint distribution of all data items factorizes:

$$p(\boldsymbol{X}|\boldsymbol{H},\boldsymbol{C}) = \prod_i \prod_j \mathcal{N}(\boldsymbol{X}_{ij}; [\boldsymbol{HC}]_{ij}, \sigma_r^2) \tag{2.8}$$

$$= \prod_i \prod_j \frac{1}{\sqrt{2\pi}\sigma_r} e^{\left(-\frac{1}{2}\left(\frac{\boldsymbol{X}_{ij}-[\boldsymbol{HC}]_{ij}}{\sigma_r}\right)^2\right)} \tag{2.9}$$

Taking the logarithm yields the log-likelihood,

$$\ln p(\boldsymbol{X}|\boldsymbol{H},\boldsymbol{C}) = -nm\ln\left(\sqrt{2\pi}\sigma_r\right) - \underbrace{\frac{1}{2\sigma_r^2}\sum_i\sum_j(\boldsymbol{X}_{ij}-[\boldsymbol{HC}]_{ij})^2}_{D_E(\boldsymbol{X},\boldsymbol{HC})} \tag{2.10}$$

Maximum likelihood (ML) estimation: Maximizing the right hand side of Eq. (2.10) w.r.t. $\boldsymbol{H}$ and $\boldsymbol{C}$ is equivalent to minimizing the squared Euclidean distance $D_E(\boldsymbol{X}, \boldsymbol{HC})$ for MF.

MF framework to explain classical representation learning models: The framework of MF can be used to interpret the classical models in Section 2.3.1, such as the popular data clustering technique (i.e., *K-means*), and the unsupervised feature transformations (i.e., *PCA* and *ICA*).

K-means interpreted in MF-based framework: K-means can be re-written in the MF framework. It is a special case of matrix factorization to partition all documents in the corpus in K disjoint clusters, whereas the matrix factorization generally allows each document characterized by one or more topics. In other words, matrix factorization better models the problem

where each document may belong to several different clusters. Bauckhage has shown that the problem of k-means clustering can be understood as a constrained matrix factorization problem in the following form [70]:

$$\min_{\boldsymbol{Z}} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{Z}^\top \left(\boldsymbol{Z}\boldsymbol{Z}^\top\right)^{-1} \boldsymbol{Z}\|^2 \tag{2.11}$$

$$\text{s.t. } z_{ij} \in \{0, 1\} \text{ and } \sum_i z_{ij} = 1 \tag{2.12}$$

Eq. (2.12) means that among all the clusters (or classes), each data point only belongs to one cluster.

PCA interpreted in MF-based framework: To discover the patterns in data that cover major variance, PCA can be formulated as,

$$\min_{\boldsymbol{U},\boldsymbol{\Sigma},\boldsymbol{V}} \|\boldsymbol{X} - \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top\|_F^2 \tag{2.13}$$

$$\text{subject to: } \boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}, \boldsymbol{V}^\top\boldsymbol{V} = \boldsymbol{I}, \text{ and } \boldsymbol{\Sigma} \in \text{diag}^+ \tag{2.14}$$

where the first few rows in U and the first few columns in V are major patterns in the data-instance space and the feature space, respectively. Starting from the first principal component (PC), succeeding PCs find linear combinations of variables that correspond to the direction with highest variance under the constraint of it being orthogonal (uncorrelated) to preceding ones. This is basically a coordinate transformation, where the first few newly formed coordinate axes (variables) captures most of the variance present in the data. By absorbing $\boldsymbol{\Sigma}$ into $\boldsymbol{U}$, we can write $\boldsymbol{X} \approx \boldsymbol{U}\boldsymbol{V}^\top$.

ICA interpreted in MF-based framework: The mixture matrix $\boldsymbol{A}$ that mixes signals $\boldsymbol{S}$ and outputs observations $\boldsymbol{X}$ is unknown to us:

$$\boldsymbol{X} = \boldsymbol{A}\boldsymbol{S} \tag{2.15}$$

ICA recovers the signal $\hat{\boldsymbol{S}}$ by finding the matrix $\boldsymbol{W}$,

$$\hat{\boldsymbol{S}} = \boldsymbol{W}\boldsymbol{X} \tag{2.16}$$

where $\boldsymbol{W} = \boldsymbol{A}^{-1}$

The NMF with KL-divergence has a Bayesian formalization as the Gamma-Poisson model (GaP), and the GaP model is a form of ICA [71]. The similarity between NMF and ICA is also shown for face recognition [72].

Comparisons—NMF, PCA, and K-means: NMF has been compared with K-means and PCA by Lee and Seung [59], and their relationships were discussed by Ding et al. [73]. Based on unary coded prototypes, K-means allows only one hidden topic to be attributed to each data point. This limitation makes k-means not very useful for analysis. PCA allows the activation of multiple hidden variables, but it lacks obvious interpretation of the data. This is because PCA allows the arbitrary signs of the entries in the matrices, whereas subtractions may not make sense in context of some applications.

PCA retains orthogonality while relaxing non-negativity, and NMF forces non-negativity while relaxing orthogonality. In contrast to K-means and PCA, NMF only allows additive combinations of non-negative entries. The non-negativity constraints form the part-based representation, which naturally uncovers the inherent data structure.

Beside the classical representation learning models above, the probabilistic models such as pLSA and LDA can also be explained and interpreted in the MF-based framework. We will show more details in Section 2.3.2.6.

Most matrix factorization algorithms originate from minimizing an objective function in the simple form of $\|\boldsymbol{X} - \boldsymbol{HC}\|_F^2$ (squared Frobenius-norm of the difference between $\boldsymbol{X}$ and its approximation $\boldsymbol{HC}$) with respect to $\boldsymbol{H}$ and $\boldsymbol{C}$, and derive update rules by either a gradient-based method [2, 1], or a multiplicative rule-based method [3]. Section 2.3.2.1 reviews these methods in details for the case of NMF.

2.3.2.1 Nonnegative matrix factorization

By applying non-negative constraints of the coefficient matrix $\boldsymbol{C}$, the nonnegative matrix factorization (NMF) enables additive combination of the latent components $\mathbf{H}_{\cdot i}$'s according to weights $\mathbf{c}_i$.

$$\min_{\boldsymbol{H},\boldsymbol{C}\geq 0} \|\boldsymbol{X} - \boldsymbol{HC}\|_F^2 \tag{2.17}$$

$$\text{equivalent to: } \min_{\boldsymbol{H},\mathbf{c}_i\geq 0} \sum_i^m \|\mathbf{x}_i - \boldsymbol{H}\mathbf{c}_i\|_F^2 \tag{2.18}$$

Probabilistic interpretation of NMF: Let $\boldsymbol{X}_{ij}$ denote the elements in matrix $\boldsymbol{X}$. Assume a latent variable representation, $\boldsymbol{X}_{ij}$ can be written as,

$$\boldsymbol{X}_{ij} = \sum_k S_{ikj} \tag{2.19}$$

where S_{ikj}'s are random variables with densities $p(s_{ikj}|\boldsymbol{H}_{ik}, \boldsymbol{C}_{kj})$. NMF can

assume S_{ikj} to follow a Poisson distribution, i.e.,

$$S_{ikj} \sim \mathcal{PO}(s_{ikj}; \boldsymbol{H}_{ik}\boldsymbol{C}_{kj}) \tag{2.20}$$

and the latent sources $S_k = \{S_{ikj}\}$ can be analytically marginalized out to obtain the marginal likelihood $\log p(\boldsymbol{X}|\boldsymbol{H}, \boldsymbol{C})$ [74]. Cemgil solves the maximum-likelihood problem for $\{\boldsymbol{H}, \boldsymbol{C}\}$ using EM algorithm, which arrives at exactly the multiplicative update rules in Alg. 4 below.

Existing algorithms of NMF: The algorithms for solving the NMF problem fall in three main categories—(1) the gradient-based methods, (2) the alternating least squares methods, and (3) the multiplicative update rules. These algorithms are based on some update rules derived from different objective functions [3]. This section reviews these algorithms in details for the case of NMF.

Alg. 1 shows the projected gradient method. The $\alpha_k > 0$ denotes the step size, and it can be determined by a line search procedure. The $P[\cdot]$ implements the gradient projection onto the nonnegative surface with,

$$\text{Projection, } P[u_i] = \begin{cases} u_i, \text{ if } u_i \geq 0 \\ 0, \text{ otherwise} \end{cases} \tag{2.21}$$

Following the cost function defined by Euclidean distance in Eq. (2.18), an alternating non-negative least square algorithm for NMF can be derived (Alg. 2). It exploits the fact that, while the optimization problem of Eq. (2.18) is not convex, by fixing one factor matrix, solving the other is convex and can be done as a simple least squares computation with a non-negativity constraint.

Algorithm 1 Projected gradient method [1]

Initialize $\boldsymbol{H}^0$, $\boldsymbol{C}^0$ and iteration index $k = 0$
repeat
$\boldsymbol{H}^{k+1} \leftarrow P\left[\boldsymbol{H}^k - \alpha_k \frac{\partial \|\boldsymbol{X} - \boldsymbol{H}^k \boldsymbol{C}^k\|_F^2}{\partial \boldsymbol{H}^k}\right]$,
$\boldsymbol{C}^{k+1} \leftarrow P\left[\boldsymbol{C}^k - \alpha_k \frac{\partial \|\boldsymbol{X} - \boldsymbol{H}^k \boldsymbol{C}^k\|_F^2}{\partial \boldsymbol{C}^k}\right]$,
$k = k + 1$
until some stopping criterion

Algorithm 2 Alternating non-negative least square [2]

Initialize $\boldsymbol{H}$ and $\boldsymbol{C}$
repeat
Solve for $\boldsymbol{C}$ in matrix equation $\boldsymbol{H}^\top \boldsymbol{H}\boldsymbol{C} = \boldsymbol{H}^\top \boldsymbol{X}$.
$\boldsymbol{C} \leftarrow P[\boldsymbol{C}]$
Solve for $\boldsymbol{H}$ in matrix equation $\boldsymbol{C}\boldsymbol{C}^\top \boldsymbol{H}^\top = \boldsymbol{C}\boldsymbol{X}^\top$.
$\boldsymbol{H} \leftarrow P[\boldsymbol{H}]$
until some stopping criterion

Derived from the Euclidean distance-based cost function, a simple algorithm of NMF with multiplicative update rules was proposed by Lee and Seung (Alg. 3) [3]. Besides the Euclidean distance, the cost function of NMF can also be defined by Kullback–Leibler divergence as follows:

$$D(\boldsymbol{X}||\boldsymbol{H}\boldsymbol{C}) = \sum_{i=1}^{m}\sum_{j=1}^{n} \boldsymbol{X}_{ij} \log \frac{\boldsymbol{X}_{ij}}{(\boldsymbol{H}\boldsymbol{C})_{ij}} - \boldsymbol{X}_{ij} + (\boldsymbol{H}\boldsymbol{C})_{ij} \equiv \mathcal{L}_{\mathbf{NMF\text{-}KL}} \quad (2.22)$$

Based on Eq. (2.22), a different algorithm can be derived (Alg. 4) [3]. The advantage of the multiplicative update rule-based algorithms is that if the initial values of elements in matrices $\boldsymbol{H}$ and $\boldsymbol{C}$ are all non-negative, then the $\boldsymbol{H}$ and $\boldsymbol{C}$ can never contain negative values. However, one drawback is that once an element in $\boldsymbol{H}$ or $\boldsymbol{C}$ becomes 0 it must remain 0.

Algorithm 3 Multiplicative rules for Euclidean distance [3]

Initialize $\boldsymbol{H}^0_{n\times K}$, $\boldsymbol{C}^0_{K\times m}$ and iteration index $k = 0$
repeat
$\boldsymbol{H}^{k+1}_{i\kappa} \leftarrow \boldsymbol{H}^k_{i\kappa} \frac{(\boldsymbol{X}(\boldsymbol{C}^k)^\top)_{i\kappa}}{(\boldsymbol{H}^k\boldsymbol{C}^k(\boldsymbol{C}^k)^\top)_{i\kappa}}$,
$\boldsymbol{C}^{k+1}_{\kappa j} \leftarrow \boldsymbol{C}^k_{\kappa j} \frac{((\boldsymbol{H}^{k+1})^\top\boldsymbol{X})_{\kappa j}}{((\boldsymbol{H}^{k+1})^\top\boldsymbol{H}^{k+1}\boldsymbol{C}^k)_{\kappa j}}$,
$k = k + 1$
until some stopping criterion

Algorithm 4 Multiplicative rules for KL-divergence [3]

Initialize $\boldsymbol{H}^0_{n\times K}$, $\boldsymbol{C}^0_{K\times m}$ and iteration index $k = 0$
repeat
$\boldsymbol{C}^{k+1}_{\kappa j} \leftarrow \boldsymbol{C}^k_{\kappa j} \frac{\sum_{i=1}^n \boldsymbol{H}^k_{i\kappa}\boldsymbol{X}_{ij}/(\boldsymbol{H}^k\boldsymbol{C}^k)_{ij}}{\sum_{q=1}^n \boldsymbol{H}^k_{q\kappa}}$,
$\boldsymbol{H}^{k+1}_{i\kappa} \leftarrow \boldsymbol{H}^k_{i\kappa} \frac{\sum_{j=1}^m \boldsymbol{C}^k_{\kappa j}\boldsymbol{X}_{ij}/(\boldsymbol{H}^k\boldsymbol{C}^k)_{ij}}{\sum_{p=1}^m \boldsymbol{C}^k_{\kappa p}}$,
$k = k + 1$
until some stopping criterion

To prevent the model from overfitting the high-dimensional small datasets, various regularization terms can be applied to penalize the model complexity. For example, a sparsity constraint achieved through l_0- or l_1-norm regularizations enforces the resulting representation with some indices set to zero (see Section 2.3.2.2). This is motivated by observations in natural images or natural language documents where each image (or document) may be described as the superposition of a small number of atomic elements such as edges and corners (or topics). This section also presents the graph-based regularization that ensures the learned new representation to preserve consistent neighborhood to the original feature space (see Section 2.3.2.3). Other regularizations developed to incorporate domain knowledge or to learn more powerful models

are also introduced (see Sections 2.3.2.4 and 2.3.2.5).

2.3.2.2 Sparse coding

In order to achieve part-based additive mapping from features to data instances, the factor matrices are usually constrained to be non-negative (NMF), i.e., $\boldsymbol{H} \in \mathbb{R}_{+}^{n \times k}$ and $\boldsymbol{C} \in \mathbb{R}_{+}^{k \times m}$. In order to produce meaningful outcome feature representations, techniques of sparse coding are usually applied. Olshausen and Field proposed sparse data representation in 1996 [75]. They believe that for an observation, only a small fraction of the possible factors are relevant. This could be represented by features that are often zero or by the fact that most extracted features are insensitive to small variations of the observation. There are plenty of successful applications of sparse coding [4, 5, 76].

l_0-norm regularization: High sparsity of the extracted factor matrices $\boldsymbol{H}$ and $\boldsymbol{C}$ can be achieved through minimizing an objective function with a regularization term being l_0 [77], l_1 [78], or $l_{p,q}$ norms [79, 80]. However, it is an intractable problem to find the sparsest representation of a signal through l_0 norm $\Big(\text{e.g.,} \sum_j \mathbb{1}(\boldsymbol{c}_j \neq 0)\Big)$, because there is a combinatorial increase in the number of local minima as the number of candidate basis vectors increases.

l_1-norm regularization: More often sparsity is introduced via an l_1-norm of coefficient matrix $\boldsymbol{C}$ $\Big(\text{e.g.,} \sum_j \|\boldsymbol{c}_j\|_1\Big)$, which also results in a sparse representation and removes noise. With an l_1-norm, the sparse non-negative matrix factorization finds a basis to capture underlying semantics in the original data matrix, and learns sparse coefficients of the basis. Having an l_1-

norm in the objective function, a sparse coding multiplicative algorithm was developed [81]. In each iteration, the basis matrix $\boldsymbol{H}$ was updated using a gradient-based method and normalized by its columns, and the multiplicative algorithm was applied to update the coefficient matrix. In order to ease the process to derive the update rules through a gradient-based method, Kim and Park used squared l_1-norm for the sparsity of the coefficient matrix [2]. This work does not adopt squared l_1-norm, because true l_1-norm (not squared) has been demonstrated to be more effective to ensure sparsity. Hoyer et al. defined a sparseness function and presented an algorithm to constrain the columns of factor matrices $\boldsymbol{H}$ and/or $\boldsymbol{C}$ to a given sparse value [78]. To tackle the non-derivativeness of the l_1-norm, a *feature-sign search algorithm* was developed to selectively activate and update some elements of each data instance to iteratively reduce the objective function [4] (see Alg. 5).

Probabilistic interpretation of non-negative sparse coding:

Method 1: Sparse coding can be interpreted from a probabilistic perspective in a prior study [82]. Let $\mathbf{x}$ denote a single observation, which is a linear combination of k independent signal sources $\mathbf{h}_i$ with some additive noise ϵ,

$$\mathbf{x} = \sum_{i}^{k} c_i \mathbf{h}_i + \epsilon \tag{2.23}$$

To find the underlying basis vectors $\mathbf{h}_i$'s that best explain all observations $\boldsymbol{X}$, we can minimize the KL-divergence between the probability distri-

bution of observed signals $p^\star(\boldsymbol{X})$ and that of signals generated by the model $p(\boldsymbol{X}|\boldsymbol{H})$,

$$KL(p^\star(\boldsymbol{X})\|p(\boldsymbol{X}|\boldsymbol{H})) = \int p^\star(\boldsymbol{X})\log\left(\frac{p^\star(\boldsymbol{X})}{p(\boldsymbol{X}|\boldsymbol{H})}\right)d\boldsymbol{X} \tag{2.24}$$

which is equivalent to maximizing $p(\boldsymbol{X}|\boldsymbol{H})$ due to $p^\star(\boldsymbol{X})$ being constant,

$$p(\boldsymbol{X}|\boldsymbol{H}) = \int p(\boldsymbol{X},\boldsymbol{C}|\boldsymbol{H})d\boldsymbol{C} = \int p(\boldsymbol{X}|\boldsymbol{C},\boldsymbol{H})p(\boldsymbol{C})d\boldsymbol{C} \tag{2.25}$$

where the probability density of $p(\boldsymbol{X}|\boldsymbol{C},\boldsymbol{H})$ is a Gaussian distribution by assuming ϵ a Gaussian white noise with variance σ^2,

$$p(\boldsymbol{X}|\boldsymbol{C},\boldsymbol{H}) = \frac{1}{\sqrt{2\sigma^2\pi}}e^{-\frac{\|\boldsymbol{X}-\boldsymbol{H}\boldsymbol{C}\|_2^2}{2\sigma^2}} \tag{2.26}$$

By assuming the independence of the sources in the prior distribution $P(\boldsymbol{H})$ and parameterizing the priors for convenience, we obtain:

$$p(\boldsymbol{C}) = \prod_{i=1}^{k} p(\mathbf{c}_i) = \prod_{i=1}^{k}\frac{1}{Z}e^{-\beta S(\mathbf{c}_i)} \tag{2.27}$$

where $S(\cdot)$ is a function to characterize the prior distribution.

Since the integral over $\boldsymbol{C}$ in Eq. (2.25) is intractable, one can approximate $p(\boldsymbol{X}|\boldsymbol{H})$ by maximizing $p(\boldsymbol{X},\boldsymbol{C}|\boldsymbol{H})$ across all choices of $\boldsymbol{C}$, which is equivalent to minimizing an energy function,

$$E(\boldsymbol{X}, \boldsymbol{C}|\boldsymbol{H}) = -\log(p(\boldsymbol{X}|\boldsymbol{H}, \boldsymbol{C})p(\boldsymbol{C})) \tag{2.28}$$

$$= \sum_{j=1}^{m} \left(\|\mathbf{x}^{(j)} - \sum_{i=1}^{k} c_i^{(j)} \mathbf{h}_i\|^2 + \lambda \sum_{i=1}^{k} S(c_i^{(j)}) \right) \tag{2.29}$$

where the l_1-norm can be achieved by selecting a Laplacian prior in Eq. (2.27).

Method 2 [83]: In MF settings, we can associate different models with different k (The optimal value of k can be chosen through cross-validation). Assuming a model $\mathcal{M}_k$ with complexity k, the posterior distribution of parameters times the evidence which the data provide for model $\mathcal{M}_k$ equals the likelihood multiplied by the prior according to Bayes' rule:

$$p(\boldsymbol{H}, \boldsymbol{C}|\boldsymbol{X}, k)p(\boldsymbol{X}|k) = p(\boldsymbol{X}|\boldsymbol{H}, \boldsymbol{C})p(\boldsymbol{H}, \boldsymbol{C}|k) \tag{2.30}$$

The most probable set of parameters $\boldsymbol{H}$, $\boldsymbol{C}$, given a fixed number k of latent variables, can be estimated by maximizing the posterior:

$$p(\boldsymbol{H}, \boldsymbol{C}|\boldsymbol{X}, k) = \frac{p(\boldsymbol{X}|\boldsymbol{H}, \boldsymbol{C})p(\boldsymbol{H}, \boldsymbol{C}|k)}{p(\boldsymbol{X}|k)} \tag{2.31}$$

w.r.t. the parameters $\boldsymbol{H}$ and $\boldsymbol{C}$.

Assuming $\boldsymbol{H}$ and $\boldsymbol{C}$ are independent (i.e., $p(\boldsymbol{H}, \boldsymbol{C}|k) = p(\boldsymbol{H}|k)p(\boldsymbol{C}|k)$) and taking the logarithm on both sides of Eq. (2.31), we have:

$$\ln p(\boldsymbol{H}, \boldsymbol{C}|\boldsymbol{X}, k) = \ln p(\boldsymbol{X}|\boldsymbol{H}, \boldsymbol{C}) + \ln p(\boldsymbol{H}|k) + \ln p(\boldsymbol{C}|k) - \ln p(\boldsymbol{X}|k) \tag{2.32}$$

Continued with Eq. (2.32), now we add prior knowledge by choosing, e.g., exponential priors of the form:

$$p(\boldsymbol{C}|k; \lambda) = \prod_i \prod_\kappa \lambda e^{-\lambda \boldsymbol{C}_{i\kappa}} \tag{2.33}$$

with $\lambda > 0$, leads to:

$$\ln p(\boldsymbol{C}|k;\lambda) = -\lambda \sum_i \sum_\kappa \boldsymbol{C}_{i\kappa} \tag{2.34}$$

which constitutes the additional penalty terms used in objective:

$$\min_{\boldsymbol{H},\boldsymbol{C}} \|\boldsymbol{X} - \boldsymbol{H}\boldsymbol{C}\|_F^2 + \lambda \|\boldsymbol{C}\|_1 \tag{2.35}$$

to enforce sparse coding of the coefficient matrix $\boldsymbol{C}$.

Maximum a posteriori (MAP) estimation: Thus sparse coding can be interpreted as MAP estimation by maximizing Eq. (2.32) w.r.t. $\boldsymbol{H}$ and $\boldsymbol{C}$ and assuming independent exponential priors of the weights $\boldsymbol{C}_{i\kappa}$ and Gaussian likelihood function in Eq. (2.9).

Alg. 5 presents the feature-sign search algorithm to learn the new representation for each data instance $\mathbf{x}_j$. For simplicity, the data instance $\mathbf{x}_j$ and its new representation $\mathbf{c}_j$ are denoted $\mathbf{x}$ and $\mathbf{c}$ respectively. The dimensions in $\mathbf{c}$ are indexed by i. Alg. 5 optimizes,

$$\min_{\mathbf{c}>0} f(\mathbf{c}) = \|\mathbf{x} - \boldsymbol{H}\mathbf{c}\|^2 + \lambda \|\mathbf{c}\|_1 \tag{2.36}$$

where step 2 is to select the dimensions of $\mathbf{c}$ with the worst approximation to $\mathbf{x}$ for update. The solution in step 3 can be obtained either by setting the derivative of objective equal to zero and solve for $\mathbf{c}$ or by unconstrained quadratic programming (QP).

Algorithm 5 Feature-sign search algorithm (for each data instance $\mathbf{x}_j$) [4]

1. Initialize $\mathbf{c} = \mathbf{0}$, $\theta = \mathbf{0}$, and active set $A = \emptyset$, where $\theta_i \in \{-1, 0, 1\}$ denotes $\text{sign}(\mathbf{c}_i)$.
2. From zero coefficients of $\mathbf{c}$, select $i = \arg\max_i \left| \frac{\partial \|\mathbf{x} - \boldsymbol{H}\mathbf{c}\|^2}{\partial \mathbf{c}_i} \right|$. Activate $\mathbf{c}_i$ by adding i to set A only if it improves the objective, namely:
   If $\frac{\partial \|\mathbf{x} - \boldsymbol{H}\mathbf{c}\|^2}{\partial \mathbf{c}_i} > \lambda$, then set $\theta_i = -1$, $A = A \cup \{i\}$.
   If $\frac{\partial \|\mathbf{x} - \boldsymbol{H}\mathbf{c}\|^2}{\partial \mathbf{c}_i} < -\lambda$, then set $\theta_i = 1$, $A = A \cup \{i\}$.
3. Feature-sign step:
   Let $\hat{\boldsymbol{H}}$ denote the submatrix of $\boldsymbol{H}$ with only columns regarding A.
   Let $\hat{\mathbf{c}}$, $\hat{\theta}$ be the subvectors of $\mathbf{c}$, θ with only dimensions regarding A.
   The solution to $\min_{\hat{\mathbf{c}}} \|\mathbf{x} - \hat{\boldsymbol{H}}\hat{\mathbf{c}}\|^2 + \lambda \hat{\theta}^\top \hat{\mathbf{c}}$ can be derived as $\hat{\mathbf{c}}_{new} = (\hat{\boldsymbol{H}}^\top \hat{\boldsymbol{H}})^{-1}(\hat{\boldsymbol{H}}^\top \mathbf{x} - \lambda(\hat{\theta})/2)$.
   Perform a line search on the segment from $\hat{\mathbf{c}}$ to $\hat{\mathbf{c}}_{new}$ to update $\hat{\mathbf{c}}$ and A.
      Check the objective at all points for sign changes.
      Update $\hat{\mathbf{c}}$ (and entries in $\mathbf{c}$) to the point with the lowest objective.
   Remove zero coefficients of $\hat{\mathbf{c}}$ from A and update $\theta = \text{sign}(\mathbf{c})$.
4. Check the optimality conditions:
   (a) $\frac{\partial \|\mathbf{x} - \boldsymbol{H}\mathbf{c}\|^2}{\partial \mathbf{c}_j} + \lambda \text{ sign}(\mathbf{c}_j) = 0$, $\forall \mathbf{c}_j \neq 0$. If satisfied, $\Rightarrow$ (b); or, $\Rightarrow$ 3.
   (b) $\left| \frac{\partial \|\mathbf{x} - \boldsymbol{H}\mathbf{c}\|^2}{\partial \mathbf{c}_j} \right| \leq \lambda$, $\forall \mathbf{c}_j = 0$. If satisfied, return $\mathbf{c}$; or, $\Rightarrow$ 2.

The feature-sign search algorithm maintains an active set to only update the dimensions of $\mathbf{c}$ that approximate $\mathbf{x}$ poorly, which makes the algorithm efficient. The proof of convergence can be found in the original paper [4].

The sparsity constraint is in line with the assumption made in our studies that each data instance is related to only a few hidden topics in experts' mind, and the low rank approximation (of NMF) and sparse representation altogether save a great deal of storage. Our studies built on this prior work's idea of sparse coding and devised a strategy to discover a set of latent con-

cepts. The desired latent concepts capture high-level medical concepts of the spoken narratives that can be used to recover the original term vectors of the narratives. Since a narrative is used to describe a specific medical image, it is common for the narrative to concentrate on a small number of medical concepts. Stated differently, a spoken narrative is expected to be only related to a small subset of latent concepts. Therefore, a desired latent concept set can be identified through NMF with sparsity constraint. See 3.3 for more details.

2.3.2.3 Graph-regularized NMF

Based on the *feature-sign search algorithm*, Zheng et al. developed the graph-regularized non-negative matrix factorization (GrNMF), which adds a graph-regularizer to sparse NMF [5]. The nonnegativity, sparsity, and a graph regularizer altogether form the Laplacian sparse coding [84].

$$\begin{aligned} \min_{\boldsymbol{H},\boldsymbol{C}\geq 0} \|\boldsymbol{X}-\boldsymbol{H}\boldsymbol{C}\|_F^2 + \alpha Tr(\boldsymbol{C}\boldsymbol{L}\boldsymbol{C}^\top) + \beta \sum_{j=1}^{m} \|\mathbf{c}_j\|_1 \\ s.t. \|\mathbf{h}_i\|^2 \leq a, i = 1, ..., k \end{aligned} \tag{2.37}$$

where $\boldsymbol{X} \in \mathbb{R}^{n\times m}$ represents the m data instances in original n feature space. The $\mathbf{h}_i, i = 1, ..., k$ denotes the basis vectors. The norm constraints on the size of the basis vectors (i.e., $\|\mathbf{h}_i\|^2 \leq a$) avoid arbitrarily large basis vectors that keep $\boldsymbol{H}\boldsymbol{C}$ unchanged while making $\mathbf{c}_j$ arbitrarily close to zero. The $\|\cdot\|$ is the vector l_2-norm and a is a positive constant number. The basis matrix can be updated using Lagrange dual and the coefficient matrix can be updated through the feature-sign search algorithm.

The graph-regularizer (i.e., $Tr(\boldsymbol{CLC}^\top)$) was introduced by a weighted graph of the data points represented in the input matrix. Let those data points be denoted as $\mathbf{x}_1$, ..., $\mathbf{x}_m$. A nearest neighbor graph G with m vertices can be constructed. The element $\boldsymbol{W}_{ij}$ in the neighboring matrix $\boldsymbol{W}$ of the graph G can be computed using a heat kernel [85].

$$\boldsymbol{W}_{ij} = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|}{\sigma}} \tag{2.38}$$

If $\mathbf{x}_i$ and $\mathbf{x}_j$ are identical, then W_{ij} equals 1; and if they are extremely different, then $\boldsymbol{W}_{ij}$ asymptotically approaches 0. The degree of $\mathbf{x}_i$ is defined as $d_i = \sum_{j=1}^{m} \boldsymbol{W}_{ij}$, and $\boldsymbol{D} = \text{diag}(d_1, ..., d_m)$. $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$, is a Laplacian matrix used to minimize the Laplacian item $Tr(\boldsymbol{CLC}^\top)$ in the objective function (Eq. (2.37)). The α and β are the weights of the regularizers, and they balance between the two types of regularization. The optima of (α, β) can be chosen through parameter tuning.

Similarly, other graph-weighting strategies can be adopted, such as *0-1 weighting* in Eq. (2.39), *histogram intersection kernel weighting* in Eq. (2.40), and *dot-product weighting* in Eq. (2.41), depending on the feature attributes. They can be used together to achieve superior clustering result [86].

$$\boldsymbol{W}_{ij} = \begin{cases} 1, & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are close} \\ 0, & \text{otherwise} \end{cases} \tag{2.39}$$

$$\boldsymbol{W}_{ij} = \begin{cases} \sum_{d=1}^{D} \min(\mathbf{x}_{di}, \mathbf{x}_{dj}), & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are close} \\ & (d \text{ represents each feature dimension}) \\ 0, & \text{otherwise} \end{cases} \tag{2.40}$$

$$\boldsymbol{W}_{ij} = \mathbf{x}_i^\top \mathbf{x}_j \tag{2.41}$$

Zheng et al. developed a graph-regularized sparse coding based on the feature-sign search algorithm in Alg. 5 [5]. Their modified feature-sign search algorithm is shown in Alg. 6, where the reconstruction error and the graph regularizer are substituted by functions for simplicity:

$$\begin{aligned} \mathcal{R}(\mathbf{c}_i) &= \|\mathbf{x}_i - \boldsymbol{H}\mathbf{c}_i\|^2 \\ \mathcal{G}(\mathbf{c}_i) &= \alpha \sum_{i=1}^{m} \sum_{j=1}^{m} \boldsymbol{L}_{ij} \mathbf{c}_i^\top \mathbf{c}_j \end{aligned} \tag{2.42}$$

where subscript i in $\mathbf{x}_i$ and $\mathbf{c}_i$ indexes the data points (not the dimensions). The θ_j (sign in the j-th dimension of $\mathbf{c}_i$) is omitted in the algorithm, because in the non-negative matrix factorization setting, the elements in $\mathbf{c}_i$ are never negative. The non-negativity is achieved through a projection step, where all negative elements in $\mathbf{c}_i$ are assigned value zero.

Algorithm 6 A modified feature-sign search algorithm for GrNMF [5]

1. Initialize $\mathbf{c}_i = \mathbf{0}$, and active set $A = \emptyset$.
2. From zero coefficients of $\mathbf{c}_i$, select $j = \arg\max_j \left| \frac{\partial\left(\mathcal{R}^{(j)}(\mathbf{c}_i)+\mathcal{G}^{(j)}(\mathbf{c}_i)\right)}{\partial \mathbf{c}_i^{(j)}} \right|$. Activate $\mathbf{c}_i^{(j)}$ by adding j to set A only if it improves the objective, namely:
   If $\frac{\partial\left(\mathcal{R}^{(j)}(\mathbf{c}_i)+\mathcal{G}^{(j)}(\mathbf{c}_i)\right)}{\partial \mathbf{c}_i^{(j)}} < -\beta$, then $A = A \cup \{j\}$.
3. Feature-sign step:
   Let $\hat{\boldsymbol{H}}$ denote the submatrix of $\boldsymbol{H}$ with only columns regarding A.
   Let $\hat{\mathbf{c}}_i$ denote the subvector of $\mathbf{c}_i$ with only dimensions regarding A.
   The solution to $\min_{\hat{\mathbf{c}}_i} \mathcal{R}(\hat{\mathbf{c}}_i) + \mathcal{G}(\hat{\mathbf{c}}_i) + \beta\hat{\mathbf{c}}_i$ can be derived as $\hat{\mathbf{c}}_i^{new} = (\hat{\boldsymbol{H}}^\top \hat{\boldsymbol{H}} + \alpha \boldsymbol{L}_{ii}\boldsymbol{I})^{-1}(\hat{\boldsymbol{H}}^\top \mathbf{x}_i - \alpha \sum_{j\neq i} \boldsymbol{L}_{ij}\hat{\mathbf{c}}_j - \beta/2)$.
   Perform a line search on the segment from $\hat{\mathbf{c}}_i$ to $\hat{\mathbf{c}}_i^{new}$ to update $\hat{\mathbf{c}}_i$, A.
4. Check the optimality conditions:
   (a) $\frac{\partial\left(\mathcal{R}^{(j)}(\mathbf{c}_i)+\mathcal{G}^{(j)}(\mathbf{c}_i)\right)}{\partial \mathbf{c}_i^{(j)}} + \beta = 0$, $\forall \mathbf{c}_i^{(j)} \neq 0$. If satisfied, $\Rightarrow$ (b); or, $\Rightarrow$ 3.
   (b) $\left| \frac{\partial\left(\mathcal{R}^{(j)}(\mathbf{c}_i)+\mathcal{G}^{(j)}(\mathbf{c}_i)\right)}{\partial \mathbf{c}_i^{(j)}} \right| \leq \beta$, $\forall \mathbf{c}_i^{(j)} = 0$. If satisfied, return $\mathbf{c}_i$; or, $\Rightarrow$ 2.

The modified feature-sign search algorithm serves as a backbone algorithm in various chapters of this dissertation, including verbal narrative clustering approach in Section 3.3, multimodal data fusion framework in Section 4.4, and interactive image grouping paradigm in Section 5.2. This also shows the flexibility of NMF-based models to be generalized for multimodal data fusion and to be modified for incorporating expert knowledge.

We developed a multimodal variant of GrNMF to fuse multimodal human data. The constraints used are the non-negativity constraint, sparsity constraint, and similarity-graph-based constraint [87]. Our multimodal GrNMF was implemented by modifying the GrNMF into a multimodal version

(see Section 4.4). In particular, we derived a multimodal feature-sign search algorithm to update the coefficient matrix $\boldsymbol{C}$ [87]. The learned unified data representation achieved better performance than each original single-modal representation. It also outperformed many other linear representation learning techniques. We also developed an expert-driven interface for interactive learning of multimodal image organization. This interface was developed using the same constraints mentioned above and it has a few extensions to update the model based on user interactions. One of the update rules is to modify the neighboring matrix using Eq. (2.38), and this update rule can be found in Section 5.2.5.

Neighboring graph-based approach, however, has an unavoidable limitation. Basing the modeling of manifolds on training set neighborhood relationships is risky statistically in high-dimensional spaces (sparsely populated due to the curse of dimensionality) as, for example, most Euclidean nearest neighbors risk having too little in common semantically. The nearest neighbor graph is simply not sufficiently densely populated to map out satisfyingly the wrinkles of the target manifold. It can also become problematic computationally to consider all pairs of data points, which scales quadratically with training set size.

2.3.2.4 Medical knowledge-regularized NMF

GrNMF uses the data structure in the original space as a constraint. This approach takes advantage of the global similarity from the data points them-

selves, but does not use any domain-related knowledge. However, per our need to learn a data representation for expert understanding of medical images, other constraints need to be derived from the domain knowledge base, so that more prior knowledge (than the observations of term frequencies) can be used as well. For example, Ji et al. explored semantic relatedness between domain-related terms, and developed a constraint to ensure that the documents sharing very related terms are close to each other in the new representation [88]. Let $\boldsymbol{T} \in \mathbb{R}^{k \times k}$ be the term-term correlation matrix in which $\boldsymbol{T}_{pq}$ measures the correlation between the p-th and the q-th term. They solved the following optimization problem:

$$\min_{\boldsymbol{H},\boldsymbol{C} \geq 0} \|\boldsymbol{X} - \boldsymbol{H}\boldsymbol{C}\|_F^2 + \lambda_1 \|\boldsymbol{H}\|_F^2 + \lambda_2 \sum_{(p,q) \in G} g(\boldsymbol{T}_{pq}) \|\mathbf{h}_p - \text{sign}(\boldsymbol{T}_{pq})\mathbf{h}_q\|^2 \tag{2.43}$$

where G denotes the index set with the magnitudes of the corresponding entries in $\boldsymbol{T}$ above a pre-specified threshold μ, i.e., $G = \{(p,q) : |\boldsymbol{T}_{pq}| \geq \mu\}$, and $g(\boldsymbol{T}_{pq})$ is the weight for the regularization on terms p and q. The $\mathbf{h}_p$ and $\mathbf{h}_q$ are terms p and q represented in new space. Highly-similar terms are constrained to have similar representations, and highly-dissimilar terms are constrained to have dissimilar representations. In their work, $g(\boldsymbol{T}_{pq}) = \boldsymbol{T}_{pq}^2$. The ontological information they used to develop the constraint is based on a term-term correlation coefficient, which is similar to the semantic relatedness illustrated in Figure 2.1 using the data in this dissertation.

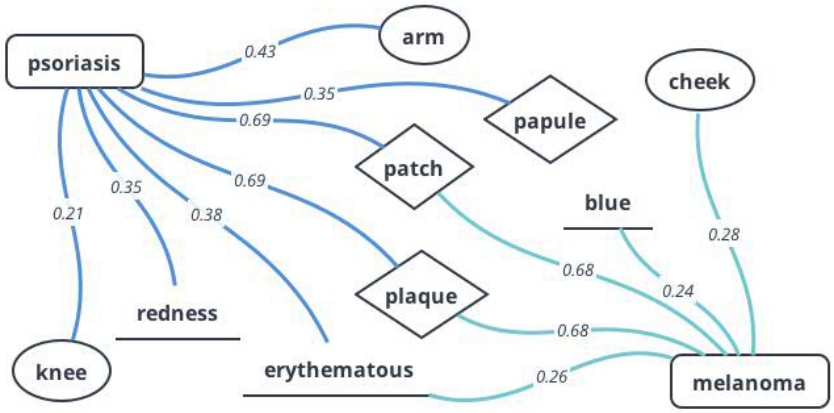


Figure 2.1: The term-term interaction graph computed using the semantic relatedness score in UMLS (details in Section 3.4.2). The vertices represent terms and edges the relatedness scores between terms. Disease names are in rounded rectangles, primary lesion types in diamonds, body locations in ellipses, and lesion colors in underlines.

2.3.2.5 Regularization for efficient non-linearity

The research problem in this dissertation is to learn expert domain knowledge from expert-derived data. This involves levels of abstractions, and hence it may require a multi-layer representation to explain the variances at different semantic levels. The deep learning models such as stacked autoencoder [89] and deep RBMs [90] have achieved success in various application domains. For example, deep learning outperforms the state-of-the-art models to classify images [60, 91] or to classify the digits in the MNIST dataset [92, 93, 94]. However, the advances in deep learning models cannot be directly used to solve the problem in this dissertation: Natural scene images can rely on object detectors to recognize important semantics in the images, whereas the image content in medical images are not clear-cut objects and there are no medical

object detectors that perform well without physician supervision. The key features to distinguish between digits $0-9$ in the MNIST dataset are the image pixels and their spatial configurations, whereas the key features in the medical images reside in domain experts' interpretations. Such differences between the research problems inform that the deep learning models cannot be directly applied to the research problem in this dissertation. In additions, given the small datasets for this dissertation, the number of hidden layers needs to be small (in case of overfitting), in comparison with the success of deep convolutional neural networks (CNNs) to classify dermatology images with a large dataset [95].

To allow arbitrary layers of non-linear transformations depending on the data size, there is a variant of sparse coding, namely *Predictive Sparse Decomposition (PSD)* [96]. It contains an extra regularizer to enforce the sparse codes to be nearly computable by a smooth and efficient encoder. The objective function is shown in Eq. (2.44). Once PSD is trained, the resulting representations used to feed a classifier are computed from $f(\cdot)$, which is fast and can then be optimized along with following stages of a deep architecture.

$$\min_{\boldsymbol{H},\boldsymbol{C}\geq 0} \|\boldsymbol{X}-\boldsymbol{H}\boldsymbol{C}\|_F^2 + \sum_{j=1}^{m}(\lambda\|\mathbf{c}_j\|_1 + \|\mathbf{c}_j - f(\mathbf{x}_j)\|_F^2) \tag{2.44}$$

This approach can be used to smooth the transformation obtained through sparse coding—the sparse coding criterion does not guarantee the smoothness of encoding process (i.e., minor changes of the input data $\boldsymbol{X}$ may

result in dramatic changes of its new representation).[1]

2.3.2.6 Summary and Discussions on Other Models

As shown above in this chapter, various assumptions on the data generation can be applied to the matrix factorization-based framework in order to arrive at an intended feature representation [97, 81]. Moreover, other constraints for the MF-based framework can be formulated through user interactions to obtain semantically meaningful features (see Chapter 5).

Below are some widely-used models for representation learning, some of which are related to NMF and/or can be explained using the matrix factorization framework. These models include probabilistic latent semantic indexing/analysis (pLSI or pLSA), latent Dirichlet allocation (LDA), and probabilistic matrix factorization (PMF). We will review some basics and their relationships with NMF and explain the reasons we adopt and improve NMF-based models instead of them in this dissertation.

Probabilistic latent semantic indexing/analysis, also known as *aspect model*, is a method arising in natural language processing to model co-occurrence data (i.e., documents in linguistic corpora). Similar to the use case of NMF in language data clustering, pLSA models each word in a document as a sample from a mixture model, where mixture components are multinomial

[1]The codes obtained by complete optimization of sparse coding can be highly non-smooth or even non-differentiable. Even the graph-based regularizer described in Section 2.3.2.3 does not guarantee the smoothness of the encoding process, because neighboring graph only ensures the local neighborhood between observed data points.

random variables that can be viewed as representations of topics.

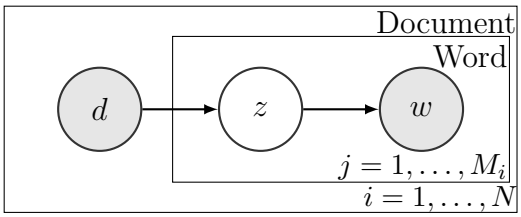


Figure 2.2: Probabilistic latent semantic analysis.

Following the notations for NMF (i.e., document d, and word w), the d and w are conditionally independent given an unobserved topic z:

$$p(d, w) = p(d) \sum_z p(w|z)p(z|d) \tag{2.45}$$

$$= \sum_z p(z)p(w|z)p(d|z) \tag{2.46}$$

The pLSA maximizes the likelihood, $\mathcal{L}_{pLSA}$:

$$\mathcal{L}_{pLSA} = \sum_i^m \sum_j^n \boldsymbol{F}_{ij} \log p(d_i, w_j) \tag{2.47}$$

where the matrix $\boldsymbol{F}$ is the word-frequency matrix or tf-idf matrix. According to [98, 99], this is equivalent to minimizing the objective function of NMF with KL-divergence as shown in Eq. (2.22). The underlying models are essentially the same, and the only difference between pLSA and NMF is how inference proceeds.

Latent Dirichlet allocation is a generative probabilistic model of a corpus. It represents documents as random mixtures over latent topics, where each topic is characterized by a distribution over words.

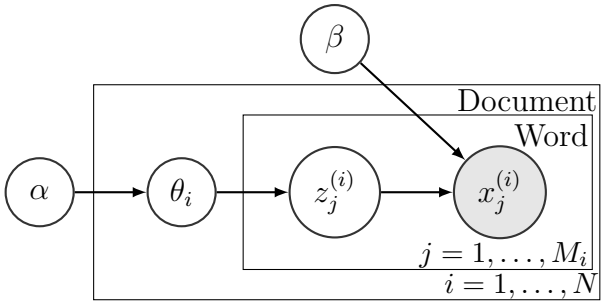


Figure 2.3: Latent Dirichlet allocation.

For each document i in a corpus, $\theta_i \sim \text{Dir}(\alpha)$. For each of the M_i words j: choose a topic $z_j^{(i)} \sim \text{Mult}(\theta_i)$, and choose a word $x_j^{(i)}$ from $p(x_j^{(i)}|z_j^{(i)}, \beta)$.

Blei et al. pointed out that the pLSI (and hence its equivalent counterpart, NMF) learns the topic mixtures only for the documents in the training set, and there is no natural way to use it to assign probabilities to a previously unseen document [100]. However, in our use case we have a fixed number of diagnostic verbal narratives, due to the high cost of recruiting medical experts for experiments. This allows the NMF to work for our scenario, where topic probabilities remain fixed per narrative.

Additionally, we currently have relatively small datasets (only tens of images and hundreds of narratives). LDA does not have enough information for the likelihood $p(\text{data}|\text{params})$, which keeps the learned posterior $p(\text{params}|\text{data})$ around the prior $p(\text{params})$. In this case, the additional variability coming from the hyperpriors is too much. In this sense NMF better suits our particular research problem and datasets, although it is prone to *overfitting*.

Besides, since LDA works naturally on word counts, it largely relies on stopword removal of the high frequency (but non-topic-specific) terms that may appear in many learned topics [101]. On the contrary, the NMF typically use tf-idf weighting scheme to solve this issue more elegantly—less ad-hoc than constructing a corpus-specific stoplist [102].

Despite that the NMF with KL-divergence approximates the LDA model under a uniform Dirichlet prior [103], existing studies that use topic modeling approaches have different preferences for LDA vs. NMF. For example, Choo et al. claims that there is inconsistency of topics learned with LDA between successive iterations in a single run, and hence prefers NMF for interactive topic modeling [104]. However, existing studies incorporate domain knowledge into topic modeling via Dirichlet forest priors, in spite of its complexity [105]. Contrary to Choo et al., Stevens et al. mentions the incoherent topics learned with NMF across multiple runs [106], which suggests LDA for interactive learning purposes. Although the uses of NMF vs. LDA for interactive learning purposes remain debatable (i.e., in terms of topic coherence [106], and topic consistency [104]), our extension with expert inputs as additional constraints in Chapter 5 uses NMF-based framework for its flexibility to receive expert inputs as additional constraints and its support for global and high-level tasks (i.e., to allow expert users working with images, instead of terms).

Probabilistic Matrix Factorization is a probabilistic model with Gaussian observation noise. Given input data $\boldsymbol{R}$, it seeks a low rank approxi-

mation with two factors $\boldsymbol{U}$ and $\boldsymbol{V}$ [97] as:

$$p(\boldsymbol{R} \mid \boldsymbol{U}, \boldsymbol{V}, \sigma^2) = \prod_{i=1}^{N} \prod_{j=1}^{M} \mathcal{N}(\boldsymbol{R}_{ij} \mid \boldsymbol{U}_i^\top \boldsymbol{V}_j, \sigma^2) \tag{2.48}$$

where $\mathcal{N}(x \mid \mu, \sigma^2)$ is the probability density function of the Gaussian distribution with mean μ and variance σ^2. Through adding zero-mean Gaussian priors on factors $\boldsymbol{U}$ and $\boldsymbol{V}$, Eq. (2.48) enables maximizing the log-posterior as:

$$\ln p(\boldsymbol{U}, \boldsymbol{V} \mid \boldsymbol{R}, \sigma^2, \sigma_{\boldsymbol{U}}^2, \sigma_{\boldsymbol{V}}^2) = \ln p(\boldsymbol{R} \mid \boldsymbol{U}, \boldsymbol{V}, \sigma^2) + \ln p(\boldsymbol{U} \mid \sigma_{\boldsymbol{U}}^2) + \ln p(\boldsymbol{V} \mid \sigma_{\boldsymbol{V}}^2) + C \tag{2.49}$$

where $\sigma_{\boldsymbol{U}}^2$ and $\sigma_{\boldsymbol{V}}^2$ are variances of the priors on $\boldsymbol{U}$ and $\boldsymbol{V}$. C is a constant that does not depend on the parameters.

$$E = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{M} I_{ij} \left(\boldsymbol{R}_{ij} - \boldsymbol{U}_i^\top \boldsymbol{V}_j \right)^2 + \frac{\lambda_{\boldsymbol{U}}}{2} \sum_{i=1}^{N} \|U_i\|_F^2 + \frac{\lambda_{\boldsymbol{V}}}{2} \sum_{j=1}^{M} \|\boldsymbol{V}_j\|_F^2 \tag{2.50}$$

where $\lambda_{\boldsymbol{U}} = \sigma^2 / \sigma_{\boldsymbol{U}}^2$, $\lambda_{\boldsymbol{V}} = \sigma^2 / \sigma_{\boldsymbol{V}}^2$, and $\| \cdot \|_F^2$ denotes the Frobenius norm.

The PMF is mostly used to build a recommender system by only constraining on the sparse observed data indicated by I_{ij} and regarding other entries as *missing values*. However, it can also explain a full observed data matrix by simply setting all I_{ij}'s with value 1.

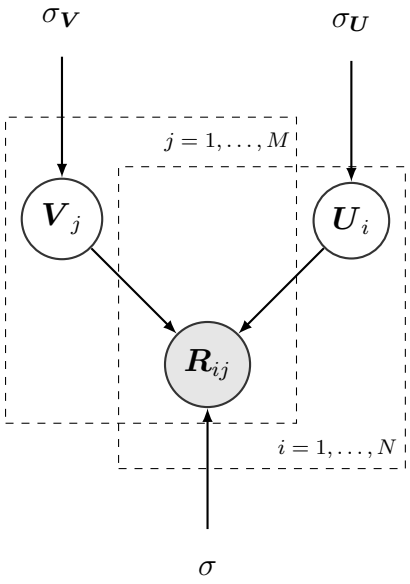


Figure 2.4: Graphical illustration of PMF.

2.3.3 Technical connections between representation learning models

The relationships between representation learning models are shown in Figure 2.5. The NMF derived from KL-divergence is equivalent to the pLSA, and it is related to ICA under the interpretation of a Gamma-Poisson (GaP) model. There is a debate in terms of using LDA or NMF for interactive machine learning purposes. This dissertation prefers NMF for two reasons—(1) the intended interactions are at high level (experts' image manipulations), and (2) due to limited number of images available, in this dissertation the inter-

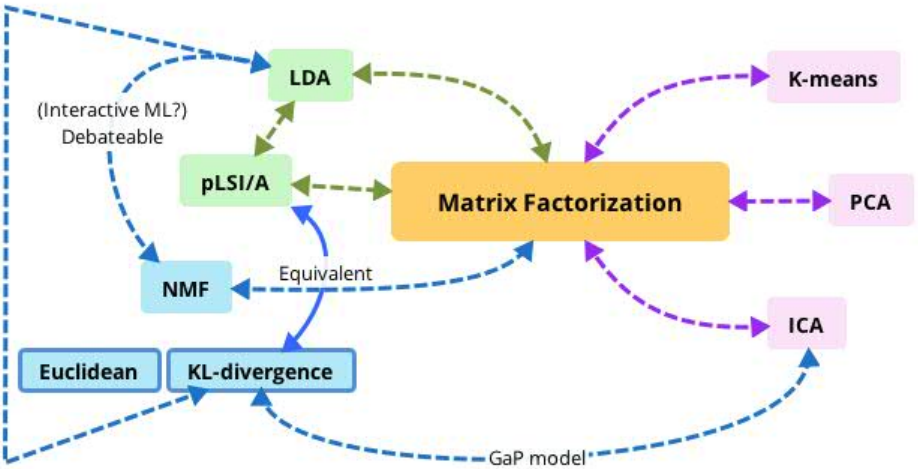


Figure 2.5: Relationships between representation learning models in the framework of matrix factorization.

actions do not involve new images.[2] All these representation learning models can be interpreted by the framework of matrix factorization.

2.3.4 Sequential Models

Despite that the Bag-of-Word (BoW) models (in Section 2.3) discover high-level representations (i.e., topic mixtures) using global information, they do not consider word order. The transcribed verbal narratives are essentially sequences of discrete observations (medical terms), so this dissertation also applies sequential models. For example, hidden Markov models (HMMs) are used to model documents and speech data whose word order contains key

[2]In case that the database will be scaled up, the future work includes developing variants for the LDA model.

information [107].

2.3.4.1 Hidden Markov model

Given an observation sequence $\mathbf{x} = \{x_1, x_2, ..., x_T\}$, the model learns initial state probabilities G, state transition probabilities $\boldsymbol{\pi}$, observation symbol (emission) probabilities $\boldsymbol{\phi}$, and a corresponding state sequence $\mathbf{z} = \{z_1, z_2, ..., z_T\}$, where each time step t chooses a state index from $\{1, 2, ..., K\}$. Below is how the model generates an output (observation):

1. The state probability distribution G is used to choose an initial state, where $\sum_i^K G_i = 1$.
2. Starting from the initial state, the following states are selected one by one per transition probabilities $\boldsymbol{\pi}$, which is a stochastic matrix $\sum_{j=1}^K \pi_{ij} = 1$, for $\forall i$.
3. Depending on our problem domain to model medical language data, we use the discrete HMMs whose observations are chosen from a finite countable set $V = \{v_1, v_2, ..., v_{|V|}\}$, s.t., $x_t \in V$, $t = 1, 2, ..., T$.

This process shows the *Markov assumption*, i.e., the distribution of the current state z_t depends only on the previous state z_{t-1}, and the *conditional independence*, i.e., the observation x_t is independent of all other observations and hidden states, conditional on the current hidden state z_t.

The *forward-backward algorithm* provides an *exact* solution to finding the posterior marginals of all hidden state variables given a sequence of observations (emissions), $p(z_t|\mathbf{x}, \boldsymbol{\pi}, \boldsymbol{\phi}, G)$ for $t = 1, 2, \ldots, T$. In the algorithm, both

the forward and backward passes are based on dynamic programming, which makes the algorithm computationally efficient.

Let the parameters of a hidden Markov model be denoted as $\theta = (\boldsymbol{\pi}, \boldsymbol{\phi}, G)$. To evaluate the probability of $p(\mathbf{x}|\theta)$ and find the most probable state at time step t, the forward-backward algorithm can be used.

Algorithm 7 The forward algorithm

1. Define $\alpha_t(i) = p(x_1, x_2, \cdots, x_t, z_t = i|\theta)$.
2. Initialize $\alpha_1(i) = G_i F(\phi_i(x_1))$, $1 \leq i \leq K$.
3. For $t = 2, \cdots, T$, derive:
   $\alpha_t(j) = \sum_{i=1}^{K} \alpha_{t-1}(i)\pi_{ij} F(\phi_j(x_t))$, $1 \leq j \leq K$.

In this manner, the $p(\mathbf{x}|\theta) = \sum_i^N \alpha_T(i)$ covers all possible states i's for each time t and hence is irrelevant to $\mathbf{z}$.

Algorithm 8 The backward algorithm

1. Define $\beta_t(i) = p(x_{t+1}, \cdots, x_T|z_t = i, \theta)$.
2. Initialize $\beta_T(i) = 1$, $1 \leq i \leq K$.
3. For $t = T-1, \cdots, 1$, derive:
   $\beta_t(i) = \sum_{j=1}^{K} \pi_{ij} F(\phi_j(x_{t+1}))\beta_{t+1}(j)$, $1 \leq i \leq K$.

Algorithm 9 The forward-backward algorithm

1. Define $\gamma_t(i) = p(z_t = i|\mathbf{x}, \theta)$.
2. Derive $\gamma_t(i) = p(z_t = 1|x_1, \cdots, x_T, \theta) = \frac{p(z_t=i, x_1, \cdots, x_t, x_{t+1}, \cdots, x_T|\theta)}{p(x_1, \cdots, x_T|\theta)}$
   $= \frac{p(z_t=i, x_1, \cdots, x_t|\theta)p(x_{t+1}, \cdots, x_T|z_t=1, \theta)}{p(x_1, \cdots, x_T|\theta)}$
   $= \frac{\alpha_t(i)\beta_t(i)}{p(\mathbf{x}|\theta)}$
   $= \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{K} \alpha_t(i)\beta_t(i)}$

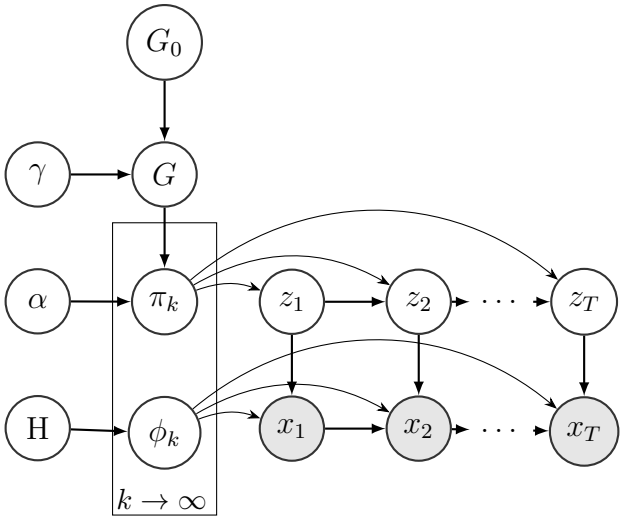


Figure 2.6: The hierarchical Dirichlet process-hidden Markov model that learns from a single observation sequence $\{x_t\}_{t=1,2,\cdots,T}$.

To find the most probable state sequence, the *Viterbi algorithm* can be used, which essentially computes $\delta_t(j) = \max_{z_1,\cdots,z_{t-1}} p(x_1, \cdots, x_t, z_1, \cdots, z_{t-1}, z_t = j|\theta)$ through forward induction and backtracking of the state sequence. Both the forward-backward and Viterbi algorithms are derived through recursive probability computation and implemented based on dynamic programming.

2.3.4.2 Infinite hidden Markov model

HMMs involve three learning tasks—inferencing the hidden state sequence $\mathbf{z}$, learning the model parameters $\boldsymbol{\pi}$, $\boldsymbol{\phi}$ and G, and selecting the optimal number of hidden states. The HDP-HMM (a.k.a, infinite HMM, or iHMM) conducts them all (especially the third above, as opposed to the canonical HMM) automatically.

In an infinite HMM, G denotes the global measure, and it is distributed according to a Dirichlet process $DP(\gamma, G_0)$ with G_0 as the base measure and γ the concentration parameter. Each π_k is a transition probability distribution from a state indexed by k, and the π_k's are conditionally independent given G. This hierarchical construction can be formulated as follow:

$$G \mid G_0 \sim DP(\gamma, G_0) \tag{2.51}$$

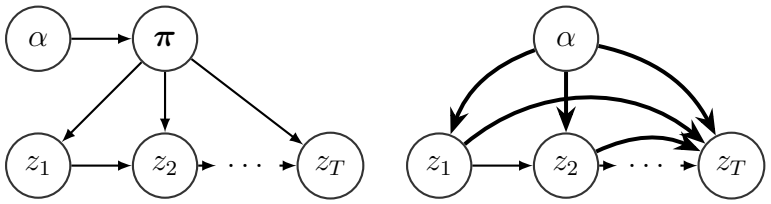
$$\begin{aligned} \pi_k \mid G &\sim DP(\alpha, G) \\ k &= 1, 2, \ldots, \infty \end{aligned} \tag{2.52}$$

In the sequence, each transition probability distribution $\{\pi_{z_{t-1}, z_t=k}\}_{k=1,2,\cdots,\infty}$ of the hidden Markov model at the lower level governs the transition toward hidden states ϕ_k's.

$$z_t \mid z_{t-1}, \pi_{z_{t-1}} \sim \pi_{z_{t-1}} \tag{2.53}$$

$$x_t \mid z_t, \phi_{z_t} \sim F(\phi_{z_t}) \tag{2.54}$$

The forward-backward algorithm does not apply to the case to solve infinite HMM, because the number of state K is infinite. This makes the *exact* Bayesian inference for the iHMM intractable. Gibbs sampling provides an approximate inference algorithm without using the forward-backward algorithm.



(a) Before: $p(z_t|z_{t-1}, \boldsymbol{\pi})$ (b) After: $p(z_t|z_{-t}, \alpha)$

Figure 2.7: Integrating out $\boldsymbol{\pi}$. For simplicity, the global measure G is omitted.

Statistical inference for iHMM with collapsed Gibbs sampling: The Gibbs sampler is a technique for generating random variables from a marginal distribution directly [108]. It replaces the difficult calculations of the marginal density with a sequence of straightforward calculations. By simulating a large enough sample, the characteristic of the marginal density can be calculated to the desired degree of accuracy.

In the case of solving the iHMM, the collapsed Gibbs sampler can be used [109], which provides an approximate solution by integrating out the hidden variable $\boldsymbol{\pi}$ and sampling each hidden state variable conditioned on all other variables. Figure 2.7 illustrates the partial graphical models for sampling the hidden states, where the subfigures (a) and (b) present the variable dependencies before and after integrating out $\boldsymbol{\pi}$. Likewise, $\boldsymbol{\phi}$ can also be integrated out.

The Gibbs sampler only changes one hidden state conditioned on all other states, whereas the Beam sampler samples the whole sequence and hence efficiently addresses the slow mixing problem of Gibbs sampler for sequential data with strongly correlated hidden states.

Statistical inference for iHMM with beam sampling: The beam sampling algorithm was proposed by Van Gael et al. [110]. It is essentially a slice sampling that uses dynamic programming to provide an exact inference. It designs some auxiliary variables to partition the transition probability distribution into subsets for efficient sampling of the hidden state sequence. More specifically, for the observed sequence $\{x_t\}_{t=1,2,\ldots,T}$, auxiliary variables $\{u_t\}_{t=1,2,\ldots,T}$ are designed and sampled with probability density,

$$\begin{aligned} p(u_t|z_{t-1}, z_t, \boldsymbol{\pi}) &= \frac{\mathbb{1}(0 < u_t < \pi_{z_{t-1},z_t})}{\pi_{z_{t-1},z_t}} \\ \mathbb{1}(C) &= \begin{cases} 1, \text{ if } C \text{ is true} \\ 0, \text{ otherwise} \end{cases} \end{aligned} \tag{2.55}$$

where each u_t serves as a dynamic threshold at t to partition the probability distribution $\{\pi_{z_{t-1},k}\}_{k=1,2,\ldots,\infty}$ into a finite set of entries larger than u_t and an infinite set smaller than u_t. Only the states k's within the finite set are considered when sampling z_t that transits out of state z_{t-1} during dynamic programming. This reduces the number of potential states to consider and hence makes the inference efficient.

$$\begin{aligned} & p(z_t|x_{1:t}, u_{1:t}, \boldsymbol{\pi}, \boldsymbol{\phi}) \propto p(z_t, u_t, x_t|x_{1:t-1}, u_{1:t-1}, \boldsymbol{\pi}, \boldsymbol{\phi}) \\ =& \sum_{z_{t-1}} p(x_t|z_t, \boldsymbol{\phi}) p(u_t|z_{t-1}, z_t, \boldsymbol{\pi}) p(z_t|z_{t-1}, \boldsymbol{\pi}) \\ & p(z_{t-1}|x_{1:t-1}, u_{1:t-1}, \boldsymbol{\pi}, \boldsymbol{\phi}), \text{and by applying Eq. (2.55), we have:} \\ =& p(x_t|z_t, \boldsymbol{\phi}) \sum_{z_{t-1}:\pi_{z_{t-1},z_t} > u_t} p(z_{t-1}|x_{1:t-1}, u_{1:t-1}, \boldsymbol{\pi}, \boldsymbol{\phi}) \end{aligned} \tag{2.56}$$

A proof for the existence of a finite set of potential states is straightforward: Assume there exists an infinite set of hidden states k's, each can be transited from state z_{t-1} with probability $p(k) = \pi_{z_{t-1},k}$ larger than u_t. Summing up the probabilities of these states, we obtain $\sum_{k:p(k)>u_t}^{\infty} p(k) > \sum_{k:p(k)>u_t}^{\infty} u_t$. Since the overall probabilities from state z_{t-1} equals 1 ($\sum_{k=1}^{\infty} p(k) = 1$), $\sum_{k:p(k)>u_t}^{\infty} p(k) < 1$. This makes $\sum_{k:p(k)>u_t}^{\infty} u_t < 1$, which satisfies only when $u_t \to 0_+$. However, this cannot be true, because $u_t \sim \text{Uniform}(0, \pi_{z_{t-1},z_t})$. Furthermore, Eq. (2.55) implies that the finite set is not an empty set. There is at least one element z_t with transitional probability larger than u_t, because u_t is sampled from the uniform distribution: $\text{Uniform}(0, \pi_{z_{t-1},z_t})$.

As shown in Figure 2.8, no variables are dependent on the auxiliary variables $\mathbf{u}$, so the $\mathbf{u}$ do not change the marginal distribution over other variables. This guarantees that the sampler converges to true posterior. Moreover, the benefit of the auxiliary variables u_t's is that they adaptively truncate the infinitely large transition matrix and hence sparsify the dynamic programming (forward-backward space).

Specifically G is sampled proportional to an additional set of auxiliary variables $\{m_{\cdot k}\}_{k=1,2,\ldots,K}$, where each $m_{\kappa k}$ is independent of others given $\mathbf{z}$, G, and α.

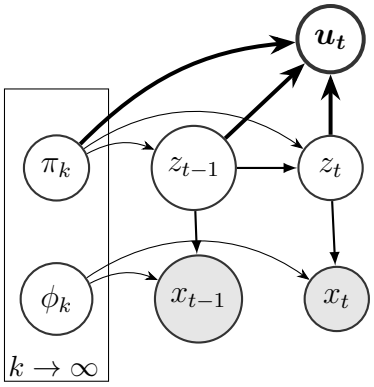


Figure 2.8: The auxiliary variables u depends on z and π.

$$
\begin{aligned}
G &= (G_1 \dots G_K, \sum_{k'=K+1}^{\infty} G_{k'}) \sim \mathrm{Dir}(m_{\cdot 1} \dots m_{\cdot K}, \gamma) \\
m_{\cdot k} &= \sum_{\kappa=1}^{K} m_{\kappa k} \\
p(m_{\kappa k} = m | \mathbf{z}, G, \alpha) &\propto S(n_{\kappa k}, m)(\alpha G_k)^m
\end{aligned}
\tag{2.57}
$$

where $S(\cdot, \cdot)$ denotes Stirling numbers of the first kind. Summing over the infinite many states that never occur in the hidden state sequence $\{z_t\}$, the conditional distribution $\pi_{k\cdot}$ given its Markov blanket $\mathbf{z}$, G, and α is

$$
\begin{aligned}
\pi_{k\cdot} &= (\pi_{k1} \dots \pi_{kK}, \sum_{k'=K+1}^{\infty} \pi_{kk'}) \\
&\sim \mathrm{Dir}(n_{k1} + \alpha G_1 \dots n_{kK} + \alpha G_K, \alpha \sum_{k'=K+1}^{\infty} G_{k'})
\end{aligned}
\tag{2.58}
$$

where $n_{k\kappa}$ denotes the transition counts from state k to κ. Each $\phi_{k\cdot}$ depends on the state sequence $\{z_t\}$, the observed concept sequence $\{x_t\}$, and the prior distribution H, and the $\phi_{k\cdot}$'s are independent given $\mathbf{z}$, $\mathbf{x}$, and H.

$$\phi_{k\cdot} \sim \text{Dir}(l_{k1} + H_1 \ldots l_{k|V|} + H_{|V|}) \tag{2.59}$$

where l_{kv} denotes the emission counts from state k to medical concept v. The vocabulary set is V. The hyper-parameters α and γ can be sampled according to [111].

In Section 3.5.2, we develop a sampling solution that considers using multiple concept sequences with different lengths as observations. Our model extends the infinite HMM and the beam sampler and it is used to automatically summarize the diagnostic cognitive states from physicians' verbal narratives.

Chapter 3

Modeling Diagnostic Verbal Narratives for Medical Conceptual Topics

3.1 Background

Image-based problem-solving is an important research area in imaging, computing, cognitive sciences, and application domains. It involves the domain-specific concepts present in images, human vision and psychophysics, and knowledge discovery. In order to perform computational image understanding tasks (e.g., object detection [112, 113, 114], shape estimation [115], and depth estimation [116]), human knowledge (e.g., the sky is blue and may appear in the upper half of an image) is used in designing computer vision algorithms. Such static and limited use of human knowledge is not sufficient for image analysis and classification in specialized domains, as there is no direct transformation from domain knowledge to particular image features and algorithms behave differently from domain experts engaged in decision-making. For example, content-based image retrieval (CBIR) fails to work well in the medical domain, because it is subject to the *semantic gap* between visual image features and the richness of human understanding [117, 118, 119].

Since parsing complex semantics in medical images requires domain

knowledge and perceptual expertise in pattern recognition, it is necessary to learn high-level image representations from physicians' inputs for computational image classification [120]. Language is the primary conduit to express meaning, so researchers are currently integrating verbal metadata with image features [39] and moving toward new directions in CBIR, such as association-based image retrieval (ABIR) and perception-based image retrieval (PBIR) [121, 122, 123]. We take these insights in new directions by using novel linguistic data (i.e., physicians' diagnostic verbal narratives) for medical image understanding.

Medical images are inherently complex and noisy due to both photographic inconsistency and different presentations of even the same medical condition. Organizing relevant medical images into semantically related and meaningful groups has been a long-standing challenge. In spite of recent efforts to incorporate domain knowledge in image clustering [44], understanding physicians' use of knowledge (for the sake of medical image clustering) remains a challenging problem that must be studied in the cognitive and biomedical domains. Our elicited physicians' diagnostic verbal narratives contain expert domain knowledge obtained from years of clinical training. A basic assumption in this study is that the meaning of a medical image is mirrored by the spoken narrative of a physician describing the image content. Based on such rationale, we exploit expert knowledge from diagnostic verbal narratives to facilitate medical image grouping, and our methodology is more natural, objective and automated than current research [22, 46, 47].

This chapter does not aim to construct an all-inclusive medical topic dictionary. Rather, it aims at analyzing experts' expressions of domain knowledge and developing an automated framework to extract an interpretable conceptual representation for medical images, so as to facilitate medical image analysis and grouping.

3.2 Medical Term Extraction

Table 3.1: A diagnostic narrative corresponding to Figure 1.1 with time stamps and tokens. There is a multiword expression (*basal cell carcinoma*) boxed in the middle rows. Besides all the uttered words, tokens also include spoken disfluency markers (e.g., *um*, *uh*, repetition, and edits), as well as pause markers, which constitute regular features of speech. These special tokens are explored in another study and are ignored in this dissertation [17].

Start time (sec)	End time (sec)	Token uttered
start 1 (0)	*end* 1 (1.40)	*token* 1 (SIL)
start 2 (1.40)	*end* 2 (2.27)	*token* 2 (um)
start 3 (2.27)	*end* 3 (2.52)	*token* 3 (this)
start 4 (2.52)	*end* 4 (2.69)	*token* 4 (is)
start 5 (2.69)	*end* 5 (3.19)	*token* 5 (a)
⋮	⋮	⋮
start i (25.23)	*end* i (25.76)	*token* i (basal)
start $i+1$ (25.76)	*end* $i+1$ (26.05)	*token* $i+1$ (cell)
start $i+2$ (26.05)	*end* $i+2$ (26.30)	*token* $i+2$ (um)
start $i+3$ (26.30)	*end* $i+3$ (27.20)	*token* $i+3$ (carcinoma)
⋮	⋮	⋮
start $n-2$ (46.48)	*end* $n-2$ (47.58)	*token* $n-2$ (and)
start $n-1$ (47.58)	*end* $n-1$ (48.19)	*token* $n-1$ (surrounding)
start n (48.19)	*end* n (48.64)	*token* n (telangiectasias)

A script was written to extract medical terms from the collected verbal narratives using the UMLS and MetaMap. The UMLS constitutes the largest existing semantic network of medical terms and lexical relations [18, 28], and

it has been used as a knowledge source for medical terminology research and information retrieval [27]. The MetaMap, developed at the National Library of Medicine, is a knowledge-intensive tool that can annotate biomedical text tokens by UMLS Metathesaurus concepts [31]. In our data, multiword expressions are quite common. For example, in the sentence *This is a basal cell um carcinoma*, where *um* is a disfluency marker, the multiword expression *basal cell carcinoma (BCC)* is detected by MetaMap as a medical concept. To handle such cases for the whole dataset, the script detects and reconstructs medical multiwords by joining adjacent medical words and discarding disfluency markers (e.g., *um* and *uh*) and pause markers (i.e., *SIL*) within a medical multiword. The corresponding time intervals are combined into one. This process transforms each spoken narrative into a sequence of words and multiword expressions. In the example *(SIL) um this is a basal cell um carcinoma* shown in Table 3.1, the multiword after *this is a* is detected as *basal cell carcinoma* with the disfluency marker *um* discarded. An illustration of a narrative after medical multiword reconstruction is in Table 3.2. The analyses and modelings in the following sections and chapters use these preprocessed spoken narratives from physicians.

Table 3.2: An illustration of the narrative in Table 3.1 after the detection of a medical multiword expression, *basal cell carcinoma*.

Start time (sec)	End time (sec)	Token uttered
⋮	⋮	⋮
start j(25.23)	*end j*(27.20)	*multiword j*(*basal cell carcinoma*)
⋮	⋮	⋮

3.3 Clustering Verbal Narratives

Intuitively, physicians use similar sets of terms to describe cases within a category and dissimilar ones for between categories. To validate this assumption we first visualize the narratives in a narrative-term matrix to stress the outstanding groups of our data, and then cluster the diagnostic narratives based on physicians' use of medical terms (i.e., words and multiwords). The representation learning algorithms we used for clustering outperforms its counterparts, and it facilitates the use of domain experts' input for medical image grouping.

3.3.1 Ground truth for narrative clustering

If shown similar images, the physicians should describe these images similarly. Since we used our first image set which represents a wide range of dermatology diagnoses, the ground truth for narrative clustering consists of 48 image labels, which are the correct diagnoses of the 48 dermatology images. In other words, the narratives corresponding to the same image are labeled the same in the ground truth.

3.3.2 Narrative processing and visualization

To visualize the distinct uses of terms for describing different images, we organize the medical terms in the dataset into a narrative-term matrix. This matrix specifies each term's frequency of occurrence in a specific narrative. The values in the matrix are tf-idf scores (term frequency times inverse document frequency) [124] as illustrated in Figure 3.1. The 768 narratives that

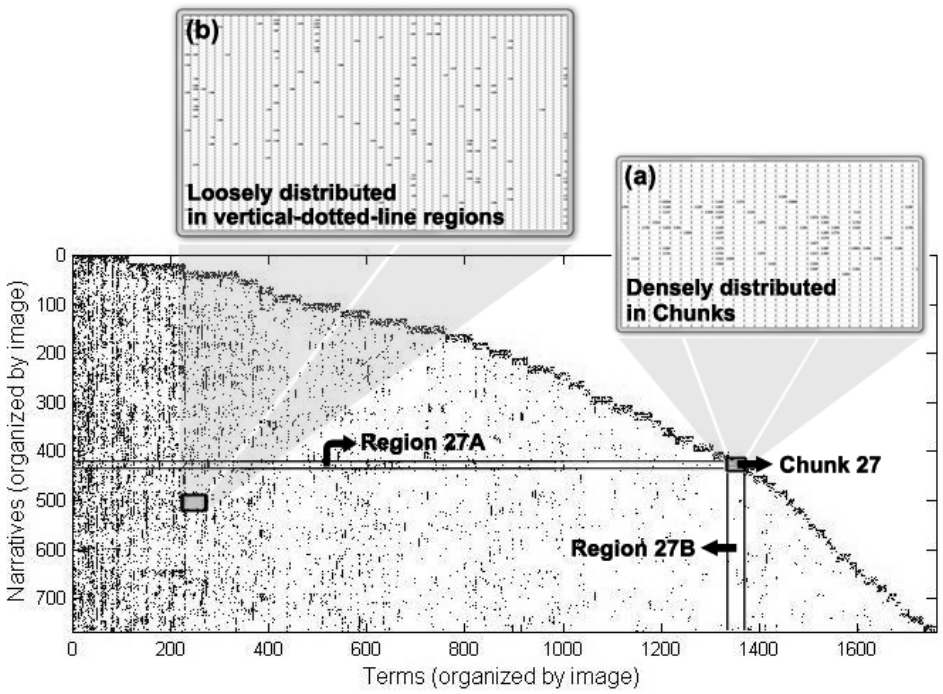


Figure 3.1: The narrative-term matrix with tf-idf scores is organized by image. The zero scores are plotted in white and others in dark grey. The "chunks" with dense non-zero scores along the pseudo-diagonal indicate that these terms are associated with 16 consecutive narratives (corresponding to the same image), and there are 48 such chunks. Region 27A, combined with chunk 27, displays all the terms used to describe narratives 417-432 (image 27); and, region 27B, with chunk 27, displays all the narratives involving terms 1339-1364. Magnification (a) shows the region of chunk 27 and its surroundings. Magnification (b) shows a small part of the region that contains vertical dotted lines.

correspond to the 48 dermatology images in Experiment I are expressed by more than 1700 medical terms (words and multiwords). In Figure 3.1 (a), chunk 27 contains non-zero scores that are densely distributed. Chunk 27 is powerful for differentiating image 27 from the rest. In Figure 3.1 (b), the

region contains loosely distributed non-zero values that form vertical dotted lines. These dotted lines represent the occurrence of medical terms that can be used to describe the manifestation of multiple different diseases, such as *papules* and *plaques*, which indicate the primary morphology of dermatological lesions, or *scale* and *erosion* which indicate the secondary morphology. These terms are also helpful in differentiating the images, since they each are shared by a subset of diseases, though they are not tightly linked to and specifically used for describing one particular disease.

The terms useful for image clustering should only occur in a few images but in a large enough number out of all the narratives corresponding to these few images. For example, the names of the correct diagnoses and their synonyms are medical terms that are labeled as the correct diagnoses of the 48 images in our experiment, such as *lichen planus*, *BCC*, and *melanoma*. Each of these terms occurs in a small number of images, because the 48 images in our experiment cover a wide range of dermatology diagnoses. Some of them only occur in a small number of narratives, since not all physicians managed to utter the correct diagnosis. The primary morphology terms, such as *papule*, *plaque*, *nodule*, and *patch*, are also useful for distinguishing among medical images. Primary morphology is highly informative in categorizing a medical image into a certain disease type. Other terms describe information such as patient demographics, lesion body location, lesion distribution, and secondary morphology. The occurrences of major categories of medical terms in our experiment are analyzed and visualized in Figure 3.2.

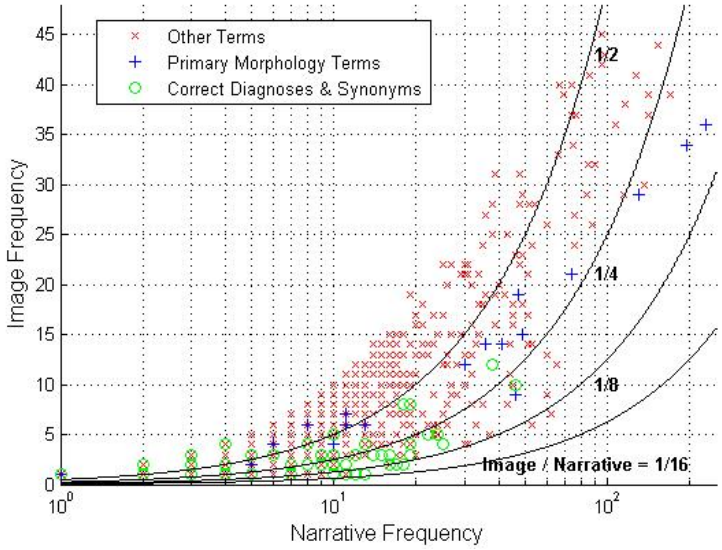


Figure 3.2: An analysis of the occurrence of medical terms in narratives and distribution across images. Primary morphology terms are highlighted as blue crosses. The dark solid curves in this figure represent the ratio of number of images to number of narratives that include a term, being 1/2, 1/4, 1/8, and 1/16, respectively. All terms are on one side of line 1/16, because there are 16 subjects describing each image. The terms between curves 1/8 and 1/16 are the word tokens uttered by a majority of participating physicians when describing the same image. The terms above line 1/2 are considered to be least useful.

3.3.3 Topic modeling the narratives

Since the spoken narratives contain physicians' image understanding, we exploit these narratives to incorporate domain knowledge in medical image grouping. Following the convention of document clustering, each narrative $\mathbf{x}_i$ is denoted by a term vector of length $|V|$, in which $\mathbf{x}_i(j)$ is set to the normalized frequency (or other metrics such as tf-idf) of term t_j if $t_j \in \mathbf{x}_i$ and 0 otherwise. The length of $\mathbf{x}_i$ equals the size of the vocabulary V, which consists of all

medical terms in the dataset. As different physicians may have distinct medical backgrounds and preferences for term choices, various naming conventions may be used by different narratives. This variation results in a large number of distinct terms especially when scaling to a massive number of images.

Partitioning-based clustering algorithms (e.g., K-means) have good clustering performance and can scale to a large number of medical images. Two term vectors are regarded as similar if they share a large enough number of terms. Directly applying these algorithms to spoken narrative clustering leads to poor clustering quality, because the term vectors for narrative descriptions can be very sparse and hence less likely to share common terms. Advanced algorithms, such as those based on matrix factorization (e.g., SVD co-clustering [125] and NMTF [126]), have been demonstrated to be more effective in dealing with the limitations of data sparseness. However, the high complexity of these algorithms leads to a computational bottleneck when clustering large-scale collections of medical images.

Inspired by the advances in latent factor models and sparse coding [127, 5], we devise a novel strategy to discover a set of latent medical concepts, referred to as *anchor concepts*. The anchor concepts aim to capture the high-level medical concepts of the spoken narratives. Ideally, they are expected to map to the true medical concepts in the spoken narratives. For example, an anchor concept that corresponds to a body part may include terms like arm, leg, and face. Similarly, another anchor concept may correspond to the primary morphology, which covers terms like papule and plaque. Nonetheless, due to

the various naming conventions and the diverse usage of terms by different physicians, such a perfect mapping is almost impossible for real-world verbal descriptions of medical images. Therefore, the anchor concepts are better understood as latent medical concepts that capture the underlying semantics of the verbal descriptions. In this regard, the anchor concepts are in line with the *quality dimensions* in Gärdenfors' conceptual spaces theory [128, 129]. In particular, by computing the anchor concepts, we are able to discover the *hidden conceptual space* that corresponds to the underlying semantics of the verbal descriptions.

Specifically, we compute a matrix $\boldsymbol{H} = \{\mathbf{h}_1, ..., \mathbf{h}_k\}$, where each $\mathbf{h}_i \in \mathbb{R}^n$ denotes an anchor concept and is a linear combination of a set of term vectors (represented by a matrix $\boldsymbol{X}$), where the j-th column denotes the term vector of narrative $\mathbf{x}_j$:

$$\boldsymbol{H} = \{\mathbf{h}_1, ..., \mathbf{h}_k\} = \boldsymbol{X}\boldsymbol{W} \tag{3.1}$$

$$\mathbf{h}_i = \boldsymbol{X}\mathbf{w}_i = \sum_{j=1}^{m} \boldsymbol{W}_{ij}\mathbf{x}_j, \forall i = 1, ..., k \tag{3.2}$$

where $\boldsymbol{W} = \{\mathbf{w}_1, ..., \mathbf{w}_k\} \in \mathbb{R}^{m \times k}$ is a weight matrix. Each entry $\boldsymbol{W}_{ij}$ denotes how much narrative $\mathbf{x}_j$ contributes to anchor concept $\mathbf{h}_i$.

The desired anchor concepts are expected to capture high-level medical concepts of the spoken narratives that can be used to recover the original term vectors of the narratives. Meanwhile, since a narrative is used to describe a specific medical image, it is common for the narrative to concentrate on a small number of medical concepts. Stated differently, a spoken narrative is

expected to be only related to a small subset of anchor concepts. Therefore, a desired anchor concept set can be identified by optimizing the following objective function:

$$\min_{\boldsymbol{H},\mathbf{c}_i\geq 0} \sum_{i=1}^{m} \|\mathbf{x}_i - \mathbf{H}\mathbf{c}_i\|^2 + \lambda||\mathbf{c}_i||_1 \quad (3.3)$$
$$\text{subject to } ||\mathbf{h}_j||^2 \leq a, \forall j = 1, ..., k$$

where $\mathbf{c}_i \in \mathbb{R}_+^k$ is the coefficient vector with $\mathbf{c}_i(j)$ signifying the correlation between $\mathbf{x}_i$ and anchor concept $\mathbf{h}_j$, and a is a constant. Since each narrative should concentrate on a small number of medical concepts, $\mathbf{c}_i$ is expected to be sparse. We achieve sparse $\mathbf{c}_i$ $(i = 1, ..., m)$ by incorporating an l_1-norm (i.e., $||\mathbf{c}_i||_1$) and λ is the penalty parameter that controls the sparsity ratio. The l_1-norm has been demonstrated to be effective in finding sparse solutions. Hence, the second term of Eq. (3.3) corresponds to a sparsity constraint on $\mathbf{c}_i$. The constraint on the size of the anchor concept, i.e., $||\mathbf{h}_j||^2 \leq a$ where $||\cdot||$ is the vector l_2-norm, avoids arbitrarily large anchor concept vectors that keep $\boldsymbol{H}\mathbf{c}_i$ unchanged while making $\mathbf{c}_i$ arbitrarily close to zero.

The coupling between $\boldsymbol{H}$ and $\boldsymbol{C}$ in the first term of Eq. (3.3) makes the overall objective function non-convex. Nonetheless, we can resort to an efficient iterative algorithm to find a local optimum. More specifically, by fixing $\boldsymbol{C}$, Eq. (3.3) is convex in $\boldsymbol{H}$, and an optimal $\boldsymbol{H}^*$ can be found by solving a constrained least square problem. Then, $\boldsymbol{H}^*$ is kept fixed, which turns Eq. (3.3) into a convex function over $\boldsymbol{C}$. An optimal $\boldsymbol{C}^*$ can be achieved by

solving an l_1 regularized least square problem. This iterative process continues until a local optimum is achieved.

Once $\boldsymbol{H}$ and $\boldsymbol{C}$ are computed, the coefficient vector $\mathbf{c}_i$ can be regarded as the projection of term $\mathbf{t}_i$ onto the anchor concept space $\boldsymbol{H} = \{\mathbf{h}_1, ..., \mathbf{h}_k\}$. As the number of anchor concepts is much smaller than the number of terms, the anchor concept space $\boldsymbol{H}$ essentially forms a lower dimensional space to represent the verbal descriptions. This makes it similar to other dimensionality reduction techniques, such as LSA [130], which is widely employed in language understanding. The key difference between the proposed anchor concept-based approach and existing dimensionality reduction technique lies in the addition of the sparsity constraint to the coefficient vectors. In particular, the coefficient vector $\mathbf{c}_i$ captures the relevance between the i-th narrative and all the k anchor concepts. The sparsity constraint on $\mathbf{c}_i$ ensures that each narrative is only related to a small number of anchor concepts. This is in line with the fact that when a physician describes a specific medical image, his/her spoken narrative is usually focused on a small number of medical concepts. Furthermore, spoken narratives can be easily separated based on their distinct relationships with the anchor concepts. Clustering of spoken narratives can be achieved by applying a simple clustering algorithm (e.g., K-means) to the coefficient matrix $\boldsymbol{C}$.

3.3.4 Narrative clustering performance evaluation

We adopt two metrics to measure the clustering quality: accuracy (i.e., AC) and mutual information (i.e., MI). Both AC and MI are widely used metrics

to assess the performance of clustering algorithms.

- **Accuracy:** For a given narrative $\mathbf{x}_i$, assume that its cluster label is l_i and its ground truth label is g_i. The AC metric is defined as follows:

$$AC = \frac{\sum_{i=1}^{m} \delta(g_i, map(l_i))}{m} \tag{3.4}$$

  where m is the total number of narratives in the dataset. The function $\delta(x, y)$ is a delta function that equals 1 if $x = y$ and equals 0 otherwise. Moreover, $map(l_i)$ is the permutation mapping function that maps each assigned cluster label to the equivalent ground truth label. The best mapping between the two sets of labels is achieved by the Kuhn-Munkres algorithm [131].

- **Mutual Information:** Let $\mathcal{D}$ be the set of true narrative groups and $\mathcal{C}$ be the narrative clusters obtained from a clustering algorithm. The mutual information metric $MI(\mathcal{D}, \mathcal{C})$ is defined as follows:

$$MI(\mathcal{D}, \mathcal{C}) = \sum_{d_i \in \mathcal{D}, c_j \in \mathcal{C}} p(d_i, c_j) \log_2 \frac{p(d_i, c_j)}{p(d_i)p(c_j)} \tag{3.5}$$

  where $p(d_i)$ and $p(c_j)$ are the probabilities that a randomly selected narrative belongs to group d_i and cluster c_j, respectively. Furthermore, $p(d_i, c_j)$ is the joint probability that the randomly selected narrative belongs to both group d_i and cluster c_j.

To demonstrate the effectiveness of the proposed narrative clustering strategy, we include two widely used clustering algorithms, K-means and spectral clustering [132], for comparison purposes. We also include one widely used

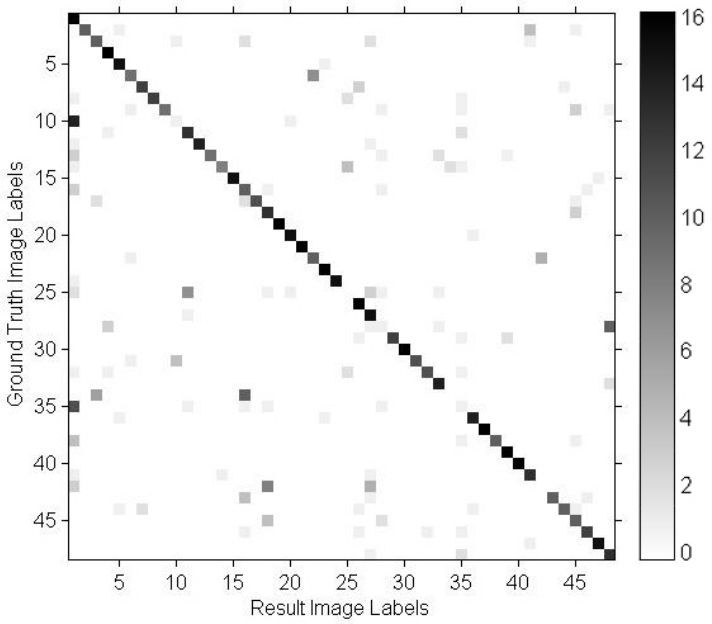


Figure 3.3: Confusion matrix of clustering results based on the anchor concept algorithm. The darkness of a block indicates the number of narratives that are in this block. For example, all 16 narratives describing image 1 are correctly clustered, so the block in the upper left corner is dark. There are few dark blocks off the diagonal, showing that only a small number of narratives are misclassified. As an example, some of the narratives describing image 10 are mislabeled as image 1.

dimensionality reduction algorithm for comparison: LSA [130]. The K-means algorithm is then applied to the low-dimensional space to achieve clustering. Table 3.3 summarizes the clustering results from all four algorithms. The proposed anchor concept-based clustering achieves the best results in both AC and MI categories. A confusion matrix visualization based on the anchor concept algorithm is shown in Figure 3.3. The labels are mostly aligned with ground truth. More detailed discussions of the clustering results are in [37].

Table 3.3: Narrative clustering performance.

Algorithms	AC (%)	MI (%)
K-means	51.69	68.53
Spectral clustering	35.68	63.15
LSA-based clustering	54.04	68.35
Anchor Concept-based clustering	**70.70**	**80.62**

3.4 Developing Lexical Metrics

Two metrics, namely *lexical consensus score* and *top N relatedness scores*, are developed for the verbal narratives in order to answer various questions that benefit medical image understanding.

3.4.1 Lexical consensus score

In order to know whether physicians agree in their image understanding and description, and to use this information as a new feature for image grouping, we propose a *lexical consensus score* S_C of an image to evaluate the degree of agreement among descriptions by all physicians. This approach was adapted in modified form from cohesion scores typically used in clustering [133]. High agreement among K physicians leads to a high consensus score.

We define the lexical consensus score of an image to be the average pairwise cosine similarity among descriptions given by K physicians $\{p_1, p_2, ..., p_K\}$ in terms of their use of medical terms. We adopt cosine similarity in our study, because cosine similarity is widely used in text analysis [134]. We assume that we have a set of images $\{m_1, m_2, ..., m_R\}$. For each image m_i from the R images, we gather the spoken narratives from K physicians and form a vocab-

ulary V_i, which contains all the medical words/multiwords used for describing this image. This vocabulary is treated as a feature space for this image, so that the K corresponding narratives could be expressed using feature vectors in this space. Because a physician cannot be compared to him/herself, K physicians form $K(K-1)/2$ pairs, thus $K(K-1)/2$ cosine similarity scores can be computed. Since the feature spaces may not be of the same length across R images, the similarity scores per image are normalized by the length of feature vector, $|V_i|$. We average the $K(K-1)/2$ normalized similarity scores and define it to be the lexical consensus score of the image, which measures K physicians' level of agreement on describing the image.

$$S_C = \frac{2}{K(K-1) \times |V_i|} \sum_{j=1}^{K} \sum_{k=j+1}^{K} sim(\mathbf{x}_j, \mathbf{x}_k) \tag{3.6}$$

where $\mathbf{x}_j$ and $\mathbf{x}_k$ are narratives from the narrative set $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_K\}$ that describes the same image, and $sim(\mathbf{x}_j, \mathbf{x}_k)$ refers to the similarity between $\mathbf{x}_j$ and $\mathbf{x}_k$.

3.4.2 Top N relatedness scores

We defined the top N overall relatedness scores S_{R_N} of a spoken narrative at the concept level, in contrast to the lexical consensus score at the token level, to capture the level of semantic relevance of a physician's description. In a case study, we relate each narrative to the correct diagnosis of the corresponding image. More targets other than correct diagnoses can also be used for other purposes, e.g., using *primary morphology* of dermatological lesions to group

dermatology images by primary morphology.

Now assume we have a target medical concept with which to relate. Since we need to know how a narrative is related to a target concept, an intuitive way is to first know how each concept within the narrative is related to the target concept and average their relatedness. For this purpose, we used an open source software package UMLS::Similarity [135], which uses UMLS to calculate the semantic relatedness between two medical concepts. The semantic relatedness quantitatively measures the degree to which the semantic features overlap between the two terms. To result in high relatedness, two terms may be related through collocation (i.e., occurring more often than by chance) as *needle* and *thread* but not necessarily be synonyms or hyponym/hypernym in a hierarchical semantic relationship [136]. For example, the sensation of *itchy* can be caused by *rashes*, so these two terms are highly related. Examples can be explored with this tool's web interface (http://atlas.ahc.umn.edu/cgi-bin/umls_similarity.cgi). This package builds upon UMLS to calculate similarity and relatedness between medical concepts and has been used for studies involving senses and relations of medical terms [137, 138]. In order to return a normalized score in the range [0, 1] so that the scores can be effectively compared across or averaged by narratives, we select the *vector measure* method [139]. Generally, this method utilizes the definitions of both medical concepts in UMLS and their second-order co-occurrence. This captures the chance that both concepts will occur in the same context, given a large medical corpus (see discussion of details by Liu et al. [139]).

Specifically for broader coverage and an empirical setting, we used the extended definitions of concept unique identifiers (CUIs) for both terms. The definition extensions are based on *has parent/child* relationship, as well as the relationship of *has a broader/narrower*, given the complete Metathesaurus vocabulary inventory, which includes resources such as *MeSH* and *SNOMED CT*. According to the above configuration, the relatedness score is within the range of [0, 1], 0 being not relevant at all and 1 being the same concept (either the exact word/multiword or its synonym).

Because of homonymy (i.e., two words having the same spelling but different meanings) and polysemy (i.e., one word having different but related meanings), there are situations where a word or a multiword has multiple medical meanings indexed by multiple CUIs in UMLS. In these situations, word sense disambiguation [140] is necessary. Before calculating the relatedness score to the image's correct diagnosis, we disambiguate the intended sense. We employ a package, UMLS::SenseRelate [141], for this purpose. Since not all words in narratives are medical terms, the scores of non-medical terms are set to zero. To infer which words and multiwords represent medical terms, we used UMLS packages to search for the CUI of each word/multiword in the narrative. If there is a CUI (or CUIs) related to this word/multiword, then we consider it as a medical term; otherwise we do not, although the UMLS is not fine-tuned in our context of the dermatology domain. Since each narrative can be annotated by several medical concepts that are, more or less, related to the image content, we define the top N relatedness scores to capture multiple

levels of semantic relevance.

The top 1 relatedness score in a narrative is the highest score calculated from the relatedness of all medical concepts to the correct diagnosis within a given narrative. This score thus corresponds to the most relevant medical concept from the physician describing this image and can intuitively serve as a good indicator of a physician's understanding of a particular image case. Likewise, the top N relatedness score is the average of the highest N scores in this narrative. We use the top N relatedness score to measure, for a specific narrative, the subject's understanding of the image s/he described. The reason for calculating the top N relatedness score, rather than averaging all the scores within one narrative, is that different physicians use different narration styles [34], either comprehensive or brief. What we want here is to reduce the variation among subjects to identify what they share.

After calculating the relatedness score for each term, we retrieved the top N relatedness scores from each narrative. Averaging the top N relatedness scores from K narratives $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_K\}$ that correspond to a same image gives us the top N relatedness score for that image. After computing the top N relatedness scores for all images, we rank these images by their top N relatedness scores arranged in order.

$$S_{R_N} = \frac{1}{K \times |N_j|} \sum_{j=1}^{K} \sum_{n_i \in N_j} rel(n_i, c) \tag{3.7}$$

where N_j is the set of terms in narrative $\mathbf{x}_j$ that have the top N relatedness

scores. For example, if the size of the set N_j is 1, only the top term is included. If the size of the set is 3, the top 3 terms are included. Additionally, n_i is a term in this set; $rel(n_i, c)$ is the relatedness score of n_i to the target concept c. For example, in the case study the target concept is the correct diagnosis of the image case.

3.4.3 Evaluation of the lexical metrics

We conducted a case study (i.e., *challenge levels* of image-based diagnoses), in which we applied the metrics of lexical consensus score and top N relatedness score, in order to evaluate the metrics. This case study was designed based on an intuition that the images which are easier to diagnose would be described with more consensus lexical choices by physicians, and the concepts referred to by physicians' utterances would be more closely related to the correct diagnosis.

We have generated two different ranked lists as ground truth. One list was generated from a dermatologist who did not participate in the data collection experiment as a subject. This physician ranked difficulty levels of image cases based on personal expertise without knowing the performance of the participating physicians in the prior experiment. The other ground truth list was obtained from data analysis of judgments (made by the same dermatologist and two other non-involved dermatologists) of the performance of the 16 participating physicians, including the correctness of their described primary morphology, differential diagnosis, and final diagnosis. Prior inter-annotator

analysis with the kappa metric indicates good inter-annotator agreement of correctness judgments for differential diagnosis as well as final diagnosis but less agreement on medical lesion morphology [8]. This ranking is linked to participants' performance based on data analysis. We compared the two ranked lists and found that they are highly correlated. The Spearman correlation score between the two lists is 0.77, and the Kendall correlation score is 0.56. The equations of Spearman and Kendall correlation measurements are shown below, respectively:

- Spearman coefficient: $r_s = \frac{1}{2\sqrt{S_X S_Y}} \{S_X + S_Y - \sum_i d_i^2\}$, where X and Y are ranked lists, X_i and Y_i are ordinals, $i = 1...N$. $S_X = \sum_i x_i^2$, $x_i = X_i - \bar{X}$, $\bar{X} = \frac{1}{N}\sum_i X_i$, and d_i is the difference between ordinals X_i and Y_i [142, 143].
- Kendall coefficient: $\tau = \frac{C - D}{C + D}$, where C is the total number of concordant pairs between the two ranked lists and D is the total number of discordant pairs [144].

Although gained from different resources, the correlation scores derived from both measurements suggest high correlation, and the two ranked lists are both useful as ground truth of the challenge levels of evaluating different image cases. These two ranked lists are referred to as *difficulty* and *correctness*.

The ranking of the 48 images based on the consensus score is positively correlated with both the difficulty-based ground truth and the correctness-based ground truth. The Spearman correlation score between the ranking

based on consensus scoring and the difficulty ground truth is 0.64 (well correlated), and that between the consensus-based ranking and correctness-based ground truth is 0.57. According to Kendall correlation metrics, the consensus-based ranking is well correlated to both ground truth rankings. See Table 3.4.

Since each narrative contains more than one medical concept, we defined the top N (with $N = 1, 3, 5, 10$) relatedness scores to capture multiple levels of semantic relevance. We compared the rankings of 48 images based on the top 1, 3, 5, and 10 relatedness scores with the rankings suggested by non-participating physicians. The correlation scores between the relatedness-based rankings and the ground truth rankings are also measured by Spearman and Kendall matrices and listed in Table 3.4. More detailed analyses of the study of image-based diagnostic challenge levels can be found in my journal paper [37].

To incorporate domain knowledge in the objective of learning a representation, the pairwise concept relatedness scores can be used as a constraint. For example, the verbal narratives that contain terms that are synonyms or highly-related will be constrained to be similar. This approach exploits the term relations to compute narrative similarities, as opposed to simply using term occurrences. Figure 3.4 illustrates the relatedness between each pair of two medical terms. Such information can be used to derive new constraints.

$$\min_{\boldsymbol{H},\boldsymbol{C}\geq 0} \|\boldsymbol{X} - \boldsymbol{H}\boldsymbol{C}\|_F^2 + \lambda_1 \|\boldsymbol{H}\|_F^2 + \lambda_2 \sum_{p\neq q} \boldsymbol{R}_{pq} \|\mathbf{h}_p - \mathbf{h}_q\|^2 \tag{3.8}$$

Table 3.4: Correlation among different image rankings based on the Spearman and Kendall methods, respectively. The first two rows and columns are ground truth rankings based on difficulty and correctness, respectively. The third row and column are consensus rankings, and the rest are rankings based on top N relatedness scoring. The well-correlated pairs of rankings are highlighted in bold.

(Spearman)	Difficulty	Correctness	Consensus	Top 1	Top 3	Top 5	Top 10
Difficulty	1	**0.77**	**0.64**	0.53	-0.13	-0.29	-0.36
Correctness	-	1	0.57	**0.75**	0.01	-0.16	-0.25
Consensus	-	-	1	0.34	-0.33	-0.46	-0.51
Top 1	-	-	-	1	0.54	0.37	0.26
Top 3	-	-	-	-	1	**0.97**	**0.91**
Top 5	-	-	-	-	-	1	**0.97**
Top 10	-	-	-	-	-	-	1
(Kendall)	Difficulty	Correctness	Consensus	Top 1	Top 3	Top 5	Top 10
Difficulty	1	**0.56**	**0.45**	0.38	-0.08	-0.17	-0.22
Correctness	-	1	**0.41**	**0.56**	0.02	-0.10	-0.17
Consensus	-	-	1	0.26	-0.21	-0.31	-0.35
Top 1	-	-	-	1	0.38	0.26	0.18
Top 3	-	-	-	-	1	**0.87**	**0.76**
Top 5	-	-	-	-	-	1	**0.88**
Top 10	-	-	-	-	-	-	1

where $\boldsymbol{R} \in \mathbb{R}^{\|v\| \times \|V\|}$ is the term-term relatedness matrix in which $\boldsymbol{R}_{pq}$ measures the semantic relatedness score between the p-th and the q-th term. The $\mathbf{h}_p$ and $\mathbf{h}_q$ are terms p and q represented in new space. Highly-related terms are constrained to have similar representations, and other terms may have dissimilar representations.

3.5 Modeling for Diagnostic Narration Patterns

Natural language processing models, such as bag-of-word and N-gram, have been used to analyze clinical texts. Since different practitioners express similar

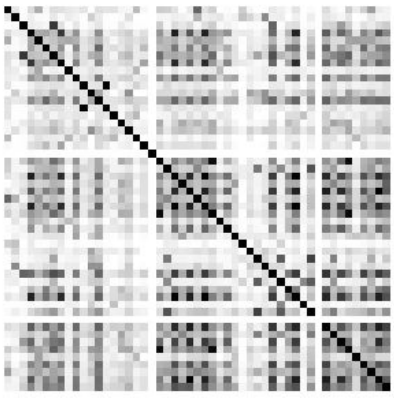


Figure 3.4: Medical term relatedness: Darkness illustrates the relatedness between two terms. The diagonal shows self-relatedness. The black elements off the diagonal show synonyms. Only a subset of the full matrix is visualized here as the overall number of terms is too large.

meanings in a variety of ways both syntactically and lexically, the medical datasets tend to be sparse. To tackle this, topic modeling approaches, such as latent semantic analysis (LSA) [130] and latent Dirichlet allocation (LDA) [100], transform original vocabulary to latent variables (a.k.a., topics) whose mixtures summarize the documents more abstractly. Despite the advantage to produce high-level representations, topic modeling approaches do not consider the word order within each document.

To recognize temporal patterns in the documents and speech data, hidden Markov model (HMM) can be used [107]. It learns a sequential structure of hidden states (i.e., patterns), each being a probability distribution over the vocabulary. We propose to use HMMs to model physicians' spoken narratives, because the order of thoughts (i.e., cognitive states) is crucial to diagnostic decision-making [19, 145]. To automatically determine the optimal number of

hidden states, Teh et al. developed a Bayesian non-parametric HMM using hierarchical Dirichlet processes (HDP-HMM) [111] and Van Gael et al. developed the *beam sampler* for it to limit the computational costs [110]. We extend the HDP-HMM (a.k.a., infinite HMM or iHMM) model and the beam sampler to allow learning from a group of medical concept sequences. This results in a desired sequential representation, based on which we train classifiers to differentiate narrative groups.

3.5.1 Gold standard

Thought units: Medical doctors are recruited to partition and label diagnostic reasoning records into meaningful units of thought (see Table 1.1) [8]. These thought units cover the terminology to standardize the description of skin lesions, including lesion arrangement, distribution, texture, color, primary lesion type, and diagnosis [10]. We use this thought unit labeling as a gold standard in our study to evaluate and interpret the patterns discovered by the model. Since our image set contains a wider range of diagnoses, their thought unit labels do not cover our whole vocabulary.

Diagnostic correctness levels: We recruit three dermatologists to evaluate the narratives from the 16 participating physicians in Experiment I in terms of their diagnostic correctness. A correctness score is assigned to each narrative, which balances the correctness of described Type II thoughts (see the lower half of Table 1.1; i.e., primary lesion type, differential diagnosis, and final diagnosis). This score ranges from 0 to 3 and its distribution across narratives

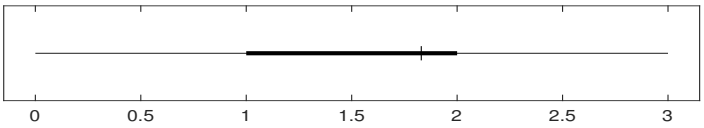


Figure 3.5: The correctness score distribution across all narratives.

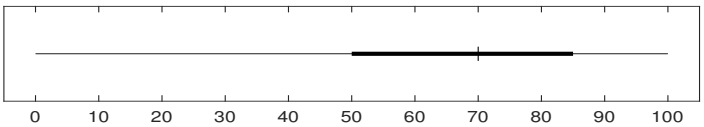


Figure 3.6: The self-reported diagnostic confidence score distribution across all narratives.

is in Figure 3.5. We define the correctness score below 1 (inclusive) as low-correctness and that above 2 (inclusive) high-correctness. These two levels of narratives in Experiment I are classified using the patterns discovered by the model, and the differences in narration patterns between two classes are visualized.

Diagnostic confidence levels: During Experiment II, each participating physician is required to self-report her diagnostic confidence at the end of narration. Figure 3.6 shows how the diagnostic confidence scores are distributed across all narratives. We define the bottom quartile (0%–50% confidence, inclusive) as low confidence and the top quartile (85%–100% confidence, inclusive) high confidence.

3.5.2 Model description

Since the preprocessing does not affect the sequential order of the remaining medical concepts in the narratives, we use HMMs as the likelihood to characterize the temporal dynamic nature of the medical concept sequences. In

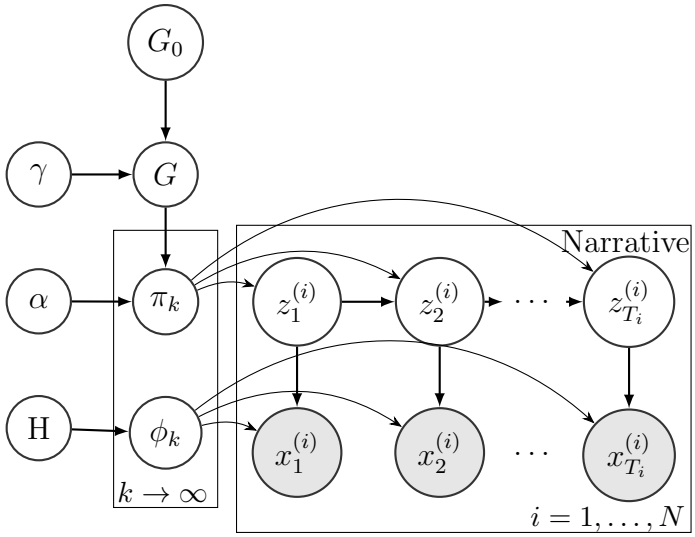


Figure 3.7: The hierarchical Dirichlet process-hidden Markov model that learns from multiple narratives as a group.

Figure 3.7, each learned hidden state sequence $\{z_t^{(i)}\}_{t=1,2,\ldots,T_i}$ presents a subset of all the hidden states that particularly corresponds to the observed medical concept sequence $\{x_t^{(i)}\}_{t=1,2,\ldots,T_i}$. We use the hierarchical Dirichlet processes proposed by Teh et al. as a prior distribution of the model to flexibly discover more hidden states as additional narratives are observed [111]. All narratives in each experiment are used to learn such a hierarchically-structured dynamical model.

We utilize the hierarchical prior in the following specification based on our problem scenario. Let G denote the global measure of an experiment (I or II), and it is distributed as $DP(\gamma, G_0)$ with G_0 the base measure and γ the concentration parameter. Each π_k is conditionally independent given G. This hierarchical construction can be formulated as,

$$G \mid G_0 \sim DP(\gamma, G_0) \tag{3.9}$$

$$\begin{aligned} \pi_k \mid G &\sim DP(\alpha, G) \\ k &= 1, 2, \ldots, \infty \end{aligned} \tag{3.10}$$

In the i^{th} narrative, each transition probability distribution $\{\pi_{z_{t-1}, z_t=k}\}_{k=1,2,\ldots,\infty}$ of the hidden Markov model at the lower level governs the transitions toward hidden states ϕ_k's.

$$z_t^{(i)} \mid z_{t-1}^{(i)}, \pi_{z_{t-1}} \sim \pi_{z_{t-1}} \tag{3.11}$$

$$x_t^{(i)} \mid z_t^{(i)}, \phi_{z_t} \sim F(\phi_{z_t}) \tag{3.12}$$

3.5.3 Inference algorithm

We use a Markov chain Monte Carlo sampler to do the posterior inference over this model. In one iteration of the sampler, each latent variable is visited and assigned a value by drawing from the distribution of that variable conditional on the assignments to all other latent variables as well as the observation. In particular, based on the sampling algorithm proposed by Van Gael et al. [110], we develop a sampling solution that uses multiple concept sequences with arbitrary lengths as observations. Specifically for each concept sequence $\{x_t^{(i)}\}_{t=1,2,\ldots,T_i}$, auxiliary variables $\{u_t^{(i)}\}$ are sampled with probability density,

$$p(u_t^{(i)}|z_{t-1}^{(i)}, z_t^{(i)}, \boldsymbol{\pi}) = \frac{\mathbb{1}(0 < u_t^{(i)} < \pi_{z_{t-1}^{(i)}, z_t^{(i)}})}{\pi_{z_{t-1}^{(i)}, z_t^{(i)}}}$$
$$\mathbb{1}(C) = \begin{cases} 1, \text{ if } C \text{ is true} \\ 0, \text{ otherwise} \end{cases} \tag{3.13}$$

where each $u_t^{(i)}$ serves as a dynamic threshold at $t^{(i)}$ to partition the probability distribution $\{\pi_{z_{t-1},k}\}_{k=1,2,\ldots,\infty}$ into a finite set of entries larger than $u_t^{(i)}$ and an infinite set smaller than $u_t^{(i)}$. Only the states k's within the finite set are considered when sampling $z_t^{(i)}$ that transits out of state $z_{t-1}^{(i)}$ during dynamic programming. This reduces the number of potential states to consider and hence makes the inference efficient.

$$\begin{aligned} & p(z_t^{(i)}|x_{1:t}^{(i)}, u_{1:t}^{(i)}, \boldsymbol{\pi}, \boldsymbol{\phi}) \propto p(z_t^{(i)}, u_t^{(i)}, x_t^{(i)}|x_{1:t-1}^{(i)}, u_{1:t-1}^{(i)}, \boldsymbol{\pi}, \boldsymbol{\phi}) \\ =& \sum_{z_{t-1}^{(i)}} p(x_t^{(i)}|z_t^{(i)}, \boldsymbol{\phi}) p(u_t^{(i)}|z_{t-1}^{(i)}, z_t^{(i)}, \boldsymbol{\pi}) p(z_t^{(i)}|z_{t-1}^{(i)}, \boldsymbol{\pi}) \\ & p(z_{t-1}^{(i)}|x_{1:t-1}^{(i)}, u_{1:t-1}^{(i)}, \boldsymbol{\pi}, \boldsymbol{\phi}), \text{and by applying Eq. (3.13), we have:} \\ =& p(x_t^{(i)}|z_t^{(i)}, \boldsymbol{\phi}) \sum_{z_{t-1}^{(i)}: \pi_{z_{t-1}^{(i)}, z_t^{(i)}} > u_t^{(i)}} p(z_{t-1}^{(i)}|x_{1:t-1}^{(i)}, u_{1:t-1}^{(i)}, \boldsymbol{\pi}, \boldsymbol{\phi}) \end{aligned} \tag{3.14}$$

Beside resampling the auxiliary variables $\{u_t^{(i)}\}$ and the state sequences $\{z_t^{(i)}\}$ in each iteration, the algorithm also resamples the shared DP base measure G, the hyper-parameters α and γ, the emission probabilities $\boldsymbol{\phi}$, and the transition probabilities $\boldsymbol{\pi}$. Specifically G is sampled according to Eq. (2.57). Summing over the infinite many states that never occur in any hidden state

sequences $\{z_t^{(i)}\}$, the conditional distribution $\pi_{k\cdot}$ given its Markov blanket $\mathbf{z}$, G, and α is

$$\begin{aligned}\pi_{k\cdot} &= (\pi_{k1} \dots \pi_{kK}, \sum_{k'=K+1}^{\infty} \pi_{kk'}) \\ &\sim \text{Dir}(\sum_i n_{k1}^{(i)} + \alpha G_1 \dots \sum_i n_{kK}^{(i)} + \alpha G_K, \alpha \sum_{k'=K+1}^{\infty} G_{k'})\end{aligned} \tag{3.15}$$

where $n_{k\kappa}^{(i)}$ denotes the transition counts in the i-th state sequence from state k to κ. Each $\phi_{k\cdot}$ depends on the state sequences $\{z_t^{(i)}\}$, the observed concept sequences $\{x_t^{(i)}\}$, and the prior distribution H, and the $\phi_{k\cdot}$'s are independent given $\mathbf{z}$, $\mathbf{x}$, and H.

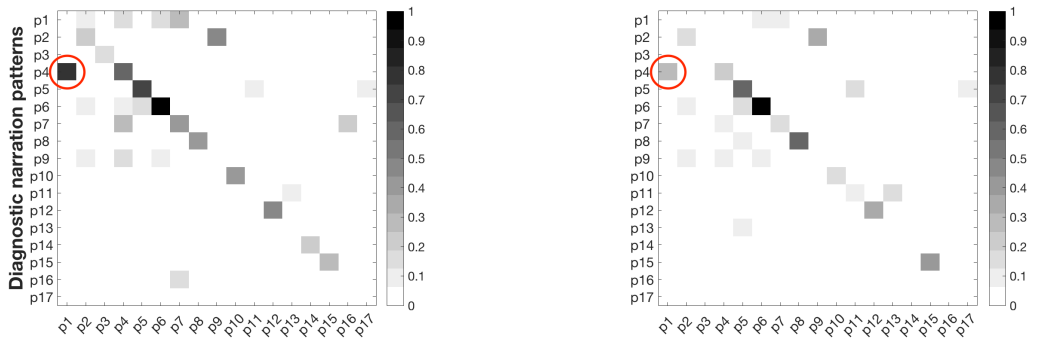
$$\phi_{k\cdot} \sim \text{Dir}(\sum_i l_{k1}^{(i)} + H_1 \dots \sum_i l_{k|V|}^{(i)} + H_{|V|}) \tag{3.16}$$

where $l_{kv}^{(i)}$ denotes the emission counts in the i-th state sequence from state k to medical concept v. The whole vocabulary set is V. We further sample the hyper-parameters α and γ according to Teh et al.'s approach [111].

In each experiment we run the sampler 20 times with random initialization of the state sequences. Each state randomly chooses between 1 and the maximum length of all sequences. We use 2000 iterations as burn-in (the beginning of the Markov chain is assumed not to accurately represent the desired distribution) and empirically choose various hyperpriors for α and γ according to the convergence behaviors in previous runs. The hidden states inferred from

the model are the diagnostic narration patterns mentioned in earlier sections, and *states* and *patterns* will be used interchangeably in the rest of this paper.

3.5.4 The discovered verbal narration patterns



(a) High-correctness group (b) Low-correctness group

Figure 3.8: Normalized state transitions in narrative groups regarding diagnostic correctness. One salient transition to discriminate both groups is from pattern 4 (the 4^{th} row) to 1 (the 1^{st} column).

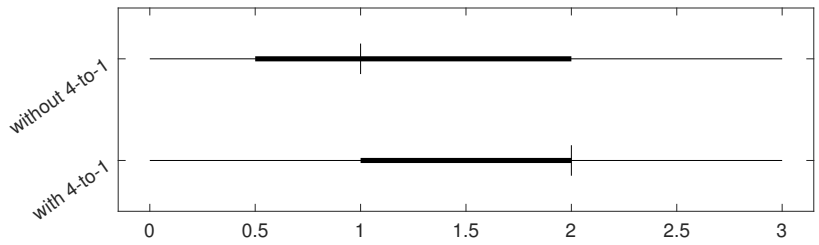
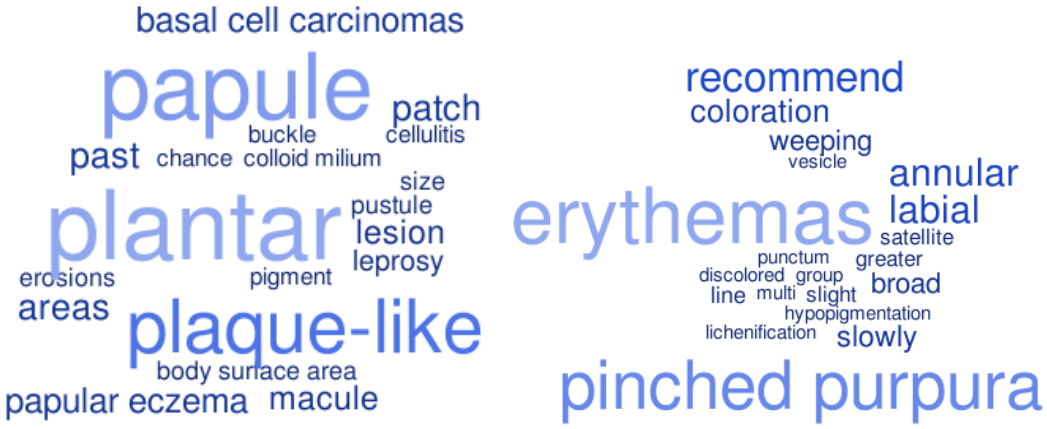


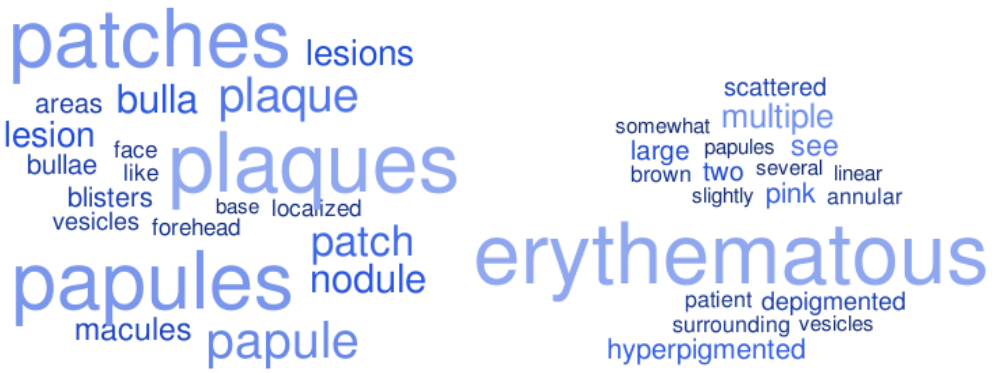
Figure 3.9: The correctness score distributions between the narratives with state transition ($4 \rightarrow 1$) and those without.

The state transition summaries for the two correctness levels in Figure 3.8 presents a salient difference (among multiple differences)—the state transition from pattern 4 to 1 (highlighted in red circles). These two patterns are in-



(a) Salient pattern 1 (PRI) in Experiment I (b) Salient pattern 4 (findings) in Experiment I

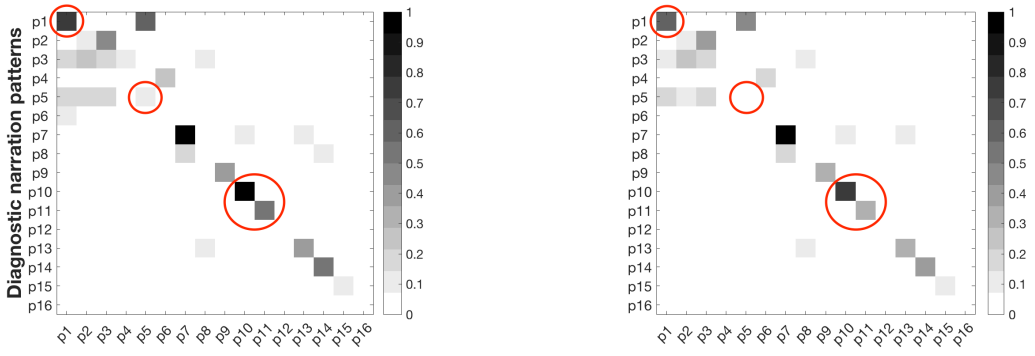
Figure 3.10: Two narration patterns learned from all narratives in Experiment I. Top 20 terms of each pattern are visualized through word cloud in which the font size indicates term frequency. Each table presents the thought unit (TU) proportion of the pattern.



(a) Pattern 5 (PRI) in Experiment II (b) Pattern 1 (findings) in Experiment II

Figure 3.11: Meaningful patterns discovered from diagnostic confidence study in Experiment II.

terpretable, and their medical concept distributions are shown in Figure 3.10. Pattern 1 can be interpreted as *primary lesion type* (PRI), and pattern 4 includes informative findings regarding color, size, shape and texture of the lesion



(a) High-confidence group

(b) Low-confidence group

Figure 3.12: Normalized state transitions in narrative groups regarding diagnostic confidence. Group (a) possesses slightly more self-transitions of 1, 5, 10 and 11 than group (b).

(a) Pattern 10 in Experiment II (None: 100%)

(b) Pattern 11 in Experiment II (DX:52%; None:38%; SEC:10%)

Figure 3.13: Two narration patterns learned from all narratives in Experiment II. Word cloud shows top terms of each pattern.

to assist determining the primary lesion type. Since the thought unit labels do not cover the whole vocabulary, there are *None* labels in both patterns. Given the meanings of these patterns, we find that the high-correctness narratives possess more transitions from describing supportive findings to mentioning

primary lesion type than the low-correctness narratives. We also consider all the narratives in Experiment I, and separately visualize the ones with and without state transition $4 \rightarrow 1$ in Figure 3.9. We notice two different distributions of correctness scores, and we find that the narratives with this key state transition generally shift towards the higher correctness end. This implies the importance of locating key clues before determining a primary lesion type in order to make a correct diagnosis. Similar patterns are discovered in parallel from the diagnostic confidence study in Experiment II (see Figure 3.11). Since these patterns appear in both experiments and match the thought units, we recognize them as *Signature Patterns*. Diagnostic confidence study presents similar state transitions between low and high confidence levels (Figure 3.12). The slight differences involve more self-transitions of patterns 1, 5, 10 and 11 in the high-confidence group. Patterns 1 and 5 presents primary lesion type and informative findings (Figure 3.11). Pattern 10 can be interpreted as *confidence marker*, and pattern 11 as *diagnosis* or *differential diagnoses* (Figure 3.13). The confidence markers in pattern 10 are *Quantitative* or *Qualitative* concepts in the UMLS.

3.5.5 Narrative correctness classification

We classify the narratives at low and high correctness levels based on various feature combinations with two classifiers, and Table 3.5 summarizes the classification performances. We use cross-validation to tune the trade-off parameter of the lasso-regularized logistic regression. Cross-validation is also used to de-

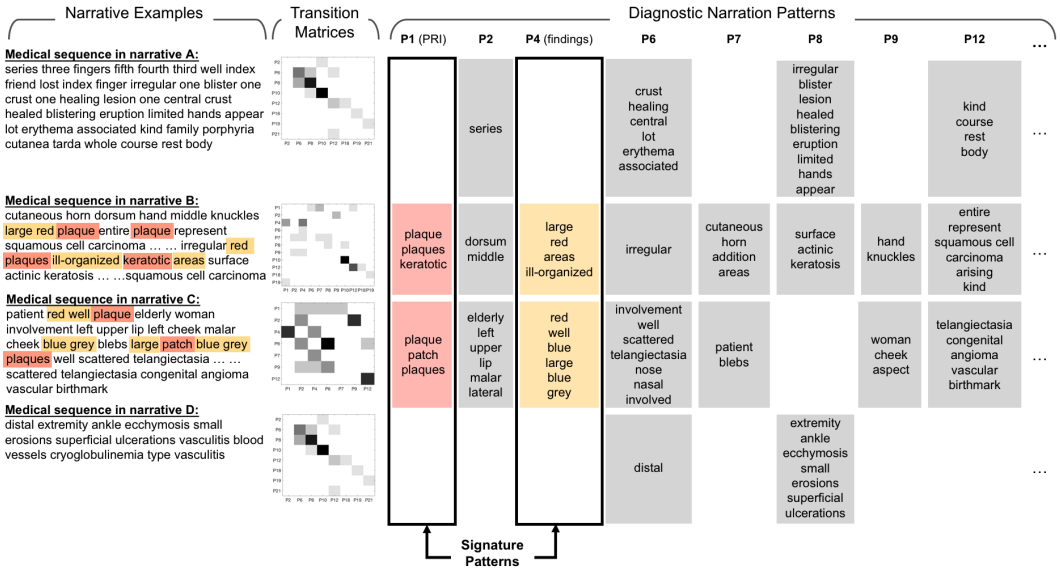


Figure 3.14: Example narratives in the diagnostic correctness study. *Left*: The remaining medical concept sequences (after preprocessing) in four narrative examples. *Middle*: The corresponding transition probability matrices out of the overall 17 patterns discovered from all narratives in Experiment I. *Right*: The shared narration pattern matrix. The two *Signature Patterns* are highlighted.

termine the optimal number of hidden states for the canonical HMM. We find that the infinite HMM works better than the canonical HMM to capture the important temporal patterns for classification. Concatenating all features boosts the classification performance, because the LDA and iHMM capture high-level information (topic-level) complementary to the fine details in tf-idf (word-level). Both classifiers suggest high importance of the patterns learned from iHMM, and Table 3.6 lists highest-ranked features and their interpretations.

In Figure 3.14, we analyze both high-correctness narratives (B and D)

Feature \ Classifier	Regularized Logistic Regression		Random Forest	
	Accuracy (%)	AUC of ROC	Accuracy (%)	AUC of ROC
tf-idf	61.8	0.63	44.9	0.65
LDA	64.0	0.63	62.9	0.67
HMM	59.6	0.62	56.2	0.58
iHMM	64.0	0.68	65.2	0.64
tf-idf + iHMM	**67.4**	0.69	**75.3**	**0.78**
tf-idf + LDA + iHMM	**67.4**	**0.71**	**75.3**	**0.78**

Table 3.5: Narrative correctness classification performances. The positive class for ROC is high-correctness.

and low ones (A and C). The narratives (A) and (B) are successfully classified. Narrative (C) is misclassified as high-correctness, because it mentions the correct primary lesion type but fails to give a correct diagnosis nor differential diagnoses. Narrative (D) is misclassified as low-correctness, because it only makes a correct diagnosis without mentioning the primary lesion type. These examples show that the features that dominate the classifier are the signature patterns (PRI and findings).

3.6 Conclusions

This chapter explores the benefits of analyzing experts' spoken inputs. Section 3.2 extracts medical terms from the diagnostic spoken narratives using

Rank	Feature (Feature interpretations in detail)
1	Pattern 4 in iHMM (erythemas, pinch purpura, annulare, ...)
2	Topic 45 in LDA (hand, dorsal, hyperkeratotic, ...)
3	Term 32 in tf-idf (papules)
...	...

Table 3.6: Ranked features by random forest classifier.

the external knowledge tool the UMLS. This processing allows analysis and modeling to work at the domain-specific semantic level. Section 3.3 clusters narratives using anchor concepts based on use of medical terms by physicians in the diagnostic verbal narratives. It differentiates the images used in our experiment by term usages. The representation learning algorithms provide a new perspective on how to group medical images. Given the usual case that only a relatively small set of annotated data/evidence is accessible compared against the big biomedical data pool, it is a natural choice to use prior knowledge, especially that derived from domain knowledge.

Section 3.4 explores the additional values of physicians' spoken narratives by using the lexical consensus metric at the token level and the top N relatedness metric of terms' relations to the correct diagnoses of images at the concept level. Experimental results show that if the use of medical terms by all experts is in agreement, the more likely the image was an easy case. In addition, the highest relatedness score computed from the most related medical concept uttered by physicians could help indicate the challenge levels of an image case. These lexical metrics can be used to develop constraints for narrative modeling.

Section 3.5 explores diagnostic decision-making by modeling physicians' utterances of medical concepts during image-based diagnoses. We develop automated approaches to discover diagnostic narration patterns from expert data. The model discovers patterns that exist in both datasets and match the expert-defined diagnostic thought units. These patterns are also impor-

tant features for diagnostic correctness classification. Since the concepts in the same narrative are essentially correlated, we plan to relax the strong Markovian assumption in the model by considering semantic relatedness of medical concepts [139, 146] and developing an HMM variant in the future. We will further explore medical image difficulty levels [37] and physicians' expertise levels [147] as they are relevant factors in diagnostic decision making and error prevention.

Chapter 4

Multimodal Data Fusion

Cognitive processing of visual information in complex domains involves multiple levels of abstraction. It is accomplished during an interaction of domain knowledge, perceptual expertise, reasoning strategies [45], and idiosyncratic visual information in the image case being inspected. Therefore, tracing back and isolating the underlying human knowledge used during image-based diagnostic reasoning is difficult.

Our datasets contain multimodal expert data elicited during in-scenario experiments, and they are critical to analyze and represent the cognitive processes. This chapter addresses the fusion of the multimodal expert data, which integrates the respective strengths of the data modalities to model human knowledge and expertise. It first reviews existing approaches to incorporating multiple data modalities in various research domains and for different purposes. And then, we present a preliminary analysis of the underlying components in eye fixations by fitting a Gaussian mixture model (GMM) (Section 4.2). The findings of this analysis encourage us to use eye fixations as a mask to highlight image features of users' interests in an information retrieval application (Section 4.3). We also combine eye tracking and verbal modalities through a

data fusion framework to discover a unified data representation in Section 4.4.

4.1 Background

Biomedical research is often accompanied by the phenomenon of *drowning in data while starving for knowledge* [148]. Plenty of data are produced daily, and various data modalities provide different information. For example, cardiologists make comprehensive diagnoses by inspecting patients' data from multiple resources, including stethoscope, cardiac computed tomography (CT), and magnetic resonance imaging (MRI). Despite that their diagnostic reasoning processes through the multimodal data are unknown to the computing researchers, it has been confirmed that combining such multimodal data produces greater findings than the sum of parts [149]. For this reason, it has become a trend in biomedical informatics to combine multiple data modalities in a number of applications, such as image classification and retrieval, knowledge discovery, and medical education. The data fusion approaches discover key information for effective feature extraction, data classification, and decision making.

Existing data fusion approaches can be categorized based on their purposes, i.e., data fusion for denoising and enhancement of the signal, and fusion for detection and classification [150].

- To de-noise and enhance signals, additional data modalities can be used for verification purposes, and some spatial-temporal or spatial-spectral

fusion algorithms can be applied. For example, in Putze et al.'s study to locate user attention, eye tracking information is combined with EEG to form fixation-related potentials (FRPs) that can select image regions based on spatial-temporal alignment [151]. Similarly, Dimigen et al. studies natural sentence reading by the co-registration of eye movements and EEG signals [152]. In Kaplan et al.'s study to enhance multispectral imagery, a fusion technique is developed [153].

- To improve detection and classification, data fusion approaches based on various machine learning techniques can be developed, including the conventional supervised learning, deep learning, active learning, and semi-supervised learning. For example, Zhou et al. improves target detection in remote sensing images by selecting more informative negative instances for classification [154]. For logo image retieval, Yan et al. developed a framework to fuse color and spatial descriptors adaptive to image quality levels [155]. To detect copy-move image forgery, Zheng et al. fused block-based and keypoint-based approaches, which overcomes the drawbacks of both [156]. Ren et al. developed a multi-view and multi-plane data fusion approach for pedestrian detection in intelligent visual surveillance [157].

Depending on the relationship between the multiple modalities, data fusion approaches can also be categorized as *exclusive*, *alternative*, *concurrent*, and *synergistic* [158].

Data fusion can be achieved through multimodal representation learning. The general idea is to learn a shared knowledge representation and construct the semantic correspondence between modalities through a shared representation. An intuitive approach is to apply matrix factorization technique for the case of multiple modalities, where a single objective function can be developed to optimize each data modality [159]. A link between data modalities is constructed through a newly-learned latent space that represents latent semantics. This achieves the feature-level data fusion, which provides a more systematic framework than decision-stage data integration [160] since the latter involves tuning the weights of decisions from different modalities. Moreover, various regularization terms (in Section 2.3.2) can be directly added in the objective function to achieve respective learning goals. Existing online update rules and visualization techniques for NMF can also be easily used for developing interactive frameworks [104, 62] (See Section 2.3.2 and Chapter 5). We have developed a data fusion algorithm based on matrix factorization to integrate multimodal data at an early stage (feature-level fusion) for our data (see Section 4.4). Besides the NMF-based learning framework, other representation learning approaches could also be extended for data fusion [161], which can also incorporate sparsity constraints [162, 163].

Deep (hierarchical) models are a collection of models capable of extracting multilevel meaningful features from data. They have been successfully applied in many application areas, including visual object recognition [164], information retrieval [165], and automatic speech recognition (ASR) [166]. Deep

learning based systems have won a number of high-profile competitions and have gained great popularity [167]. A few notable examples of deep models include deep belief networks [90], deep Boltzmann machines [168, 169], deep auto-encoders [170], and deep sparse coding [171]. Most of these architectures are constructed by a stack of feature extractors, such as RBM and auto-encoder that are flexible to structurally incorporate prior knowledge [172, 173]. Each layer in the architecture encodes features at different level of abstraction, defined as a composition of lower-level features.

As opposed to shallow architectures which are incapable of extracting some types of structure from high-dimensional input, deep models can extract complex statistical dependencies from data and efficiently learn high-level representations by re-using and combining intermediate concepts. Montavon et al. reviews and analyzes deep networks through kernel analysis [174]. There are also systematic analyses of the relationships between deep network and other basic algorithms, such as sparse coding-based algorithms [171, 65].

4.2 Mixture Components in Eye Fixations

Eye movement patterns reflect viewers' perception of the image content. For example, Figure 4.1-(a) shows a dermatology image presenting symmetrically-distributed lesions on a patient's cheeks and nose. According to Li et al.'s modeling of physicians' eye movements, the symmetric lesions in this image initiate physicians' symmetry viewing pattern [15]. In this section, we use multivariate Gaussian mixture model (GMM) with variables being fixation loca-

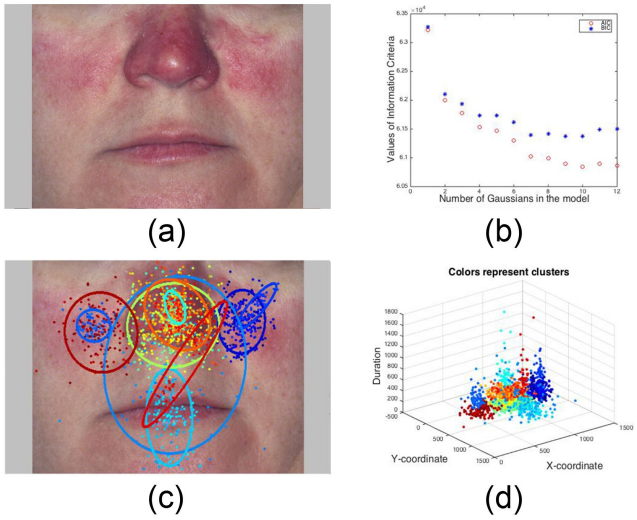


Figure 4.1: A symmetrical viewing pattern detected by GMM in eye fixations.

tion (i.e., x- and y-coordinates) and duration to mark up the expert-identified informative image regions (i.e., the Gaussian mixture components).

Similarly, in Figure 4.2 we present an image with multiple solitary lesions that trigger physicians solitary viewing patterns. In both figures, subfigure (a) shows the original image viewed by all participating physicians. Subfigure (b) shows the model selection using Akaike information criterion (AIC) as red circles and Bayesian information criterion (BIC) as blue stars. Subfigure (c) visualizes the Gaussian mixture featured by fixation duration and x-, y- coordinates overlaying the original image, and (d) visualizes these features in 3D. Colors indicate different components. Since we analyze all physicians' eye fixations of the same image, the discovered components capture their commonality (rather than variance) in eye movement behaviors which reflects the consensus understanding of the image content.

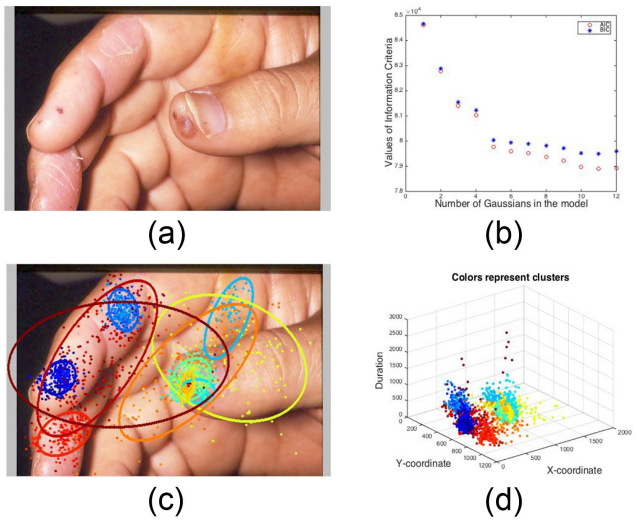


Figure 4.2: A solitary viewing pattern detected by GMM in eye fixations.

4.3 Human-centered Information Retrieval

In medical domains, highly specialized expert knowledge is needed for understanding image semantics. To advance content-based image retrieval to human-centered image retrieval (HCIR), we combine image content and human image understanding in the loop of information retrieval. In particular, we implement a prototype image retrieval system that stores physicians' verbal and gaze data and allows an end user to form a query based on verbal or gaze input. These two interaction styles initiate the retrieval processes that involve conceptual and perceptual knowledge matching respectively.

4.3.1 HCIR System design

To represent expert understanding of the candidate images in the dataset, we preprocess and store physicians' eye movements and verbal data as meta

knowledge in the system. In particular, we first use the scale-invariant feature transform (SIFT) [175] to detect local key points that highlight local covariant structures such as corners, edges and small patches in each image, and to calculate corresponding feature descriptors based on neighboring pixels' intensity gradients. Since eye fixations inform feature importance in images [6], we then construct experts' *intersection eye fixation maps* for each candidate image. These maps identify the visual content viewed the most by experts, and hence evaluate the significance of features in each image from domain experts' point of view. Since the medical terms in physicians' verbal data convey insights about their conceptual knowledge for each image case, we process all narratives and construct a medical term vector for each image.

While interacting with the system, an end user is able to: (1) select an image of interest for which she wants to retrieve similar images, then (2) inspect this image, and (3) describe it in realtime (see Figure 4.3). The system captures the user's eye movements and keyboard input to form a query. In a retrieval process, her eye movements are used for highlighting the image features in the query image, which then are matched to the important image features identified by experts' *intersection eye fixation maps* in each candidate image. Her verbal inputs regarding the query image are used to measure the semantic similarity to the medical terms associated with each candidate image. In this manner, we also cast the image matching problem as document retrieval (see Figure 4.4).

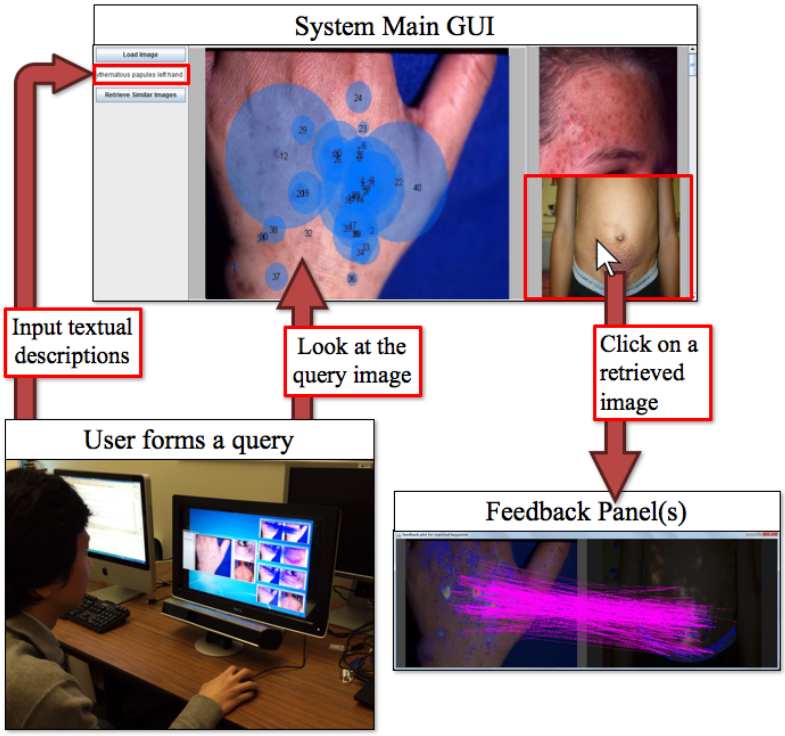


Figure 4.3: An overview of system design (user view). Human knowledge overtly expressed in verbal and eye movement data from both physicians (collected in data-elicitation experiments described in Section 1.1) and an end user (during interaction with the system) are used for image retrieval.

4.3.2 Eye tracking-based retrieval

In particular, a program was developed using JavaCV to detect the SIFT keypoints and the corresponding descriptors of each candidate image [176]. During the retrieval process, the extracted SIFT keypoints are filtered by the user's realtime eye fixations. The highlighted keypoints are used to match those from candidate images filtered by physicians' intersection fixation map.

Physicians are knowledgeable of where to look and what to look for

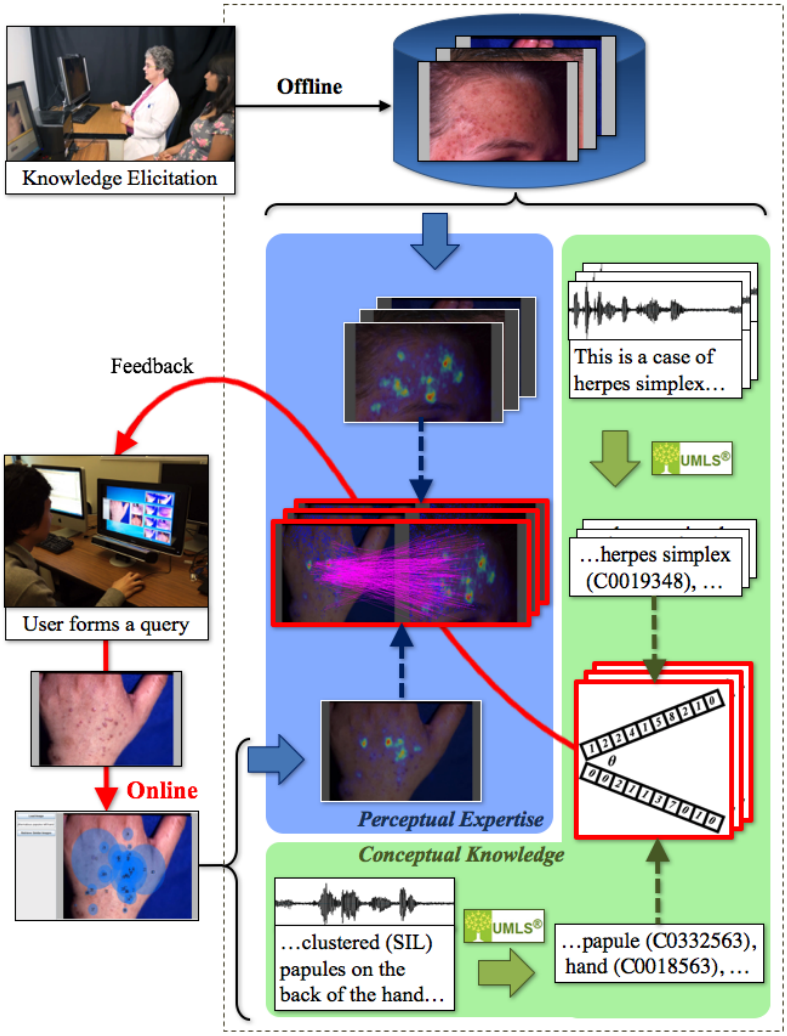


Figure 4.4: An overview of system design (system view). The verbal and eye movement data are processed and used for similarity computations.

in the images [13, 43]. In order to expose the medically-important visual features in the candidate images to the retrieval algorithm, for each image we compute an intersection fixation map, which is a normalized mixture of physicians' fixations smoothed by Gaussian kernel. According to our data-

elicitation experiments, the horizontal radius is empirically chosen to be 2 visual degrees, and the vertical radius is 3.

Both physicians' intersection fixation maps corresponding to the candidate images and the end user's fixation map for the query image are used in the following manner to filter the keypoints in images. Let (μ_i, d_i) denote one fixation from the eye movements of a human viewing an image. In particular, μ_i represents the x-y coordinate of each fixation, and d_i represents the corresponding duration time. To use the user's eye movement fixations to evaluate the significance of SIFT feature points $\{f_1, f_2, \ldots, f_N\}$ where f_n denotes the location of the feature point n, we define the perceptual importance of SIFT feature point n as

$$S(f_n) = \alpha_k \sum_i d_i P(f_n; \mu_i, \sigma^2) \tag{4.1}$$

where $P(f_n; \mu_i, \sigma^2)$ represents a Gaussian distribution centered at fixation i with σ^2 corresponding to the visual angle of the foveal field. Since the visual acuity attenuates dramatically from the center of focus [177], a light-tailed Gaussian distribution is a natural choice to model the foveal field. The value of this distribution at the feature point n denotes how likely the feature point is targeted by fixation i, and α_k is a normalization constant.

For each (*query, candidate*) pair, we generate pairs of matched keypoints by brute force matching: for each SIFT keypoint n in the query image, we find its best matched SIFT keypoint n' (its closest neighbor in feature

space) in the candidate image, and the distance between these two keypoints is recorded as $D(n, n')$. We use the perceptual importances of keypoints in both images to weight the feature distances and thus define the highlighted visual similarity score V as

$$V(q, c) = \sum_{n} \frac{S_q(f_n) S_c(f_{n'})}{D(n, n')} \tag{4.2}$$

where S_q denotes the importance of a keypoint n in the query image q, and S_c denotes that of the corresponding keypoints n' in the candidate image c. Since the distance $D(n, n')$ between two image features could be zero, $D(n, n')$ is smoothed for all n's.

Eq. (4.1) shows that the keypoints are weighted by their spatial distance to the center of Gaussian. The longer the distance, the less the importance. The keypoint similarity $1/D(n, n')$ in feature space is weighted by the perceptual importances of both keypoints $S_q(f_n)$ and $S_c(f_{n'})$, and the summation over all keypoints defines the visual similarity score for ranking the candidate images.

Experimental results show the importance of eye gaze features, and these features are used in Section 4.4 as one of the data modalities to learn a unified data representation to capture high level of meanings that better describe and cluster the images.

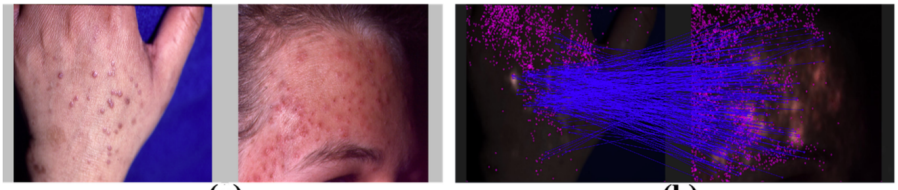


Figure 4.5: Using the end user's and physicians' eye movements as filters for visual features to retrieve images. Panel (a) is the original images (left query, and right one of the candidates) to be filtered in (b). Panel (b) shows the filtering processes: The magenta points are SIFT keypoints detected by JavaCV program. The blue lines indicate feature matching between two key points across the two images, and the two ends of the blue lines are the remaining keypoints after eye movement filtering. (Images used with the permission of Logical Images, Inc.)

4.3.3 Verbal input-based retrieval

Physicians' verbal descriptions of candidate images contain the key domain concepts involving image understanding. Therefore, we incorporate textual input from an end user to select physicians' descriptions of images in database for retrieval. In order to facilitate image retrieval by text input, we formulate a medical term vector for each query and each candidate image. The cosine similarity [134] for each (*query, candidate*) pair is computed and used to score the candidate image.

Since an end user's text input may only contain a few terms, we apply query expansion to avoid a too sparse query vector. We extract all the medical concepts in our datasets and compute pairwise relatedness scores using the UMLS::Similarity package [135]. These relatedness scores are stored in the system as a lookup table, in case a user's input may need expansion.

Only the stored concepts that are highly related (scoring above an empirically determined threshold of 0.85 in the score range of [0, 1]) to the user-input concept are selected to be query-expanded concepts. All original user-input concepts, as well as the query-expanded concepts, are then used for forming a query and computing the concept similarity. The original concepts input by the end user are given a higher weight than the query-expanded ones, so as to tune out the potential noise introduced by the query expansion.

4.3.4 Retrieval performance evaluation

Since lesion morphology (i.e., lesion type) is crucial for generating differential diagnoses and grouping dermatology images in the medical sense [178], we evaluate our prototype system at this level. We apply three retrieval strategies for evaluation, namely image-based (SIFT) features alone, image-based features highlighted by eye movement data, and verbal data alone, to compute a visual score, a highlighted visual score, and a verbal score, respectively, for each image in a test database, in order to evaluate the effectiveness of each modality for image retrieval. Given image retrieval queries, image candidates are then ranked using one of these scores, and the top images are returned as results to be evaluated. Ranked lists of retrieved images based on eye movement- and verbal data-based scores can serve as a benchmark for our full system that combines these modalities.

For evaluating the performance of our prototype system, we use a subset of the physicians' data from our data-elicitation experiment. Within the

subset, each individual physician's eye movements and verbal description can be used as a simulated human input, and play the role of an end user. This evaluation approach eliminates our need to recruit a medical user in the testing phase.

We randomly select data recorded for 4 physicians for the images in our database to generate test cases. The data consist of eye movements and verbal descriptions recorded when these physicians inspected 48 images in our data collection, resulting in data for 192 individual image inspection cases. After discarding image inspection cases for which physicians gave incorrect diagnoses, we have 131 cases. In each test case, we label each image in the database by lesion morphology as ground truth. All labels are determined by our collaborating dermatologist who has access to the metadata of the images for our data-elicitation experiment. To eliminate confusions, test cases, in which the query images present more than one morphology, are not evaluated. We use the image inspected by the physician as the *query image*, and the remaining collection of 47 images as *candidate images*; the physician's verbal descriptions are used for verbal data-based retrieval, and his/her eye movements are used for filtering visual features. The 4 chosen physicians' data are not used for generating concept vectors or intersection maps of the candidate images. For each test case, ranked lists of retrieved images based on different retrieval strategies are generated and evaluated at the level of lesion morphology.

In order to evaluate the retrieved images at the level of dermatology

lesion morphology, we count how many retrieved images share the same dermatology lesion morphology with the query image. To stress the usefulness of experts' eye movement and verbal data for image retrieval at this domain-related semantic level, we perform tests that do not involve experts' data, and compare the results to test cases based on 4 chosen physicians' data. In order to do so, we customize the definitions of *precision* and *recall* [179] for our scenario. The precision is defined as the ratio of the number of retrieved images of the same morphology as the query to the total number of images retrieved, and the recall is defined as the ratio of the number of retrieved images of the same morphology as the query to the total number of images in database that are of that morphology. The summary of retrieval performances is displayed in Table 4.1. Retrieval performance benefits both from using experts' eye fixations as a filter and using their verbal data, compared to whole-image visual features alone.

As a further test for the usefulness of eye fixations as a filter for visual features, we add 26 images to the database, while the test cases, each containing a query image and a physician's fixation map, remain the same as in Table 4.1. The intersection fixation maps of the added images are generated from 30 physicians' fixations. Since extra images are added in the database for retrieval, in each test case we retrieve top 20 similar images. With additional images in database, retrieval performance still benefits from using experts' eye fixations as a filter. Our second evaluation procedure leaves the consideration of verbal descriptions for future work.

Table 4.1: Precision (P) and Recall (R) comparison at lesion morphology level. Among 48 images in the database, there are 9 images considered as containing the morphology *macule*, 38 *papule*, 5 *bulla*, 4 *pustule*, and 1 *nodule*. Because of visual similarity, *patch* is categorized in *macule*, *vesicle* is in *bulla*, and *plaque* in *papule*. The number of test cases, in which the query image belongs to the morphology, is marked in parenthesis in each row, as well as the number of images containing this primary morphology that can be retrieved from in database. In each test case, 47 images out of 48 in the database that are not the query image can be retrieved. Top 10 similar images among the 47 are retrieved. The performances of three strategies are all listed by the primary morphology of the query image. Overall performances are listed in the last row.

Morph. (test cases, images in DB)	Visual		Visual w/ Filter		Verbal	
	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)
Macule (20, 9)	12	13.3	18	20	19.3	21.4
Papule (89, 38)	78.6	20.7	97.4	25.6	87.1	22.9
Bulla (4, 5)	15	30	10	20	20	40
Pustule (2, 4)	10	25	10	25	10	25
Nodule (3, 1)	10	100	10	100	0	0
Overall (118, 48)	**62.3**	**21.9**	**77.3**	**26.3**	**69.8**	**22.7**

Based on the results in Table 4.1 and 4.2, visual features combined with experts' eye movement filters improve retrieval performance. This suggests that human eye movements improve retrieval performance in at least two aspects. First, experts' eye movements filter out irrelevant features, such as the ones from the background (normal skins, clothes, etc.). This helps bypass the manual segmentation of the medical images and also allows us to study experts' subconscious image viewing behaviors. Second, they also highlighted diagnostic valuable features from the lesions. Besides, retrieval results are also improved based on experts' verbal descriptions, which suggests that experts' conceptual knowledge are also valuable for image retrieval at the semantic

Table 4.2: Precision (P) and Recall (R) comparison at lesion morphology level for the additional test that involve more images in the database.

Morph. (test cases, images in DB)	Visual		Visual w/ Filter	
	P (%)	R (%)	P (%)	R (%)
Macule (20, 15)	14.5	19.3	20	26.7
Papule (89, 56)	60.3	21.6	75.6	27
Bulla (4, 9)	5	11.1	21.3	47.2
Pustule (2, 5)	2.5	10	5	20
Nodule (3, 4)	0	0	5	25
Overall (118, 74)	**48.2**	**20.1**	**61.3**	**27.5**

level. This multimodal user interaction provides a natural way for the user to interact and retrieve images. To conclude, this human-centered approach initiates natural user interactions during information retrieval, and it lays a foundation for future systems that may advance the semantic similarity computations.

4.4 Multimodal Data Fusion

Image-based diagnostic reasoning involves multiple cognitive processes, including visual pattern identification and medical language expression. Since the heterogeneous data streams offer different levels of thoughts and decisions [180], our studies in previous sections tend to consider the data modalities separately. For example, image classification and retrieval can be achieved by using either gaze-weighted image features or verbal inputs [181] (Section 4.3). Chapter 3 and Section 4.2 model verbal data or eye movement data alone. Similarly, in other studies gaze patterns and speech features were separately used to measure conversational activities [182]. In this section, we address

the fusion of multimodal data by projecting them into a unified latent space. The aim here is to achieve a data representation that uncovers the hidden semantics to explain the observed data of both modalities.

As stated in Section 2.3, NMF is a latent factor-based model that finds the hidden factors and expresses the observed data using them. Caicedo and Fabio have extended it to the multimodal case to discover the latent components that capture the underlying meanings of images [159]. The resulting unified representation of the multimodal data is flexible and robust to address the limitations caused by the unavailability of one or more data modalities. Inspired by their study, we develop an approach to data fusion by extending the Laplacian sparse coding to the multimodal case. In particular, we enforce sparsity of the unified representation. To ensure the feature learning stability during space change, a graph-based regularizer is also added [5]. To solve the newly-defined multimodal version of Laplacian sparse coding problem, we also extend the feature-sign search algorithm to a multimodal version.

In particular, we used the multimodal expert data from Experiment I. The eye movement data or verbal descriptions of one physician inspecting a single image is defined as a *trial* and will be referred to in the remainder of this section. We have 48 (images) $\times$ 16 (physicians) = 768 eye movement trials and 768 verbal description trials, which form 768 data instances. As a quantitative measure of visual content perception, eye tracking features, such as *number of fixations*, *fixation duration*, and *pupil size*, are known to indicate viewers' interests [183, 50, 49, 12, 6]. To capture physicians' viewing

behaviors in different image regions, we create $21 \times 14 = 294$ grids. We count and average these 3 eye tracking features within each grid segment, and create an 882(features)-by-768(trials) matrix. Despite the over-simplicity, we use the grid-based features in this study for a proof of concept of the data fusion approach. Since experts' expressed language during image inspection provides additional information regarding the conceptual content adopted for diagnoses, we analyze the conceptual domain knowledge through physicians' transcribed diagnostic spoken narratives. The concepts in narratives are used to construct a feature space with length 1759. 768 verbal records are preprocessed to form 768 feature vectors, each of which representing the occurrences of these 1759 medical concepts. A 1759(features)-by-768(trials) matrix is thus created as physicians' conceptual image interpretations.

4.4.1 Gold standard

Similar to Section 3.3.1, we use the 48 dermatology image labels as gold standard of the narrative clusters, as we have a wide range of dermatology diagnoses in Experiment I.

4.4.2 Multimodal data fusion framework

We extend the Laplacian sparse coding [5] to uncover latent semantics that explain both eye movement data $\boldsymbol{E} \in \mathbb{R}^{n_e \times m}$ and verbal data $\boldsymbol{V} \in \mathbb{R}^{n_v \times m}$. The objective is formulated as,

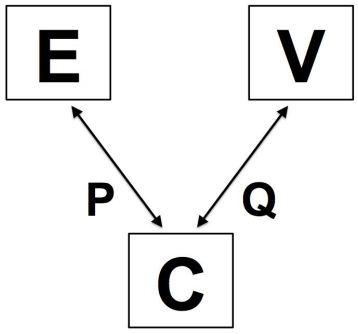


Figure 4.6: Fusing multiple data modalities. Coefficient matrix, $\boldsymbol{C}$, and basis matrices, $\boldsymbol{P}$ and $\boldsymbol{Q}$, are learned from an eye tracking data matrix, $\boldsymbol{E}$, and a verbal description data matrix, $\boldsymbol{V}$.

$$\min_{\boldsymbol{P},\boldsymbol{Q},\boldsymbol{C}\geq 0} \|\boldsymbol{E}-\boldsymbol{P}\boldsymbol{C}\|_F^2 + \|\boldsymbol{V}-\boldsymbol{Q}\boldsymbol{C}\|_F^2 + \alpha Tr(\boldsymbol{C}\boldsymbol{L}\boldsymbol{C}^\top) + \beta \sum_{j=1}^{m} \|\mathbf{c}_j\|_1 \\ s.t. \|\mathbf{p}_i\|^2 \leq a, \|\mathbf{q}_i\|^2 \leq a, i = 1, ..., k \tag{4.3}$$

where $\boldsymbol{E} \in \mathbb{R}^{n_e \times m}$ and $\boldsymbol{V} \in \mathbb{R}^{n_v \times m}$ stand for two different data modalities, eye tracking data and verbal data. $\boldsymbol{P} \in \mathbb{R}^{n_e \times k}$ and $\boldsymbol{Q} \in \mathbb{R}^{n_v \times k}$ are the corresponding basis matrices. Let $\mathbf{p}_i$ and $\mathbf{q}_i, i = 1, ..., k$ denote eye tracking basis vectors and verbal description basis vectors, respectively. Essentially, $\mathbf{p}_i$'s are expected to capture visual patterns from the eye movements, whereas $\mathbf{q}_i$'s are expected to capture high-level medical concepts of the spoken narratives. The norm constraints on the size of the basis vectors, i.e., $\|\mathbf{p}_i\|^2 \leq a$ and $\|\mathbf{q}_i\|^2 \leq a$, avoid arbitrarily large basis vectors that keep $\boldsymbol{P}\boldsymbol{C}$ and $\boldsymbol{Q}\boldsymbol{C}$ unchanged while making $\mathbf{c}_j$ arbitrarily close to zero. The $\|\cdot\|$ is the vector l_2-norm and a is a positive constant number. The coefficient matrix, $\boldsymbol{C} \in \mathbb{R}^{k \times m}$, is a new representation of the original data. In our experiment, $n_e = 882$ eye tracking

features, $n_v = 1759$ verbal features, and $m = 768$ data instances. We update the basis matrices using Lagrange dual and update the coefficient matrix by deriving a multimodal feature-sign search algorithm (see Section 4.4.3).

The α and β are the weights of graph-regularizer and l_1-regularizer, respectively. They can be set by tuning our model for an (α, β) pair that gives a minimum of objective function. Graph-regularizer was introduced by a weighted graph of the data points represented in the input matrices [5]. Let those data points be denoted as $\mathbf{x}_1$, ..., $\mathbf{x}_m$, each with the length of n. A nearest neighbor graph G with m vertices can be constructed. The element $\boldsymbol{W}_{ij}$ in the weight matrix $\boldsymbol{W}$ of the graph G equals 1, if $\mathbf{x}_i$ is among the k-nearest neighbors of $\mathbf{x}_j$ or $\mathbf{x}_j$ is among those of $\mathbf{x}_i$; otherwise, $\boldsymbol{W}_{ij}$ equals 0. The degree of $\mathbf{x}_i$ is defined as $d_i = \sum_{j=1}^{m} \boldsymbol{W}_{ij}$, and $\boldsymbol{D} = \text{diag}(d_1, ..., d_m)$. $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$, is a Laplacian matrix used to minimize the Laplacian item $Tr(\boldsymbol{CLC}^\top)$ in the objective function. This item ensures that the nearest neighboring data instances in the input matrix $\boldsymbol{X}$ are still close in the factorized matrix representation. This factorization process removes noise in data and keeps useful information through optimization.

4.4.3 Algorithm to solve the multimodal GrNMF

We extend the feature-sign search algorithm for the multimodal data fusion framework, and the extension algorithm is shown in Alg. 10. Similar to Alg. 6, Alg. 10 is also used to selectively active dimensions indexed by j in each data point $\mathbf{c}_i$.

Algorithm 10 The developed feature-sign search algorithm for multimodal GrNMF.

1. Initialize $\mathbf{c}_i = \mathbf{0}$, and active set $A = \emptyset$.
2. From zero coefficients of $\mathbf{c}_i$, select $j = \arg\max_j \left| \frac{\partial\left(\mathcal{R}_E^{(j)}(\mathbf{c}_i)+\mathcal{R}_V^{(j)}(\mathbf{c}_i)+\mathcal{G}^{(j)}(\mathbf{c}_i)\right)}{\partial \mathbf{c}_i^{(j)}} \right|$.

   Activate $\mathbf{c}_i^{(j)}$ by adding j to set A only if it improves the objective, namely:

   If $\frac{\partial\left(\mathcal{R}_E^{(j)}(\mathbf{c}_i)+\mathcal{R}_V^{(j)}(\mathbf{c}_i)+\mathcal{G}^{(j)}(\mathbf{c}_i)\right)}{\partial \mathbf{c}_i^{(j)}} < -\beta$, then $A = A \cup \{j\}$.
3. Feature-sign step:

   Let $\hat{\boldsymbol{P}}$, $\hat{\boldsymbol{Q}}$ be the submatrices of $\boldsymbol{P}$, $\boldsymbol{Q}$ with only columns regarding A.

   Let $\hat{\mathbf{c}}_i$ denote the subvector of $\mathbf{c}_i$ with only dimensions regarding A.

   The solution to $\min_{\hat{\mathbf{c}}_i} \mathcal{R}_E(\hat{\mathbf{c}}_i) + \mathcal{R}_V(\hat{\mathbf{c}}_i) + \mathcal{G}(\hat{\mathbf{c}}_i) + \beta\hat{\mathbf{c}}_i$ can be derived as $\hat{\mathbf{c}}_i^{new} = (\hat{\boldsymbol{P}}^\top\hat{\boldsymbol{P}} + \hat{\boldsymbol{Q}}^\top\hat{\boldsymbol{Q}} + \alpha\boldsymbol{L}_{ii}\boldsymbol{I})^{-1}(\hat{\boldsymbol{P}}^\top\mathbf{e} + \hat{\boldsymbol{Q}}^\top\mathbf{v} - \alpha\sum_{j\neq i}\boldsymbol{L}_{ij}\hat{\mathbf{c}}_j - \beta/2)$.

   Perform a line search on the segment from $\hat{\mathbf{c}}_i$ to $\hat{\mathbf{c}}_i^{new}$ to update $\hat{\mathbf{c}}_i$, A.
4. Check the optimality conditions:

   (a) $\frac{\partial\left(\mathcal{R}_E^{(j)}(\mathbf{c}_i)+\mathcal{R}_V^{(j)}(\mathbf{c}_i)+\mathcal{G}^{(j)}(\mathbf{c}_i)\right)}{\partial \mathbf{c}_i^{(j)}} + \beta = 0$, $\forall \mathbf{c}_i^{(j)} \neq 0$.

   If satisfied, $\Rightarrow$ (b); or, $\Rightarrow$ 3.

   (b) $\left| \frac{\partial\left(\mathcal{R}_E^{(j)}(\mathbf{c}_i)+\mathcal{R}_V^{(j)}(\mathbf{c}_i)+\mathcal{G}^{(j)}(\mathbf{c}_i)\right)}{\partial \mathbf{c}_i^{(j)}} \right| \leq \beta$, $\forall \mathbf{c}_i^{(j)} = 0$.

   If satisfied, return $\mathbf{c}_i$; or, $\Rightarrow$ 2.

4.4.4 Performance evaluation via clustering

We choose to use clustering for evaluation purpose with the rationale that images with similar meanings will have similar representations and hence can be accurately clustered.

The clustering results based on different data representations are displayed in the tables. The best performances achieved by the listed algorithms are compared using the metrics of Accuracy (i.e., AC) and Mutual Information (i.e., MI), whose formulas are in Eq. (3.4) and Eq. (3.5).

Table 4.3: Clustering performance by eye tracking data.

Algorithms	AC (%)	MI (%)
K-means	87.11	95.37
PCA	87.80	95.83
NMF	88.02	95.02
l_1-NMF	88.41	94.64
Graph-based	**90.23**	**96.76**

Table 4.4: Clustering performance by verbal data.

Algorithms	AC (%)	MI (%)
K-means	51.69	68.53
PCA	69.27	79.08
NMF	60.42	74.12
l_1-NMF	66.41	77.43
Graph-based	**70.70**	**80.62**

Table 4.5: Clustering performance by multimodal data.

Algorithms		AC (%)	MI (%)
Concat.*	K-means	87.11	95.37
	PCA	89.84	96.73
Multimodal	NMF	88.15	93.65
	l_1-NMF	89.84	94.58
	Graph-based	**91.67**	**97.76**

*The first 2 rows apply K-means and principal component analysis respectively to the matrix that concatenates eye tracking data and verbal data, because these approaches are not directly applicable to multimodal data.

The results of data instance clustering by multimodal data are in Table 4.5. In order to show the advantage of the proposed data fusion technique, the clustering results based on each single data modality are displayed in Tables 4.3 and 4.4 for comparison purposes. NMF on a single data modality is implemented using the projected gradient method [1]. We derive its multimodal version and display the clustering result in Table 4.5. The implementation of l_1-NMF (sparse NMF regularized by l_1-norm) referred to in all 3 tables is based on the feature-sign search algorithm [4]. The clustering results based on eye tracking features and the clustering results based on linguistic features are listed in Tables 4.3 and 4.4, respectively. The GrNMF approach gives the best results to cluster single-modal data. Table 4.5 displays the clustering results using both data modalities in different ways. The GrNMF-based multimodal data fusion is achieved by using the approach described above. The clustering result based on the coefficients of hidden components obtained from both eye gaze and linguistic features are shown in the last row of Table 4.5. This data fusion approach gives consistently competitive results across 3 tables, which demonstrates the effectiveness of this approach.

Intuitively, the dynamic modeling of conceptual knowledge should be enabled through interpreting verbal descriptions and the eye movement sequences jointly. These two data modalities disclosing two different aspects are also correlated temporally, i.e., an observer's change of eye movement patterns usually indicates her switch between conceptual patterns. However, the temporal correspondence is difficult to elaborate. Therefore, we fused mul-

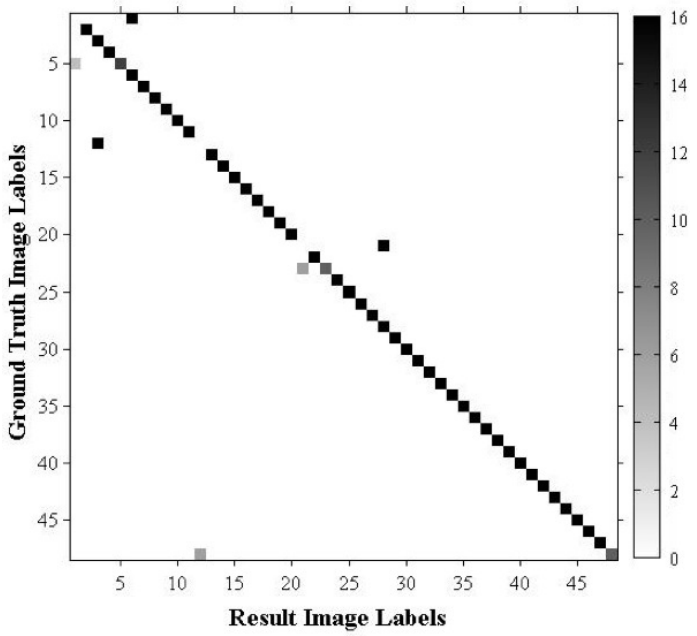


Figure 4.7: Confusion matrix of clustering trials by image based on multimodal GrNMF algorithm. The darkness of a block indicates the number of trials that are in this block. The dark diagonal indicates good clustering performance.

timodal data in order to provide one possible and plausible explanation of the factor behind them both. Particularly, we have developed a multimodal variation of matrix factorization for data fusion [87]. We also incorporate Graph-regularized NMF (GrNMF) in our data fusion model, so as to capture the global data structure configured by both modalities. Figure 4.6 represents our model, and the corresponding objective function is presented in Eq. (4.3). We also use a general-purpose prior (i.e., a sparsity constraint) to enforce each observed data point to be explained by only a few hidden variables. We use l_1-norm for sparsity rather than the intuitive choice of l_0-norm, because the l_1-norm is convex. The solution for l_1-constrained problem can approximate the

solution for the l_0-constrained problem [184]. The details of these constraints are described in Section 2.3.2.

NMF is flexible to learning representations with domain knowledge-related constraints and accept user inputs as constraints. Techniques other than NMF could be adopted instead to learn a unified representation across multimodal data, such as RBM [161]. In the work to fuse multimodal features using RBM, higher level features in different modals are learned separately, and a shared representation is learned on top from these features. However, instead of applying such an approach to our case directly, we need an adapted structure to model our data if adopting RBM. This is because out input data are higher-level human behavioral and cognitive data, which are themselves closer to the semantics than image feature representations. Irrelevant details of the images are excluded. Human image understandings are included in the data. The learning process should be different, in order to adopt RBM-like methods. Not only the reception of natural visual stimuli, clinical diagnosis contains the use of domain expertise and reasoning, which may be modeled as an extra level of hidden variables or an extra set of variables on the same level as others. Romberg et al. developed a data fusion approach with probabilistic latent semantic analysis (pLSA), which is called multilayer multimodal pLSA [185]. It naturally handles multiple data modalities and a hierarchy of abstractions. The basic idea is to apply pLSA in a first step to each mode separately, and in a second step concatenate the derived topic vectors for each modality to learn another pLSA on top to allow grasping concepts across different modalities.

Future studies for data fusion include developing new algorithms and online update rules. For example, PMF and RBM are promising, since they add stochastics and non-linearity properties, which make the model more powerful.

4.5 Conclusions

The experimental results on the unified data representation confirm the effectiveness of our proposed data fusion approach. This study benefits the knowledge-dependent computational systems that involve multiple human sensory and behavioral modalities, such as image classification and retrieval, since they require efficient representation of the multiple elements of human image understanding. As a future work, a full information retrieval system can be developed (Figure 4.8) to extend the current achievements—the multimodal human-centered prototype in Section 4.3 and the data fusion framework in Section 4.4.

Image understanding involves levels of coupled or uncoupled factors that need multiple layers of nodes to represent. Given our goal to understand and group images by expert knowledge, in future work we propose to construct a deep architecture based on observations from two resources—regional image features recognized by computer vision approaches, and language-based semantic features (e.g., the learned interpretable diagnostic narration patterns in Section 3.5), as opposed to learning from pixel-level features in other studies (e.g., object recognition in a large image set). However, empirically, the small scale of our dataset and the heterogeneity of our data modalities introduce

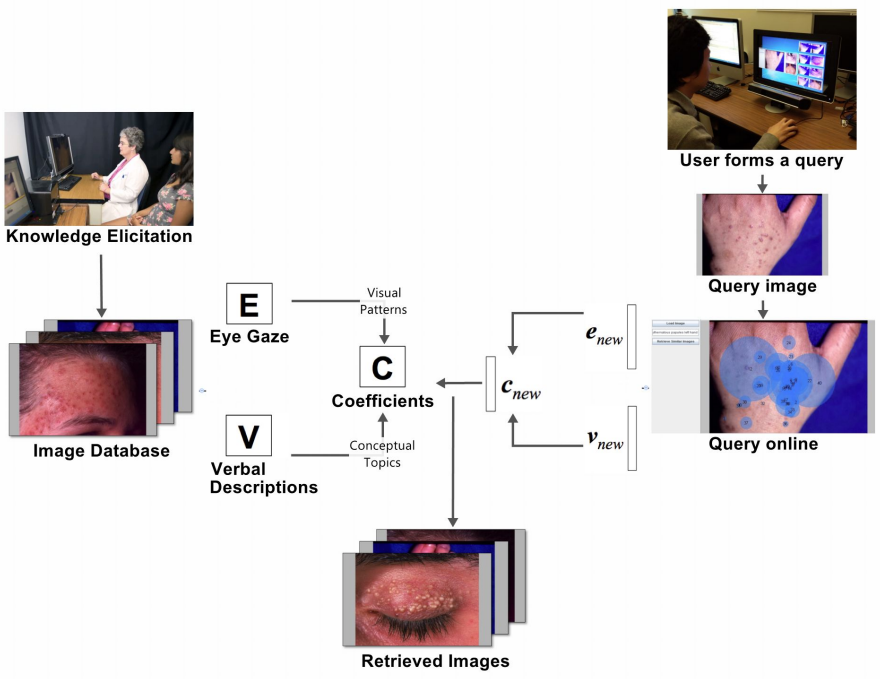


Figure 4.8: An overview of the full retrieval system design. Human knowledge overtly expressed in verbal and eye movement data from both physicians (collected in data-elicitation experiment) and an end user (during interaction with the system) are used for image retrieval. Data in both modalities at both ends are projected into the unified latent space for similarity comparison. The similarity comparison in the latent space is computationally efficient because of the low dimensionality k of the space.

additional challenges for developing an effective deep network that suits our research problem.

To recognize high-level domain-specific features/objects (e.g., lesion blobs), the eye movements can be used with computer vision algorithms that generate regional templates [186]. For example, the mostly viewed image regions can be used to create templates that detect similar regions in an image.

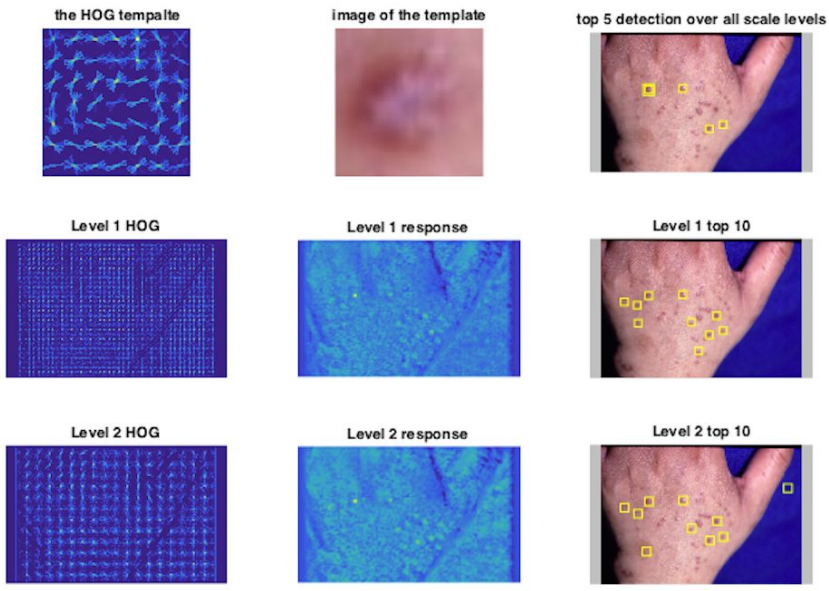


Figure 4.9: A template lesion of papule is derived from an eye fixated papule. This template can be used to detect visually similar lesions in the image.

Figure 4.9 shows an eye fixation-informed template to detect papules in a dermatology image. Since superpixels balance between spatial closeness and feature similarity, visually similar adjacent pixels are clustered as an image segment [187, 188]. Eye fixations can also be used to select or weight these superpixels, so as to represent the image with only the regional features of viewers' interests.

Chapter 5

Interactive Machine Learning for Knowledge Discovery

In visually-oriented specialized medical domains such as dermatology and radiology, physicians explore interesting image cases from medical image repositories for comparative case studies to aid clinical diagnoses, educate medical trainees, and support medical research. This image browsing and lookup could benefit from a grouping of medical images that is consistent with experts' understanding of the image content.

However, semantic image grouping in knowledge-rich domains is challenging, since domain knowledge and human expertise are key to transform image pixels into meaningful content while they tend to be tacit. Manually marking and annotating images is not only labor-intensive but also ineffective.

To use expertise while occupying minimal expert efforts, we present an interactive machine learning paradigm that considers experts as an integral part of the learning process to improve image grouping. This paradigm is designed for automatically computing and quantifying interpretable grouping of dermatological images. In this manner, the computational evolution of an image grouping model, its visualization, and expert interactions form a loop to

improve image grouping. In our paradigm, dermatologists encode their domain knowledge about the medical images by grouping a small subset of images via a carefully designed interface. Our learning algorithm automatically incorporates these manually specified connections as constraints for reorganizing the whole image dataset.

5.1 Background

Interactive learning has become an increasingly popular framework in recent years as it often significantly reduces the efforts associated with data collection on the human end—the machine presents global data patterns for user inspections, and it processes user inputs for overall model updates. The rationale behind interactive approaches is that it is extremely difficult and even undesirable to fully automate application-specific tasks. Instead, a computational design methodology allows gracefully combining automated services with direct user manipulation, so that an end-user's interactions can help solve real-world problems.

As a primary technique of interactive machine learning, *active learning* interactively selects and presents difficult-to-classify learning cases to users and receives users' labeling of them. These learning cases usually have the highest label entropy. Another technique is *reinforcement learning* [189]. It rewards good learning results and penalizes bad learning results based on user feedback. Reinforcement learning does not receive enough inputs, so it is not adopted in our case to work on knowledge-intensive domain.

Different from these existing interactive machine learning approaches, *visual analytics* emphasizes sensemaking of large, complex datasets through interactively exploring visualizations generated by statistical models. Semantic interaction seeks to enable analysts to spatially interact with the models directly within the visual metaphor using interactions that derive from their analytic process, such as searching, highlighting, annotating, and repositioning documents. Analyst can express their expert domain knowledge about the target (e.g., documents, images, etc.), for example, by simply moving them, which guides the underlying model to improve the overall layout, taking the user's feedback into account [190]. One example framework that uses visual analytics to collect user inputs for learning is based on NMF. It modifies the *reference matrices* $\boldsymbol{V}$ and $\boldsymbol{G}$ in the model through user interactions, and optimizes the objective function 5.1 to update the model.

$$\min_{\boldsymbol{H},\boldsymbol{C},\boldsymbol{D}_C \geq 0} \|\boldsymbol{X} - \boldsymbol{H}\boldsymbol{C}\|_F^2 + \|(\boldsymbol{H} - \boldsymbol{V})\boldsymbol{M}_H\|_F^2 + \|(\boldsymbol{C} - \boldsymbol{G}\boldsymbol{D}_C)\boldsymbol{M}_C\|_F^2 \quad (5.1)$$

where $\boldsymbol{H} \in \mathbb{R}_+^{m \times k}$ and $\boldsymbol{C} \in \mathbb{R}_+^{k \times n}$ are the two factor matrices learned to be as close as possible to reference matrices $\boldsymbol{V} \in \mathbb{R}_+^{m \times k}$ and $\boldsymbol{G} \in \mathbb{R}_+^{k \times n}$, respectively, while still approximating the input matrix $\boldsymbol{X}$ by $\boldsymbol{H}\boldsymbol{C}$. $\boldsymbol{M}_H$ and $\boldsymbol{M}_C$ enables such supervision to be applied selectively on a subset of columns of $\boldsymbol{H}$ and $\boldsymbol{C}$ when the corresponding diagonal entries of $\boldsymbol{M}_H$ and $\boldsymbol{M}_C$ are set to zeros. On the other hand, larger diagonal values of $\boldsymbol{M}_H$ and $\boldsymbol{M}_C$ supervise more strongly the corresponding columns of $\boldsymbol{H}$ and $\boldsymbol{C}$, respectively [104].

To achieve 2D visualization of image distributions in the learned representation, techniques such as probabilistic principal component analysis (PPCA), multidimensional scaling (MDS), and generative topographic mapping (GTM) are usually used in visual analytics. In this way, users can use the distance metaphor (the low-dimensional closeness reflects the high-dimensional similarity) to conceptually manipulate the interface [62]. Beecks et al. also discussed different similarity metrics and indexing supports for large-scale multimedia explorations [191, 192]. They are referred to as *visual embedding* approaches.

Section 4.3 builds a multimodal interactive image retrieval system to test a variety of input mechanisms for image retrieval [181]. Experts' preprocessed eye movements and verbal data were stored as meta knowledge in the prototype system, and users interact with the system using these two mechanisms as well. We identify the important image features for query-candidate matching by exploiting local invariant features within the regions of the most interests to viewers as measured by gaze, which indicates expert knowledge-informed points of view. This prototype verifies the feasibility to use verbal data and eye movements as user inputs in a retrieval system.

Since multiple modalities of expert image understanding correlate, I have built a system initialized through the data fusion framework presented in Section 4.4. This system encoded expert interactions as new forms of constraints to guide the learning process to arrive at a more preferred image grouping. The loop with expert constraints is illustrated in Figure 5.1. There are two research directions to achieve *semantic visualization* for the interac-

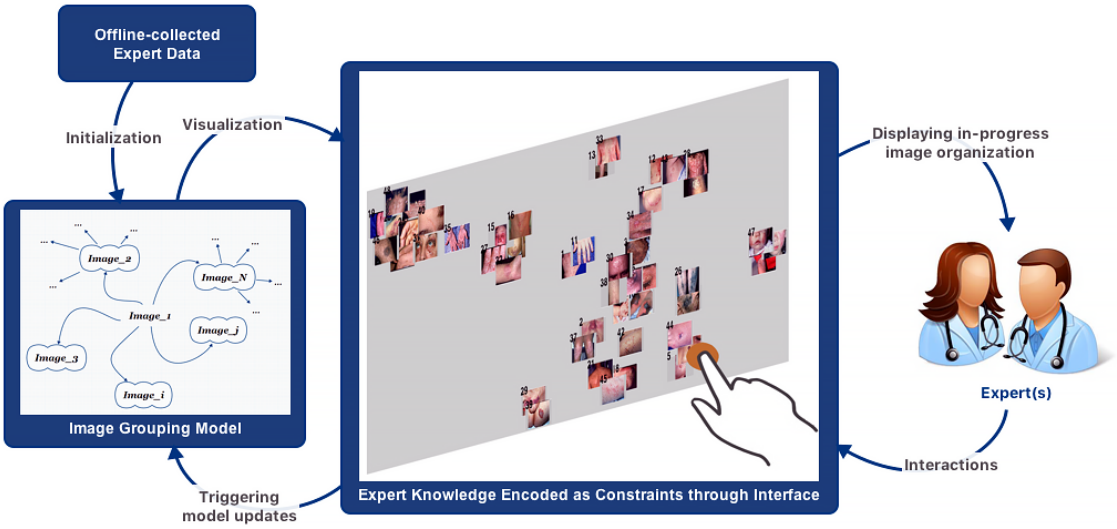


Figure 5.1: Overview of the flow chart of our *expert-in-the-loop* paradigm. An expert encodes domain knowledge as special constraints through rounds of interactions.

tive system—(1) combining a topic modeling approach with a visualization technique, such as "LDA + MDS", and "LDA + t-SNE", and (2) developing a standalone approach to model and visualize the data [193]. For simplicity, the proposed interactive learning paradigm in this chapter follows the former.

5.2 Interactive Image Grouping Paradigm

Since the image grouping learned through general-purpose machine learning algorithms usually does not reflect ideal expert image understanding [194], we also develop an expert-driven interface that allows interactive learning of image organization by merging computational approaches with user interactions [180]. We designed a framework that allows users to justify or correct the

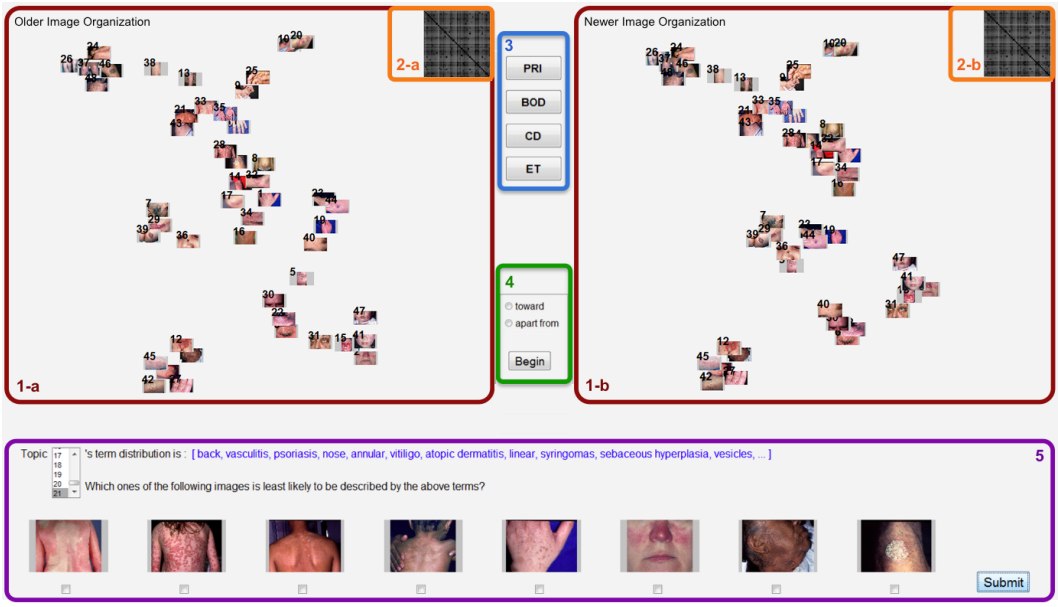
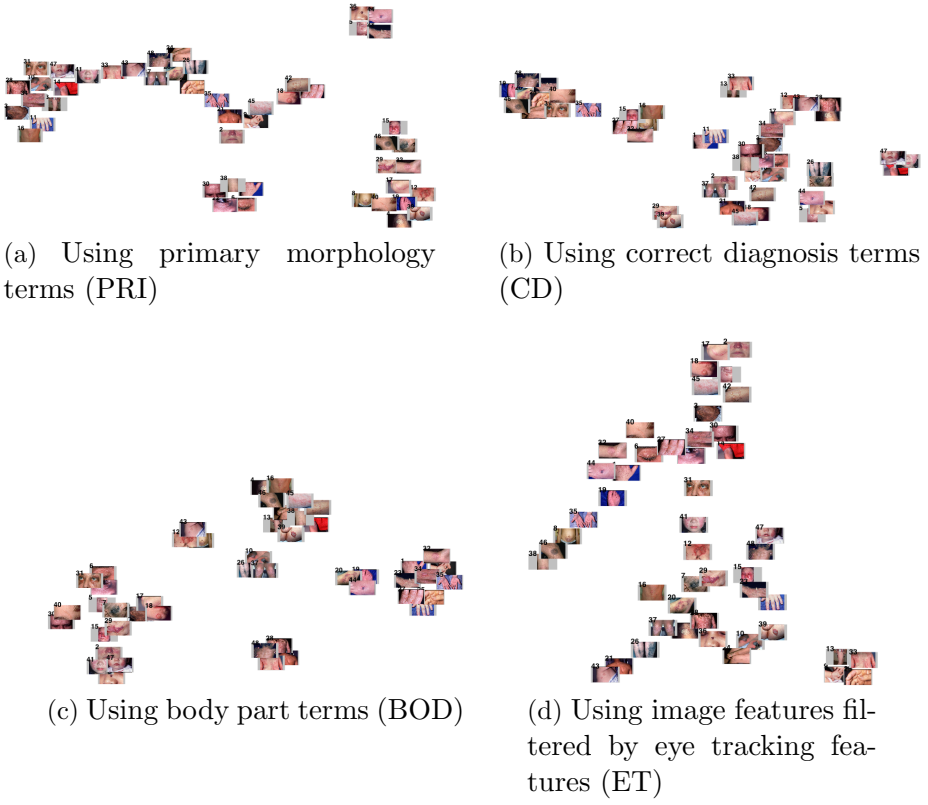


Figure 5.2: Image grouping interface (Details of algorithms behind this interface are described below): Panel (1-a) visualizes the image grouping before each round of expert image manipulation, and panel (1-b) visualizes the resulting image grouping afterwards. Experts are allowed to select multiple images in (1-a) for manipulation. Panels (2-a) and (2-b) are matrix views corresponding to (1-a) and (1-b), respectively, to show global pairwise image similarities. A button set (3) pops up new windows (shown in Figure 5.3 (a-d)) to visualize image grouping initialized using various subsets of features, such as primary morphology terms (PRI). BOD stands for body parts, CD for correct diagnoses, and ET for eye gaze-filtered image features. Panel (4) allows experts to specify the direction to manipulate the selected images. Panel (5) lists the top key terms in each topic and allows experts to disconnect images from a topic.

unsupervised representation learning (i.e., multimodal GrNMF) results, and the user inputs are used as a guidance for further learning. See Figure 5.2 as an illustration of the interface with which users can manipulate the data structure and constrain learning of the image representations.



(a) Using primary morphology terms (PRI)

(b) Using correct diagnosis terms (CD)

(c) Using body part terms (BOD)

(d) Using image features filtered by eye tracking features (ET)

Figure 5.3: Image groupings generated using subsets of features.

5.2.1 Paradigm overview

In order to minimize human efforts and provide experts with a good starting point to group images, we create an initial image grouping using a multimodal expert dataset described in Section 1.1 [6]. This initial image grouping is learned through a multimodal data fusion algorithm (i.e., multimodal GrNMF) flexible to incorporate new images [87]. From here, the loop to improve image grouping begins (see Figure 5.1). An expert can inspect the image grouping and choose to improve it through an interface. Specifically, she encodes her

domain knowledge about the medical images by grouping a small subset of images. Our learning algorithm automatically incorporates these manually specified connections as constraints for reorganizing the whole image dataset. The rules by which the interface processes expert inputs as implicit constraints are described in Section 5.2.5. The incrementally reorganized image set is presented by the visualization techniques in Section 5.2.4. In this way, the computational evolution of an image grouping model, its visualization, and expert interactions form a loop to improve image grouping. The interface design and the supported expert image manipulations are presented in Section 5.2.3. An *expert-in-the-loop* evaluation study is described in Section 5.2.6.

5.2.2 Paradigm initialization

In particular, we create an initial image grouping using a multimodal expert dataset from our prior work [87] (Section 4.4) to minimize human efforts and provide experts with a starting point to group images. This initial image grouping is learned through a multimodal data fusion algorithm flexible to take new images [87]. From here, the loop to improve image grouping begins (see Figure 5.1). An expert can inspect the image grouping and choose to improve it through an interface. The interface then parses expert manipulations as implicit constraints by some rules and incrementally learns the model, and visualizes the new image grouping.

5.2.3 Interface design

The initial image grouping purely based on offline collected expert data is first visualized in panel (1-a) *Older Image Organization* in Figure 5.2 for experts to inspect and manipulate. In the case where domain expert users need further information on the current image grouping, we provide two extra visualizations. First, experts can see an image cluster and the top features contributing to this cluster (see Figure 5.4). Second, experts can click the buttons in Figure 5.2 (panel 3) to compare the image grouping obtained when using different subsets of features, such as only primary morphology terms (Figure 5.3a) with that using the whole feature set (Figure 5.2 (panel 1-a)). The primary morphology is one of the thought units defined by experts in our experiments in Table 1.1.

Experts have two options to improve the image grouping in each round. First, they can directly drag images toward or apart from each other in Figure 5.2 (panel 1-a). The system processes such expert inputs and incorporates them for updating the neighboring graph-based regularizer. Second, experts can select a topic from the listbox in Figure 5.2 (panel 5), and indicate the least relevant image(s) according to the vocabulary distribution of the selected topic. Based on such expert inputs, the system updates the image-topic distribution matrix. After experts interact with the interface using either option, the image grouping in the previous round is copied to Figure 5.2 (panel 1-a), and the improved one is shown in Figure 5.2 (panel 1-b). In each round, both image groupings are visualized following the approaches discussed in Section 5.2.4.

5.2.4 Visualizing image groups

To comprehensively visualize the image grouping, our interface presents both a *graph view* shown in Figure 5.2 (panel 1) and a *matrix view* shown in Figure 5.2 (panel 2). Both views are automatically updated during expert interactions.

In the graph view, we adopt the t-distributed stochastic neighborhood embedding (t-SNE) algorithm. It better visualizes the high dimensional structure of image grouping in 2D graph view than other dimensionality reduction techniques, such as principal component analysis (PCA) [104, 63]. We use a *distance metaphor* to imply to experts that more similar images are spatially closer. However, this metaphor does not proportionally reflect all pairwise image similarities[1] in high-dimensional space, because of the difficulty to retain the whole data structure for any dimensionality reduction algorithms. To tackle this issue, our interface allows experts to see an image and its high dimensional close neighbors in 2D visualization. The popup window visualizing these neighbors is illustrated in Figure 5.4.

The interface also presents a matrix view that serves to give an overview of the pairwise image similarities, because it is impractical that experts choose to see the close neighbors of all images in a 2D graph view. See Figure 5.5 for a magnified matrix view. The matrix view provides a global indexing of pairwise image similarities in the learned representation.

[1]We do not define *image similarity* for domain experts to not restrict them by layperson definitions. We use t-SNE only as a feature projection technique for low dimensional visualization.

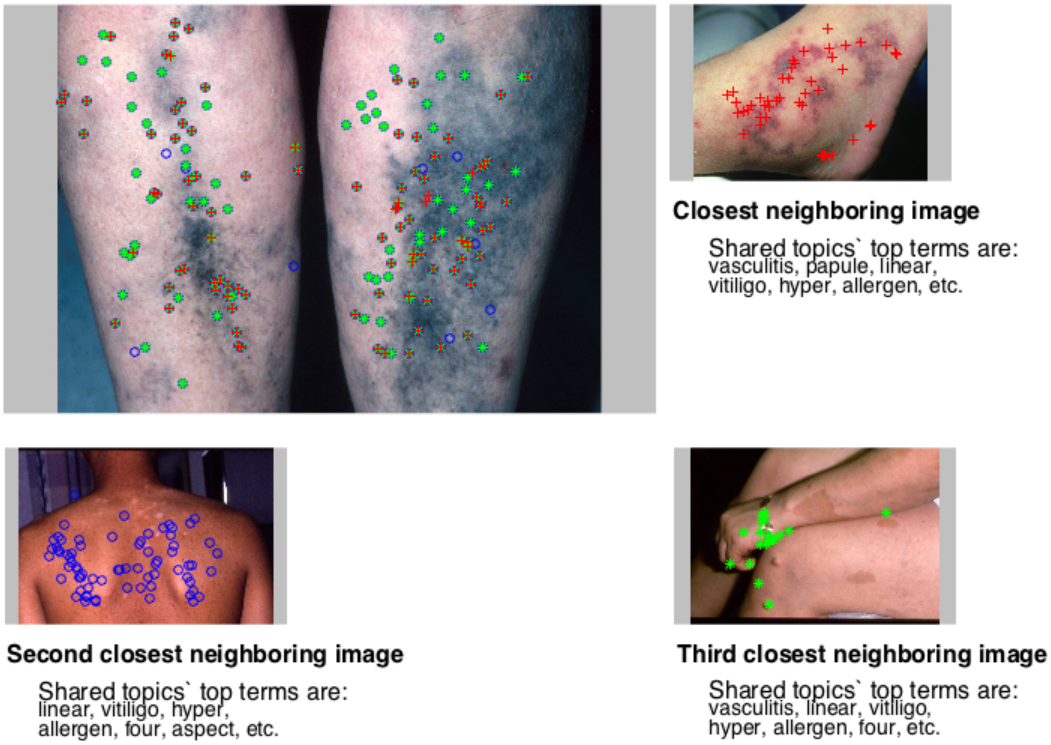


Figure 5.4: An example of the visualization in popup window after the user double-clicks an image in the main interface. The double-clicked image is shown in the top left quarter, and its top 3 neighbors in the learned unified image representation are visualized in other quarters. The shared verbal features are ranked by term frequency, and the top ones are listed below each corresponding neighbor. The shared eye tracking features are also ranked, and the grid segments containing the top ones are marked in both the clicked image and its neighbors. The markers differentiate the image pairs. (Image courtesy of Logical Images, Inc.)

5.2.5 Expert user-specified constraints

There are mainly two approaches in prior studies allowing user interactions to help improve learning a model: document-level interactions [195, 104], or topic/cluster-level interactions [196, 197]. In our scenario, to improve medical

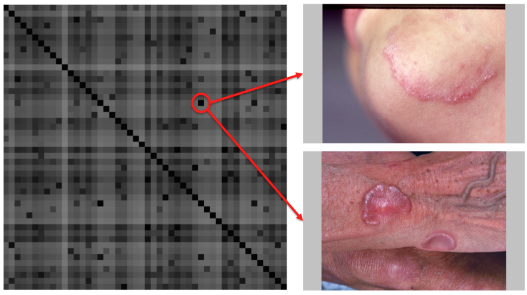


Figure 5.5: An example of the matrix view. The intensity of each block represents the similarity between corresponding images. The darker the block is, the more similar the images are. For example, the similarity between the images on the right is indicated by the dark block circled in the matrix view on the left. (Image courtesy of Logical Images, Inc.)

image grouping, the documents are images. To develop this interface, we prefer document (image)-level interactions for two reasons. On the one hand, the medical conditions are more intuitive in the form of images than texts to physicians. On the other hand, the topics we learned offline based on the multimodal expert dataset from Experiment I are not easily visualizeable nor interpretable by physicians. Below are two functions in the interface for receiving expert inputs and updating the model, both to support image-level interactions.

Constraint on neighboring matrix, $\boldsymbol{W}$: Let the images in the original feature space be denoted as $\mathbf{x}_1$, ..., $\mathbf{x}_m$. A nearest neighbor graph G with m vertices can be constructed. One usual way to compute the element $\boldsymbol{W}_{ij}$ in the neighboring matrix $\boldsymbol{W}$ of the graph G is through a heat kernel presented in Eq. (2.38) [85]. If $\mathbf{x}_i$ and $\mathbf{x}_j$ are identical, then $\boldsymbol{W}_{ij}$ equals 1; and if they are extremely different, then $\boldsymbol{W}_{ij}$ asymptotically approaches 0.

The interface can encode expert image manipulations as a transformation of the neighboring matrix $\boldsymbol{W}$. This transformation is determined by multiple factors, including previous image grouping and experts' interpretation of it. The transformation of $\boldsymbol{W}$ can be simplified as $\mathcal{F}(\cdot,\cdot)$ in Eq. (5.2) and be considered as a constraint set by experts to guide the learning process.

$$\min_{\boldsymbol{P},\boldsymbol{Q},\boldsymbol{C}\geq 0} \|\boldsymbol{E}-\boldsymbol{P}\boldsymbol{C}\|_F^2 + \|\boldsymbol{V}-\boldsymbol{Q}\boldsymbol{C}\|_F^2 + \alpha\mathcal{G}(\mathcal{F}(\boldsymbol{W},K),\boldsymbol{C}) + \beta\mathcal{S}(\boldsymbol{C}) \quad (5.2)$$

where K denotes the set of images selected by an expert in Figure 5.2 (panel 1-a). We use *hard constraints*, i.e., by moving one image toward or away from another, experts can connect or disconnect them in the model. Such expert constraint essentially sets a boundary regarding pairwise image similarities. Once an expert begins to connect these images, the system sets all $\boldsymbol{W}_{ij}$'s $(i,j \in K,\ i \neq j)$ to be 1. Likewise, $\boldsymbol{W}_{ij}$'s $(i,j \in K,\ i \neq j)$ are all set to be 0, if they should be grouped differently. This rule is designed to update the neighboring matrix $\boldsymbol{W}$ in Eq. (2.38). Once all $\boldsymbol{W}_{ij}$'s specified by the expert are updated, the algorithm will trigger the further learning process for the image representation $\boldsymbol{C}$ and the visual and verbal topics $\boldsymbol{P}$ and $\boldsymbol{Q}$ with respect to the objective function in Eq. (5.2).

Constraint on topic-coefficient matrix, $\boldsymbol{C}$: Experts can also improve the image grouping through the task illustrated in Figure 5.2 (panel 5). For each topic selected by experts in the listbox, its top terms in the topic-term distribution are listed. The list of top terms explains the gist of the topic to

experts. The images that are considered highly relevant to the selected topic by the algorithm are then displayed at the bottom. The task for experts is to submit the least relevant image(s) to the topic to disconnect its/their link(s) to the topic. After experts have indicated the least relevant image(s), the system updates the coefficient matrix $\boldsymbol{C}$ according to the constraint in Eq. (5.3).

$$\begin{aligned} \min_{\boldsymbol{P},\boldsymbol{Q},\boldsymbol{C}\geq 0} \|\boldsymbol{E}-\boldsymbol{P}\boldsymbol{C}\|_F^2 + \|\boldsymbol{V}-\boldsymbol{Q}\boldsymbol{C}\|_F^2 + \alpha\mathcal{G}(\boldsymbol{W},\boldsymbol{C}) + \beta\mathcal{S}(\boldsymbol{C}) \\ \text{s.t. } C_{ij}=0,\ i\in T, \text{ and } j\in L(i) \end{aligned} \tag{5.3}$$

where T is the collection of selected topics, and $L(i)$ represents the least relevant images for topic i. The element $\boldsymbol{C}_{ij}$ will be set to 0, if image j is selected to be least relevant to topic i. Once all $\boldsymbol{C}_{ij}$'s are updated, the algorithm begins to learn $\boldsymbol{P}$, $\boldsymbol{Q}$ and $\boldsymbol{C}$ further with respect to Eq. (5.3).

Constraint on topic-basis matrix, $\boldsymbol{Q}$: Given the dermatology image grouping domain where experts intuitively work visually, we value the interactions at the image level. However, our verbal data also suit visual text analytics approaches for updating based on experts' term-level inputs [198].

$$\begin{aligned} \min_{\boldsymbol{P},\boldsymbol{Q},\boldsymbol{C}\geq 0} \|\boldsymbol{E}-\boldsymbol{P}\boldsymbol{C}\|_F^2 + \|\boldsymbol{V}-\boldsymbol{Q}\boldsymbol{C}\|_F^2 + \alpha\mathcal{G}(\boldsymbol{W},\boldsymbol{C}) \\ +\beta\mathcal{S}(\boldsymbol{C}) + \gamma_1\|\boldsymbol{Q}-\boldsymbol{Q}_r\| \end{aligned} \tag{5.4}$$

where $\boldsymbol{Q}_r$ is a matrix consisting of expert inputs for the verbal topics. It can be used to receive experts' changes of topic-term distributions during interactions. Every time the matrix $\boldsymbol{Q}_r$ is changed due to an expert interaction, the underlying verbal topics in $\boldsymbol{Q}$ are learned toward it.

For the update of neighboring matrix $\boldsymbol{W}$, that of the topic-coefficient matrix $\boldsymbol{C}$, and that of the topic-basis matrix $\boldsymbol{Q}$, the model is learned incrementally, and it is consistent between successive interactions. In order for experts to work on consistent image groupings, we also keep the visualization consistent between successive interactions. This is achieved by storing the 2D coordinates of images and using them as the starting point in the graph view (Figure 5.2 (panel 1)) for the next interaction [63].

5.2.6 Evaluation of the paradigm

To evaluate the effectiveness of the paradigm per an expert's objectives, a domain expert was asked to provide a *reference image grouping* that best matches her overall understanding of the relationships between medical images in the database. In particular, for each image she listed its most similar images in terms of their differential diagnoses. We designed an experiment to compare the image grouping performances between the results of fully automated machine learning and our *expert-in-the-loop* paradigm. For fully automated learning (case 1), the resulting image grouping was estimated by a model without expert inputs. In our paradigm (case 2), the physician interacted with the model in the loop toward a better image grouping result. She manipulated the images based on her medical knowledge and the clinical information presented in these images. To quantitatively evaluate the image grouping performances, we retrieved the image neighbors and compared them to the corresponding reference image grouping for both cases.

Table 5.1 summarizes the performances of both cases given various modalities. The image groupings with expert interactive constraints consistently outperform the traditional learning case. In particular, our paradigm performs much better than fully automated learning with verbal feature of correct diagnosis (CD). This suggests that diagnoses are the primary factor considered by the expert to group medical images. Furthermore, learning from multimodal features achieves the best performance for both cases. We also elicited the expert's qualitative evaluation through an interview. The expert noticed the improvement of each iteration. K-means, PCA, hierarchical clustering [199], LSA, and LDA are used for comparison purposes. Since these algorithms are not easily applied to the multiple modalities, their multimodal performances are omitted. Density-based and distribution-based algorithms do not work because of the small number of data instances we have so far.

Similar to Table 5.1, Tables 5.2-5.4 show the performances within the top 10, 15, and 20 retrieved neighbors. Various algorithms are good at modeling specific data modalities, which directs our future work to model the data with different statistical properties distinctly [169]. Since we are using multimodal expert data for initialization, multiple graph-weighting strategies can also be adopted in the future to capture various data attributes [86]. Image features generally do not perform as well as verbal features. Intuitively, this is because the verbal features contain the terms that capture the domain knowledge, whereas the image features do not. Also observed is that the eye-tracking filters do not always boost the performance of image features. This

Table 5.1: Image grouping performances of fully automated learning and our paradigm. The measurement is the percentage of images in the reference list to appear within the top 5 retrieved neighbors. Different combinations of modalities include primary morphology terms only (PRI), body location terms only (BOD), correct diagnoses terms only (CD), SIFT features only, SIFT features filtered by gaze features (SIFT+Gaze), and multimodal data (overall).

		Verbal			SIFT	SIFT+Gaze	Multimodal
		PRI	BOD	CD			
case 1 (fully automated learning)	K-means	29.46	11.61	14.29	8.93	14.29	–
	PCA	41.07	16.07	45.54	16.07	15.18	–
	Hierarchical clustering	25.00	11.61	12.50	11.61	13.39	–
	Latent Semantic Analysis	33.04	12.50	47.32	–	–	–
	Latent Dirichlet Allocation	15.18	8.93	17.86	–	–	–
	Laplacian sparse coding	33.04	14.29	36.61	10.71	14.29	**52.68**
case 2 (our paradigm)		34.82	16.96	42.86	12.50	17.86	**59.82**

points us to future work involving improved use of eye movement data, such as using the high-level behavioral patterns [15]. Experimental outcomes show that Laplacian sparse coding does not always or substantially beat some learners. Therefore, our future work also includes adaptation and improvement of other approaches for interactive machine learning.

During the paradigm evaluation, we also recorded the expert's verbal labeling of the image groups. The labeling of image groups is useful to disclose her diagnostic reasoning while grouping images. This can be incorporated in future work to optimize the semantic feature space. Another important part of our future work involves implementing our paradigm on a larger dermatological image database with more experts in the loop to test our paradigm's robustness. New images with no eye-tracking trials and with no or very sparse annotations (few words of the morphology categories or the disease names) will

Table 5.2: The percentage of images in the reference list to appear within the top 10 retrieved neighbors

		Verbal			SIFT	SIFT+Gaze	Multimodal
		PRI	BOD	CD			
case 1 (fully automated learning)	K-means	33.04	20.54	29.46	21.43	16.96	–
	PCA	55.36	27.68	63.39	28.57	26.79	–
	Hierarchical clustering	29.46	17.86	25.00	25.89	25.89	–
	Latent Semantic Analysis	49.11	23.21	58.93	–	–	–
	Latent Dirichlet Allocation	22.32	17.86	25.89	–	–	–
	Laplacian sparse coding	48.21	20.54	50.00	26.79	36.61	**71.43**
case 2 (our paradigm)		48.21	26.79	62.50	26.79	34.82	**69.64**

Table 5.3: The percentage of images in the reference list to appear within the top 15 retrieved neighbors

		Verbal			SIFT	SIFT+Gaze	Multimodal
		PRI	BOD	CD			
case 1 (fully automated learning)	K-means	40.18	37.50	31.25	39.29	23.21	–
	PCA	67.86	41.07	72.32	42.86	40.18	–
	Hierarchical clustering	37.50	37.50	41.07	35.71	41.07	–
	Latent Semantic Analysis	61.61	33.04	63.39	–	–	–
	Latent Dirichlet Allocation	29.46	28.57	35.71	–	–	–
	Laplacian sparse coding	56.25	31.25	54.46	41.07	43.75	**76.79**
case 2 (our paradigm)		63.39	34.82	73.21	36.61	41.07	**77.68**

Table 5.4: The percentage of images in the reference list to appear within the top 20 retrieved neighbors

		Verbal			SIFT	SIFT+Gaze	Multimodal
		PRI	BOD	CD			
case 1 (fully automated learning)	K-means	46.43	36.61	41.96	41.07	20.54	–
	PCA	75.00	53.57	75.89	49.11	51.79	–
	Hierarchical clustering	43.75	42.86	52.68	53.57	50.89	–
	Latent Semantic Analysis	70.54	42.86	66.96	–	–	–
	Latent Dirichlet Allocation	37.50	41.07	46.43	–	–	–
	Laplacian sparse coding	59.82	42.86	63.39	55.36	55.36	**82.14**
case 2 (our paradigm)		75.89	43.75	78.57	52.68	50.00	**83.04**

be first positioned in the model simply based on visual similarities. An image hierarchy can be learned and visualized. For the ease of expert interactions, a

few representative images can be selected from each group. In the case where new images do not even have offline annotations, they can still be positioned in an existing image grouping for further improvements, since single modality features can be easily projected into the unified topic space [87].

The presentation of image groupings could also be based on experts' trade-off between various factors, such as the primary lesion morphology and the causes of the diseases. Our current visualization may not be feasible for a larger database. It is necessary to design a more effective visualization strategy to allow experts to explore both global structure and local details of image grouping. To receive experts' accurate inputs through interactions, a learning framework with feature selection can be adopted [200]. Furthermore, to minimize the offset between the neighborhood in the topic space and that in the visualization space, a joint regularization strategy can be developed [201]. It can also be envisioned that when dealing with a larger dataset a few image cases could constantly bubble up in the neighborhood and cause user fatigue while repeatedly skipping them. Our future work therefore also includes developing a penalty term to isolate the images implicitly skipped by a user.

There are both global and local constraints in our paradigm. In general, the global constraint such as the neighboring matrix in Section 5.2.5 is to make sure that the learned hidden topics best retain the relationships between the observed data points, whereas the local constraints usually refer to experts' localized changes regarding a small subset of image relations. The balance

is achieved through the interactive process when the machine and an expert finally agree upon each other's decision. For the flexibility of the model and the generalization of the paradigm for a larger dataset, our future work involves replacing the hard constraints in Eq. (5.2) with soft ones. In this way, the parameters in neighboring graph can also be learned and adapted to reflect the relative similarities among the neighbors locally. In order to balance the influences between the offline collected expert data and online expert inputs, other soft constraints could be applied by encoding expert interactions in a new penalty term. Besides, similar to updating verbal topics in Eq. (5.4), we plan to allow updating eye movement-filtered image patterns through a similar term $\gamma_2\|\boldsymbol{P} - \boldsymbol{P}_r\|$ and support such updates by adding corresponding visualizations.

In our model, expert input is transformed into constraints, which are then used to update the model. Experts have the flexibility to provide all the constraints in one round or separate them into multiple rounds. The order of these constraints does not affect the final model. In another word, the final model remains the same as long as the same set of constraints are provided (stability). However, the intermediate result of the model may affect an expert's decision making, which may lead to them to provide different constraints. Such bidirectional effects have been observed in human-agent reciprocal social interaction studies [202]. This kind of dynamics is interesting and can be studied in our future work.

In the realm of interactive machine learning, there is always a trade-

off between the power of the used model to capture the underlying semantics and the simplicity of the model to achieve good responsiveness and support realtime interactions. In a typical interaction loop based on our current implementation, the expert spent 1 min inputting her constraint and the learning algorithms (including the visualization algorithm) converged within 10 seconds in a single-core machine. We consider approximated learning rules for better responsiveness in the future and online learning algorithms to handle new data points [203]. Moreover, we may use a different learning framework for fast model updates than that for model initialization [204].

5.3 Conclusions

This chapter presents an interactive machine learning paradigm with experts in the loop for improving image grouping. We demonstrate that image grouping can be significantly improved by expert constraints through incremental updates of the underlying computational model. In each iteration, our paradigm allows to accommodate our model to experts' input. Performance evaluation shows that expert constraints are an effective way to infuse expert knowledge into the learning process and improve overall image grouping. The contributions of this chapter involves many areas, including *interaction-based visual analytics*, *knowledge discovery through interactions*, *user modeling during interactions*, and *human-centered computing*.

– **Interaction-based visual analytics:** Existing systems that allow interactive user visual analysis usually adopt topic modeling techniques

[104, 195, 196, 197, 205]. Original features are reduced to a lower-dimensional topic space, in which documents are grouped. One type of such system, including UTOPIAN [104] and iVisClustering [195], visualizes the topics, so that users can adjust the topic-term distribution at the term granularity. In contrast, our paradigm focuses experts on natural high-level image grouping tasks and encodes expert image manipulations as constraints to improve the overall image grouping. Moreover, in our domain the objects for experts to interact with are medical images rather than latent topics, which may be confusing to the experts. Another type of system, including LSAView [196] and iVisClassifier [197], involves document-level interactions. These systems require users to change the parameters of the algorithms. In contrast, our system updates the underlying topic model based on experts' natural manipulations of the images.

– **Knowledge discovery through interactions:** There are many existing visual analytic applications whose purpose is for data exploration and summarization [206, 207, 208]. The visualized data clusters can be easily interpreted. For knowledge discovery purposes, there are also applications in the domains such as geography [209], whose outcome is also straightforward. Our paradigm is presented in the medical domain where the understanding and interpretations are difficult. We elicit and use the knowledge and expertise of the medical end users through interactions.

– **User modeling during interactions:** Similar to the ReGroup system's approach to interactively tailor its suggestions [210], our paradigm allows the model and the user to learn from each other. As with more and more interactions between an expert user and the learning algorithm, the underlying model is gradually adapted to the user's mental model (how she groups the images and what her standards are). The model records her personalized considerations during the task. Having different users in the loop results in different model outputs. These outputs can evolve separately, or be weighted to achieve an overall model. We seamlessly integrate machine learning and an adaptive user interaction mechanism to collect the most useful information that is complementary to the limited data issue.

– **Human-centered computing:** The loop requires both the computational strength of machine learning algorithms and the domain knowledge from the experts. The experts are given high-level and natural tasks, and local changes made by the experts can cause global updates of the underlying model. The global constraint is to make sure that the learned hidden topics can be used to best recover the observed data points whereas the local constraints come from the expert input. The balance is achieved through the interactive process when the machine and expert finally agree upon each other's decision.

Chapter 6

Summary

6.1 Conclusions

This dissertation approaches the research topic of medical image grouping from the domain experts' point-of-view. Multimodal data are collected from physicians, and the data are modeled to discover a domain knowledge representation which exhibits physicians' understanding of the medical images. Three research directions to approach a knowledge representation are taken:

The collected physicians' diagnostic verbal narratives contain medical terms. This makes the verbal data a straightforward data modality to infer physicians' uses of domain knowledge and cognitive reasoning processes. We model these elicited diagnostic verbal narratives by developing and using machine learning approaches. In particular, two models are developed by assuming a verbal narrative either *a bag of medical concepts* or *a sequence of medical concepts*, both involving solving an inverse problem. The discovered conceptual topics facilitate grouping the medical images based on domain-specific semantics.

Experts' eye movement data are useful to inform important image regions from expert perspective. We use eye movements, along with verbal de-

scriptions, to develop a human-centered information retrieval system. These two data modalities are also fused to discover a unified data representation, where the corresponding medical images can be grouped.

To allow domain experts further refining the image grouping with minimal efforts, an interactive machine learning paradigm is developed. The underlying machine learning model supports knowledge discovery by presenting data similarities to experts, and the experts constrain the learning process through rounds of interactions. This paradigm allows the model to evolve with expert interactions, and it showcases the expert-machine collaboration.

The approaches proposed in this dissertation can facilitate education in the medical fields, research in cognition and decision-making, and medical image classification and use based on physicians' thoughts.

6.2 Future Work

As discussed in Section 2.3.2.6, complex models (e.g., deep models) cannot be easily applied on our dataset, since these models involve a larger number of parameters to learn, which requires a large dataset. Meanwhile, recruiting domain experts for medical knowledge data collection is expensive. However, we still plan some remedies for this difficulty—(1) to exploit the medical knowledge ontology additionally, (2) to carefully develop hierarchical models specific to our dataset and domain, and (3) to design an effective and intelligent interface to reduce human efforts while providing domain knowledge.

6.2.1 External knowledge resources

We have used external knowledge resource (i.e., the UMLS) for narrative processing and semantic relatedness computation in this dissertation. Future work can also use ontology resources for various purposes. For example, the UMLS can be used in the model initialization stage so that the image grouping matches as much of experts' mental model as possible. It can also be used as a mid-layer to help verbalize the learned *topics* to domain experts (in their language), which makes the machine learning results more interpretable. Furthermore, the relatedness score for each pair of medical terms computed using the UMLS in Section 3.4 can be developed as a constraint to guide the learning process.

6.2.2 Multimodal data fusion

Based on the approaches and findings in this dissertation, future work includes modeling the multimodal data with a more complex model that explains the dependencies between the latent variables learned from different data modalities.

As eye movements reflect the spatial distribution of important image regions, the eye movement patterns can be modeled as a weighted aggregation of image patterns. Likewise, the verbal description patterns can be dependent on the above two types of patterns.

6.2.3 Interactive machine learning

For the interactive machine learning to work intelligently with the expert users, another research direction could be inferring experts' intended conceptual level and encoding their inputs into the learning model at that level (change of hyper-parameters). Adaptive model structures can be learned for different individual expert's mental model (change of model structure).

Since the effectiveness of visualization in a knowledge discover system is key to elicit expert knowledge, we also plan to improve the existing interactive learning paradigm with a more natural mental mapping between behaviors and results. The envisioned challenges include (1) the projection from high dimensional data to 2D/3D visualization, (2) the space limits of user interface, and (3) a comprehensive viewpoint of data relations presented to expert users.

The envisioned full system that allows two interaction styles (see Figure 4.8) can also be implemented for semantic image retrieval.

Appendix

Appendix 1

Publications

Peer-Reviewed Journal Articles

1. Xuan Guo, Qi Yu, Rui Li, Cecilia Ovesdotter Alm, Cara Calvelli, Pengcheng Shi, and Anne R. Haake. Intelligent medical image grouping through interactive learning. *International Journal of Data Science and Analytics* 2, no. 3-4: 95-105, 2016.

2. Xuan Guo, Qi Yu, Cecilia Ovesdotter Alm, Cara Calvelli, Jeff B. Pelz, Pengcheng Shi, and Anne R. Haake. From spoken narratives to domain knowledge: Mining linguistic data for medical image understanding. *Artificial Intelligence in Medicine*, 62(2): 79–90, 2014

Peer-Reviewed Conference Papers

3. Weishi Shi, Qi Yu, and Xuan Guo. An efficient many-class active learning framework for knowledge-rich domains. In *IEEE International Conference on Data Mining (ICDM'17)*. Under Review.

4. Xuan Guo, Rui Li, Qi Yu, and Anne R. Haake. Modeling physicians' utterances to explore diagnostic decision-making. In *the International Joint Conference on Artificial Intelligence (IJCAI'17)*. Accepted.

5. Xuan Guo, Qi Yu, Rui Li, Cecilia Ovesdotter Alm, Cara Calvelli, Pengcheng Shi, and Anne R. Haake. An expert-in-the-loop paradigm for learning medical image grouping. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'16)*, pp. 477-488. Springer International Publishing, 2016.

6. Xuan Guo. Multimodal interactive machine learning for user understanding. In *Proceedings of the* 20^{th} *International Conference on Intelligent User Interfaces Companion (IUI'15)*, pp. 129-132. ACM, 2015.

7. Xuan Guo, Qi Yu, Rui Li, Cecilia Ovesdotter Alm, and Anne R. Haake. Fusing multimodal human expert data to uncover hidden semantics. In *Proceedings of the* 7^{th} *Workshop on Eye Gaze in Intelligent Human Machine Interaction: Eye-Gaze & Multimodality (ICMI'14-GazeIn)*, pages 21–26. ACM, 2014.

8. Xuan Guo, Rui Li, Cecilia Ovesdotter Alm, Qi Yu, Jeff B. Pelz, Pengcheng Shi, and Anne R. Haake. Infusing perceptual expertise and domain knowledge into a human-centered image retrieval system: A prototype application. In *Proceedings of the Symposium on Eye Tracking Research & Applications (ETRA'14)*, pages 275–278. ACM, 2014.

Bibliography

[1] Chih-Jen Lin. Projected gradient methods for nonnegative matrix factorization. *Neural Computation*, 19(10):2756–2779, 2007.

[2] Hyunsoo Kim and Haesun Park. Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method. *SIAM Journal on Matrix Analysis and Applications*, 30(2):713–730, 2008.

[3] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*, pages 556–562, 2001.

[4] Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Y Ng. Efficient sparse coding algorithms. In *Advances in Neural Information Processing Systems*, pages 801–808, 2006.

[5] Miao Zheng, Jiajun Bu, Chun Chen, Can Wang, Lijun Zhang, Guang Qiu, and Deng Cai. Graph regularized sparse coding for image representation. *IEEE Transactions on Image Processing*, 20(5):1327–1336, 2011.

[6] Preethi Vaidyanathan, Jeff Pelz, Rui Li, Sai Mulpuru, Dong Wang, Pengcheng Shi, Cara Calvelli, and Anne Haake. Using human experts' gaze data to evaluate image processing algorithms. In 10^{th} *IVMSP Workshop: Perception and Visual Signal Analysis*, pages 129–134. IEEE, 2011.

[7] David Needham and Kevin Flint. Uncovering the truth behind vygotsky's cognitive apprenticeship: Engaging reflective practitioners in the'master-apprentice'relationship. *International Journal of Learning*, 10, 2004.

[8] Wilson McCoy, Cecilia Ovesdotter Alm, Cara Calvelli, Rui Li, Jeff B Pelz, Pengcheng Shi, and Anne Haake. Annotation schemes to encode domain knowledge in medical narratives. In *Proceedings of the* 6^{th} *Linguistic Annotation Workshop*, pages 95–103. Association for Computational Linguistics, 2012.

[9] Paul Boersma and David Weenink. Praat: Doing phonetics by computer (version 5.1. 05) [computer program]. `http://www.praat.org/` (Accessed: 10 April 2014), 2009.

[10] Faye Lyons and Lisa Ellen Ousley. *Dermatology for the Advanced Practice Nurse*. Springer Publishing Company, 2014.

[11] Rui Li, Preethi Vaidyanathan, Sai Mulpuru, Jeff Pelz, Pengcheng Shi, Cara Calvelli, and Anne Haake. Human-centric approaches to image understanding and retrieval. In *Western New York Image Processing Workshop (WNYIPW)*, pages 62–65. IEEE, 2010.

[12] Rui Li, Jeff Pelz, Pengcheng Shi, Cecilia Ovesdotter Alm, and Anne R Haake. Learning eye movement patterns for characterization of perceptual expertise. In *Proceedings of the Symposium on Eye Tracking Research & Applications (ETRA)*, pages 393–396. ACM, 2012.

[13] Rui Li, Jeff Pelz, Pengcheng Shi, and Anne R Haake. Learning image-derived eye movement patterns to characterize perceptual expertise. In *Proceedings of the* 34^{th} *Annual Meeting of the Cognitive Science Society*, pages 190–195, 2012.

[14] Preethi Vaidyanathan, Jeff Pelz, Cecilia Alm, Pengcheng Shi, and Anne Haake. Recurrence quantification analysis reveals eye-movement behavior differences between experts and novices. In *Proceedings of the Symposium on Eye Tracking Research & Applications (ETRA)*, pages 303–306. ACM, 2014.

[15] Rui Li, Pengcheng Shi, and Anne R Haake. Image understanding from experts' eyes by modeling perceptual skill of diagnostic reasoning processes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2187–2194. IEEE, 2013.

[16] Rui Li, Pengcheng Shi, Jeff Pelz, Cecilia O Alm, and Anne R Haake. Modeling eye movement patterns to characterize perceptual skill in image-based diagnostic reasoning processes. *Computer Vision and Image Understanding*, 151:138–152, 2016.

[17] Kathryn Womack, Wilson McCoy, Cecilia Ovesdotter Alm, Cara Calvelli, Jeff B Pelz, Pengcheng Shi, and Anne Haake. Disfluencies as extra-propositional indicators of cognitive processing. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics (ExProM)*, pages 1–9. Association for Computational Linguistics, 2012.

[18] Olivier Bodenreider. The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267–D270, 2004.

[19] Pat Croskerry. A universal model of diagnostic reasoning. *Academic Medicine*, 84(8):1022–1028, 2009.

[20] Cathryn A Galanter and Vimla L Patel. Medical decision making: A selective review for child psychiatrists and psychologists. *Journal of Child Psychology and Psychiatry*, 46(7):675–689, 2005.

[21] Judith L Bowen. Educational strategies to promote clinical diagnostic reasoning. *New England Journal of Medicine*, 355(21):2217–2225, 2006.

[22] Andrew Evered, Darren Walker, Andrew A Watt, and Nick Perham. To what extent does nonanalytic reasoning contribute to visual learning in cytopathology? *Cancer Cytopathology*, 121(6):329–338, 2013.

[23] Yannis Kalfoglou, Srinandan Dasmahapatra, David Dupplaw, Bo Hu, Paul Lewis, and Nigel Shadbolt. Living with the semantic gap: Experiences and remedies in the context of medical imaging. In 1^{st} *International Conference on Semantics and Digital Media Technologies (SAMT)*, pages 46–47. Springer, 2006.

[24] George A Miller. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[25] Karin Kipper Schuler. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, Univ. of Pennsylvania, 2005.

[26] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. IEEE, 2009.

[27] Kin Wah Fung and MD Olivier Bodenreider. Knowledge representation and ontologies. In *Clinical Research Informatics*, pages 255–275. Springer-Verlag London Limited, 2012.

[28] Domenico M Pisanelli, Aldo Gangemi, and Geri Steve. An ontological analysis of the UMLS Metathesaurus. In *American Medical Informatics Association Annual Symposium Proceedings*, pages 810–814. AMIA, 1998.

[29] Vijay N Garla and Cynthia Brandt. Ontology-guided feature engineering for clinical text classification. *Journal of Biomedical Informatics*, 45(5):992–998, 2012.

[30] Yanpeng Li and Hongfang Liu. Learning semantic tags from big data for clinical text representation. *AMIA Summits on Translational Science Proceedings*, 2015:461, 2015.

[31] Alan R Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. In *American Medical Informatics Association Annual Symposium Proceedings*, pages 17–21. AMIA, 2001.

[32] Javier Tejedor, Doroteo T Toledano, Dong Wang, Simon King, and José Colás. Feature analysis for discriminative confidence estimation in spoken term detection. *Computer Speech & Language*, 28(5):1083–1114, 2014.

[33] Joseph Bullard, Cecilia Ovesdotter Alm, Qi Yu, Pengcheng Shi, and Anne Haake. Towards multimodal modeling of physicians' diagnostic confidence and self-awareness using medical narratives. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1718–1727. Dublin City University and Association for Computational Linguistics, 2014.

[34] Kathryn Womack, Cecilia Ovesdotter Alm, Cara Calvelli, Jeff Pelz, Pengcheng Shi, and Anne R Haake. Markers of confidence and correctness in spoken medical narratives. In *Proceedings of the* 14^{th} *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2549–2553. ACM, 2013.

[35] Limor Hochberg, Cecilia Ovesdotter Alm, Esa M Rantanen, Qi Yu, Caroline M DeLong, and Anne R Haake. Towards automatic annotation of clinical decision-making style. *The* 8^{th} *Linguistic Annotation Workshop*, page 129, 2014.

[36] Juana María Ruiz-Martínez, Rafael Valencia-García, Jesualdo Tomás Fernández-Breis, Francisco García-Sánchez, and Rodrigo Martínez-Béjar. Ontology learning from biomedical natural language documents using UMLS. *Expert Systems with Applications*, 38(10):12365–12378, 2011.

[37] Xuan Guo, Qi Yu, Cecilia Ovesdotter Alm, Cara Calvelli, Jeff B Pelz, Pengcheng Shi, and Anne R Haake. From spoken narratives to domain knowledge: Mining linguistic data for medical image understanding. *Artificial Intelligence in Medicine*, 62(2):79–90, 2014.

[38] Elia Bruni, Giang Binh Tran, and Marco Baroni. Distributional semantics from text and images. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 22–32. Association for Computational Linguistics, 2011.

[39] D.A. Forsyth, Tamara Berg, Cecilia Ovesdotter Alm, and Gang Wang. Words and pictures: Categories, modifiers, depiction, and iconography. In *Object Categorization: Computer and Human Vision Perspectives*, pages 167–181. Cambridge University Press, 2009.

[40] Lei Wu, Xian-Sheng Hua, Nenghai Yu, Wei-Ying Ma, and Shipeng Li. Flickr distance: A relationship measure for visual concepts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(5):863–875, 2012.

[41] Li-Jia Li, Chong Wang, Yongwhan Lim, David M Blei, and Li Fei-Fei. Building and using a semantivisual image hierarchy. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3336–3343. IEEE, 2010.

[42] George Kingsley Zipf. *The Psycho-Biology of Language.* Boston, MA: Houghton, Mifflin, 1935.

[43] Robert R Hoffman and Stephen M Fiore. Perceptual (re)learning: A leverage

point for human-centered computing. *Intelligent Systems, IEEE*, 22(3):79–83, 2007.

[44] Anil K Jain. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.

[45] Geoff Norman. Dual processing and diagnostic errors. *Advances in Health Sciences Education*, 14(1):37–49, 2009.

[46] Jonathan Sherbino, Kelly L Dore, Timothy J Wood, Meredith E Young, Wolfgang Gaissmaier, Sharyn Kreuger, and Geoffrey R Norman. The relationship between response time and diagnostic accuracy. *Academic Medicine*, 87(6):785–791, 2012.

[47] Marie-Claude Audétat, Valérie Dory, Mathieu Nendaz, Dominique Vanpee, Dominique Pestiaux, Noelle Junod Perron, and Bernard Charlin. What is so difficult about managing clinical reasoning difficulties? *Medical Education*, 46(2):216–227, 2012.

[48] Dario D Salvucci and Joseph H Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the Symposium on Eye Tracking Research & Applications (ETRA)*, pages 71–78. ACM, 2000.

[49] Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost Van de Weijer. *Eye tracking: A Comprehensive Guide to Methods and Measures*. Oxford University Press, 2011.

[50] Andrew Duchowski. *Eye tracking methodology: Theory and practice*, volume 373. Springer Science & Business Media, 2007.

[51] Alexander C Schütz, Doris I Braun, and Karl R Gegenfurtner. Eye movements and perception: A selective review. *Journal of Vision*, 11(5):9, 2011.

[52] Behrad Noudoost, Mindy H Chang, Nicholas A Steinmetz, and Tirin Moore. Top-down control of visual attention. *Current Opinion in Neurobiology*, 20(2):183–190, 2010.

[53] John M Henderson. Human gaze control during real-world scene perception. *Trends in Cognitive Sciences*, 7(11):498–504, 2003.

[54] Kai Kunze, Hitoshi Kawaichi, Kazuyo Yoshimura, and Koichi Kise. Towards inferring language expertise using eye tracking. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, pages 217–222. ACM, 2013.

[55] Christian Schulze, Robby Frister, and Faisal Shafait. Eye-tracker based part-image selection for image retrieval. In *International Conference on Image Processing (ICIP)*, pages 4392–4396, 2013.

[56] Alberto Faro, Daniela Giordano, Carmelo Pino, and Concetto Spampinato. Visual attention for implicit relevance feedback in a content based image retrieval. In *Proceedings of the Symposium on Eye-Tracking Research & Applications (ETRA)*, pages 73–76. ACM, 2010.

[57] Georg Buscher, Edward Cutrell, and Meredith Ringel Morris. What do you see when you're surfing?: Using eye tracking to predict salient regions of

web pages. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 21–30. ACM, 2009.

[58] Andreas Gegenfurtner, Erno Lehtinen, and Roger Säljö. Expertise differences in the comprehension of visualizations: A meta-analysis of eye-tracking research in professional domains. *Educational Psychology Review*, 23(4):523–552, 2011.

[59] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[60] Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the* 26^{th} *Annual International Conference on Machine Learning*, pages 609–616. ACM, 2009.

[61] Maneesh Bhand, Ritvik Mudur, Bipin Suresh, Andrew Saxe, and Andrew Y Ng. Unsupervised learning models of primary cortical receptive fields and receptive field plasticity. In *Advances in Neural Information Processing Systems*, pages 1971–1979, 2011.

[62] Alex Endert, Chao Han, Dipayan Maiti, Leanna House, Scotland Leman, and Chris North. Observation-level interaction with statistical models for visual analytics. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, pages 121–130. IEEE, 2011.

[63] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2579-2605):85, 2008.

[64] Ian Jolliffe. *Principal Component Analysis*. Wiley Online Library, 2002.

[65] Adam Coates, Andrew Y Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *International Conference on Artificial Intelligence and Statistics*, pages 215–223, 2011.

[66] Jorge E Camargo and Fabio A Gonzalez. Multimodal visualization based on latent topic analysis. In *IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2014.

[67] Lawrence Cayton. Algorithms for manifold learning. *Univ. of California at San Diego Tech. Rep*, pages 1–17, 2005.

[68] Seungjin Choi, Andrzej Cichocki, Hyung-Min Park, and Soo-Young Lee. Blind source separation and independent component analysis: A review. *Neural Information Processing-Letters and Reviews*, 6(1):1–57, 2005.

[69] Christopher J James and Christian W Hesse. Independent component analysis for biomedical signals. *Physiological measurement*, 26(1):R15, 2005.

[70] Christian Bauckhage. k-means clustering is matrix factorization. *arXiv preprint arXiv:1512.07548*, 2015.

[71] John Canny. GaP: A factor model for discrete data. In *Proceedings of the* 27^{th} *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 122–129. ACM, 2004.

[72] Menaka Rajapakse and Lnnce Wyse. NMF vs ICA for face recognition. In *Proceedings of the* 3^{rd} *International Symposium on Image and Signal Processing and Analysis (ISPA)*, volume 2, pages 605–610. IEEE, 2003.

[73] Chris HQ Ding, Xiaofeng He, and Horst D Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. In *SDM*, volume 5, pages 606–610. SIAM, 2005.

[74] Ali Taylan Cemgil. Bayesian inference for nonnegative matrix factorisation models. *Computational Intelligence and Neuroscience*, 2009, 2009.

[75] Bruno A Olshausen et al. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

[76] Qi Yu. Sparse functional representation for large-scale service clustering. In Chengfei Liu, Heiko Ludwig, Farouk Toumani, and Qi Yu, editors, *Proceedings of* 10^{th} *International Conference on Service-Oriented Computing (ICSOC)*, pages 468–483. Springer, 2012.

[77] Robert Peharz and Franz Pernkopf. Sparse nonnegative matrix factorization with l_0-constraints. *Neurocomputing*, 80:38–46, 2012.

[78] Patrik O Hoyer. Non-negative matrix factorization with sparseness constraints. *The Journal of Machine Learning Research*, 5:1457–1469, 2004.

[79] Taehwan Kim, Gregory Shakhnarovich, and Raquel Urtasun. Sparse coding for learning interpretable spatio-temporal primitives. In *Advances in Neural*

Information Processing Systems, pages 1117–1125, 2010.

[80] Feiping Nie, Heng Huang, Xiao Cai, and Chris H Ding. Efficient and robust feature selection via joint $l_{2,1}$-norms minimization. In *Advances in Neural Information Processing Systems*, pages 1813–1821, 2010.

[81] Patrik O Hoyer. Non-negative sparse coding. In *Proceedings of the* 12^{th} *IEEE Workshop on Neural Networks for Signal Processing*, pages 557–565. IEEE, 2002.

[82] Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37(23):3311–3325, 1997.

[83] Ganesh R Naik, editor. *Non-negative Matrix Factorization Techniques: Advances in Theory and Applications*. Signals and Communication Technology. Springer-Verlag Berlin Heidelberg, 2016.

[84] Shenghua Gao, Ivor Waihung Tsang, Liang-Tien Chia, and Peilin Zhao. Local features are not lonely–Laplacian sparse coding for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3555–3561. IEEE, 2010.

[85] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing Systems (NIPS)*, volume 14, pages 585–591, 2001.

[86] Jim Jing-Yan Wang, Halima Bensmail, and Xin Gao. Multiple graph regularized nonnegative matrix factorization. *Pattern Recognition*, 46(10):2840–2847, 2013.

[87] Xuan Guo, Qi Yu, Rui Li, Cecilia Ovesdotter Alm, and Anne R Haake. Fusing multimodal human expert data to uncover hidden semantics. In *Proceedings of the* 7^{th} *Workshop on Eye Gaze in Intelligent Human Machine Interaction: Eye-Gaze & Multimodality*, pages 21–26. ACM, 2014.

[88] Shuiwang Ji, Lei Yuan, Ying-Xin Li, Zhi-Hua Zhou, Sudhir Kumar, and Jieping Ye. Drosophila gene expression pattern annotation using sparse features and term-term interactions. In *Proceedings of the* 15^{th} *ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 407–416. ACM, 2009.

[89] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the* 25^{th} *International Conference on Machine Learning*, pages 1096–1103. ACM, 2008.

[90] Nicolas Le Roux and Yoshua Bengio. Representational power of restricted Boltzmann machines and deep belief networks. *Neural Computation*, 20(6):1631–1649, 2008.

[91] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful repre-

sentations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(Dec):3371–3408, 2010.

[92] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[93] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, 2006.

[94] Marc'Aurelio Ranzato, Christopher Poultney, Sumit Chopra, and Yann LeCun. Efficient learning of sparse representations with an energy-based model. In *Proceedings of the* 19^{th} *International Conference on Neural Information Processing Systems*, pages 1137–1144. MIT Press, 2006.

[95] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

[96] Koray Kavukcuoglu, Marc'Aurelio Ranzato, and Yann LeCun. Fast inference in sparse coding algorithms with applications to object recognition. *arXiv preprint arXiv:1010.3467*, 2010.

[97] Andriy Mnih and Ruslan Salakhutdinov. Probabilistic matrix factorization. In *Advances in Neural Information Processing Systems*, pages 1257–1264, 2007.

[98] Chris Ding, Tao Li, and Wei Peng. On the equivalence between non-negative

matrix factorization and probabilistic latent semantic indexing. *Computational Statistics & Data Analysis*, 52(8):3913–3927, 2008.

[99] Karthik Devarajan, Guoli Wang, and Nader Ebrahimi. A unified statistical approach to non-negative matrix factorization and probabilistic latent semantic indexing. *Machine learning*, 99(1):137–163, 2015.

[100] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

[101] Alexandra Schofield, Måns Magnusson, and David Mimno. Pulling out the stops: Rethinking stopword removal for topic models. *Proceedings of the* 15^{th} *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 432–436, 2017.

[102] Andrew T Wilson and Peter A Chew. Term weighting schemes for latent dirichlet allocation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 465–473. Association for Computational Linguistics, 2010.

[103] Thiago de Paulo Faleiros, Alneu de Andrade Lopes, et al. On the equivalence between algorithms for non-negative matrix factorization and latent dirichlet allocation. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, XXIV*. European Neural Network Society-ENNS, 2016.

[104] Jaegul Choo, Changhyun Lee, Chandan K Reddy, and Haesun Park. UTOPIAN: User-driven topic modeling based on interactive nonnegative matrix factorization. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):1992–2001, 2013.

[105] David Andrzejewski, Xiaojin Zhu, and Mark Craven. Incorporating domain knowledge into topic modeling via dirichlet forest priors. In *Proceedings of the* 26^{th} *Annual International Conference on Machine Learning*, pages 25–32. ACM, 2009.

[106] Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 952–961. Association for Computational Linguistics, 2012.

[107] Lawrence R Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[108] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):721–741, 1984.

[109] Jun S Liu. The collapsed gibbs sampler in bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, 89(427):958–966, 1994.

[110] Jurgen Van Gael, Yunus Saatci, Yee Whye Teh, and Zoubin Ghahramani. Beam sampling for the infinite hidden Markov model. In *Proceedings of the* 25^{th} *International Conference on Machine Learning*, pages 1088–1095. ACM, 2008.

[111] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 2006.

[112] Nima Razavi, Juergen Gall, and Luc Van Gool. Scalable multi-class object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1505–1512. IEEE, 2011.

[113] Mohammad Amin Sadeghi and Ali Farhadi. Recognition using visual phrases. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1745–1752. IEEE, 2011.

[114] Helmut Grabner, Juergen Gall, and Luc Van Gool. What makes a chair a chair? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1529–1536. IEEE, 2011.

[115] Micah K Johnson and Edward H Adelson. Shape estimation in natural illumination. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2553–2560. IEEE, 2011.

[116] Beyang Liu, Stephen Gould, and Daphne Koller. Single image depth estimation from predicted semantic labels. In *IEEE Conference on Computer*

Vision and Pattern Recognition (CVPR), pages 1253–1260. IEEE, 2010.

[117] Ying Liu, Dengsheng Zhang, Guojun Lu, and Wei-Ying Ma. A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, 40(1):262–282, 2007.

[118] Ritendra Datta, Jia Li, and James Z Wang. Content-based image retrieval: Approaches and trends of the new age. In *Proceedings of the* 7^{th} *ACM SIGMM International Workshop on Multimedia Information Retrieval*, pages 253–262. ACM, 2005.

[119] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (CSUR)*, 40(2):5, 2008.

[120] Li-Jia Li, Hao Su, Eric P Xing, and Li Fei-Fei. Object bank: A high-level image representation for scene classification and semantic feature sparsification. *Advances in Neural Information Processing Systems*, 24:1378–1386, 2010.

[121] R Priyatharshini and S Chitrakala. Association based image retrieval: A survey. In *Mobile Communication and Power Engineering*, pages 17–26. Springer, 2013.

[122] Ceyhun Burak Akgül, Daniel L Rubin, Sandy Napel, Christopher F Beaulieu, Hayit Greenspan, and Burak Acar. Content-based image retrieval in radiology: Current status and future directions. *Journal of Digital Imaging*, 24(2):208–222, 2011.

[123] B Mohammed Imran and MM Sufyan Beg. Towards perception-based image retrieval. In *Advances in Computer Science and Information Technology*, volume 86, pages 280–289. Springer, Heidelberg, 2012.

[124] Akiko Aizawa. An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, 39(1):45–65, 2003.

[125] Qi Yu and Manjeet Rege. On service community learning: A co-clustering approach. In 8^{th} *International Conference on Web Services (ICWS)*, pages 283–290. IEEE, 2010.

[126] Chris Ding, Tao Li, Wei Peng, and Haesun Park. Orthogonal nonnegative matrix t-factorizations for clustering. In *Proceedings of the* 12^{th} *ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 126–135, New York, NY, USA, 2006. ACM.

[127] Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Y Ng. Efficient sparse coding algorithms. In *Proceedings of the* 19^{th} *Annual Conference on Neural Information Processing Systems (NIPS)*, pages 801–808. MIT Press, 2006.

[128] Peter Gärdenfors. Conceptual spaces as a framework for knowledge representation. *Mind and Matter*, 2(2):9–27, 2004.

[129] Peter Gärdenfors. *Conceptual Spaces: The Geometry of Thought*. MIT Press, Cambridge, MA, USA, 2004.

[130] Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis.

Journal of the American Sociery for Information Science (JASIS), 41(6):391–407, 1990.

[131] László Lovász and Michael D Plummer. *Matching Theory, Volume 121 of North-Holland Mathematics Studies*. North-Holland Publishing Co., Amsterdam, 1986.

[132] Ulrike Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[133] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, et al. *Introduction to Data Mining*. Pearson Education India, 2007.

[134] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[135] Bridget T McInnes, Ted Pedersen, and Serguei VS Pakhomov. UMLS-interface and UMLS-similarity: Open source software for measuring paths and semantic similarity. In *American Medical Informatics Association Annual Symposium Proceedings*, pages 431–435. AMIA, 2009.

[136] Daniel Jurafsky and James H Martin. *Speech and Language Processing (Prentice Hall Series in Artificial Intelligence)*. Prentice Hall, 2008.

[137] Hongfang Liu, Stephen B Johnson, and Carol Friedman. Automatic resolution of ambiguous terms based on machine learning and conceptual relations in the UMLS. *Journal of the American Medical Informatics Association*, 9(6):621–636, 2002.

[138] Kiem-Hieu Nguyen and Cheol-Young Ock. Semantic relatedness for biomedical word sense disambiguation. In *Workshop Proceedings of TextGraphs-7 on Graph-based Methods for Natural Language Processing*, pages 25–29. Association for Computational Linguistics, 2012.

[139] Ying Liu, Bridget T McInnes, Ted Pedersen, Genevieve Melton-Meaux, and Serguei Pakhomov. Semantic relatedness study using second order co-occurrence vectors computed from biomedical corpora, UMLS and WordNet. In *Proceedings of the* 2^{nd} *ACM SIGHIT International Health Informatics Symposium*, pages 363–372. ACM, 2012.

[140] David Yarowsky. Word sense disambiguation. In *Handbook of Natural Language Processing,* 2^{nd} *Edition*, pages 315–338. CRC Press, Taylor and Francis Group, 2010.

[141] Bridget T McInnes, Ted Pedersen, Ying Liu, Genevieve B Melton, and Serguei V Pakhomov. Knowledge-based method for determining the meaning of ambiguous biomedical terms using information content measures of similarity. In *American Medical Informatics Association Annual Symposium Proceedings*, pages 895–904. AMIA, 2011.

[142] D Griffiths. A pragmatic approach to Spearman's rank correlation coefficient. *Teaching Statistics*, 2(1):10–13, 1980.

[143] Sorana-Daniela Bolboaca and Lorentz Jäntschi. Pearson versus Spearman, Kendall's tau correlation analysis on structure-activity relationships of biologic active compounds. *Leonardo Journal of Sciences*, 5(9):179–200, 2006.

[144] Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

[145] Eta S Berner and Mark L Graber. Overconfidence as a cause of diagnostic error in medicine. *The American Journal of Medicine*, 121(5):S2–S23, 2008.

[146] Danushka Bollegala, Takanori Maehara, and Ken-ichi Kawarabayashi. Embedding semantic relations into word representations. In *Proceedings of the* 24^{th} *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1222–1228. AAAI Press, 2015.

[147] Run-ze Wu, Qi Liu, Yuping Liu, Enhong Chen, Yu Su, Zhigang Chen, and Guoping Hu. Cognitive modeling for predicting examinee performance. In *Proceedings of the* 24^{th} *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1017–1024. AAAI Press, 2015.

[148] Andreas Holzinger and Igor Jurisica. Knowledge discovery and data mining in biomedical informatics: The future is in integrative, interactive machine learning solutions. In *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics*, pages 1–18. Springer, 2014.

[149] Michael Duerr-Specht, Randy Goebel, and Andreas Holzinger. Medicine and health care as a data problem: Will computers become better medical doctors? In *Smart Health*, pages 21–39. Springer, 2015.

[150] Jinchang Ren, Junwei Han, and Mauro Dalla Mura. Special issue on multimodal data fusion for multidimensional signal processing. *Multidimensional*

Systems and Signal Processing, 27(4):801–805, 2016.

[151] Felix Putze, Jutta Hild, Rainer Kärgel, Christian Herff, Alexander Redmann, Jürgen Beyerer, and Tanja Schultz. Locating user attention using eye tracking and EEG for spatio-temporal event selection. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces*, pages 129–136. ACM, 2013.

[152] Olaf Dimigen, Werner Sommer, Annette Hohlfeld, Arthur M Jacobs, and Reinhold Kliegl. Coregistration of eye movements and EEG in natural reading: Analyses and review. *Journal of Experimental Psychology: General*, 140(4):552, 2011.

[153] Nur Huseyin Kaplan, Isin Erer, and Furkan Elibol. Fusion of multispectral and panchromatic images by combining bilateral filter and ihs transform. In *Proceedings of the* 20^{th} *European Signal Processing Conference (EUSIPCO)*, pages 2501–2505. IEEE, 2012.

[154] Peicheng Zhou, Gong Cheng, Zhenbao Liu, Shuhui Bu, and Xintao Hu. Weakly supervised target detection in remote sensing images based on transferred deep features and negative bootstrapping. *Multidimensional Systems and Signal Processing*, 27(4):925–944, 2016.

[155] Yijun Yan, Jinchang Ren, Yinsheng Li, James FC Windmill, Winifred Ijomah, and Kuo-Ming Chao. Adaptive fusion of color and spatial features for noise-robust retrieval of colored logo and trademark images. *Multidimensional Systems and Signal Processing*, 27(4):945–968, 2016.

[156] Jiangbin Zheng, Yanan Liu, Jinchang Ren, Tingge Zhu, Yijun Yan, and Heng Yang. Fusion of block and keypoints based approaches for effective copy-move image forgery detection. *Multidimensional Systems and Signal Processing*, 27(4):989–1005, 2016.

[157] Jie Ren, Ming Xu, Jeremy S Smith, and Shi Cheng. Multi-view and multi-plane data fusion for effective pedestrian detection in intelligent visual surveillance. *Multidimensional Systems and Signal Processing*, 27(4):1007–1029, 2016.

[158] Laurence Nigay and Joëlle Coutaz. A design space for multimodal systems: Concurrent processing and data fusion. In *Proceedings of the INTERACT'93 and CHI'93 Conference on Human Factors in Computing Systems*, pages 172–178. ACM, 1993.

[159] Juan C Caicedo and Fabio A González. Multimodal fusion for image retrieval using matrix factorization. In *Proceedings of the* 2^{nd} *ACM International Conference on Multimedia Retrieval*, pages 56–56, New York, NY, USA, 2012. ACM.

[160] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55, 2012.

[161] Xiaoyi Li, Jing Gao, Hui Li, Le Yang, and Rohini K Srihari. A multimodal framework for unsupervised feature fusion. In *Proceedings of the* 22^{nd} *ACM*

International Conference on Information & Knowledge Management, pages 897–902. ACM, 2013.

[162] Honglak Lee, Chaitanya Ekanadham, and Andrew Y Ng. Sparse deep belief net model for visual area V2. In *Advances in Neural Information Processing Systems*, pages 873–880, 2008.

[163] Aaron C Courville, James Bergstra, and Yoshua Bengio. A spike and slab restricted boltzmann machine. In *International Conference on Artificial Intelligence and Statistics*, pages 233–241, 2011.

[164] Marc Aurelio Ranzato, Fu Jie Huang, Y-Lan Boureau, and Yann LeCun. Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2007.

[165] Ruslan Salakhutdinov and Geoffrey Hinton. Semantic hashing. *RBM*, 500(3):500, 2007.

[166] Jing Huang and Brian Kingsbury. Audio-visual deep learning for noise robust speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7596–7599. IEEE, 2013.

[167] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[168] Ruslan Salakhutdinov and Geoffrey E Hinton. Deep Boltzmann machines. In *International Conference on Artificial Intelligence and Statistics*, pages 448–455, 2009.

[169] Nitish Srivastava and Ruslan R Salakhutdinov. Multimodal learning with deep Boltzmann machines. In *Advances in Neural Information Processing Systems*, pages 2222–2230, 2012.

[170] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 35(8):1798–1828, 2013.

[171] Yunlong He, Koray Kavukcuoglu, Yun Wang, Arthur Szlam, and Yanjun Qi. Unsupervised feature learning by deep sparse coding. In *Proceedings of the SIAM International Conference on Data Mining*, pages 902–910, 2014.

[172] Kevin J Lang, Alex H Waibel, and Geoffrey E Hinton. A time-delay neural network architecture for isolated word recognition. *Neural networks*, 3(1):23–43, 1990.

[173] Jason Weston, Frédéric Ratle, Hossein Mobahi, and Ronan Collobert. Deep learning via semi-supervised embedding. In *Neural Networks: Tricks of the Trade*, pages 639–655. Springer, 2012.

[174] Grégoire Montavon, Mikio L Braun, and Klaus-Robert Müller. Kernel analysis of deep networks. *The Journal of Machine Learning Research*, 12:2563–2581, 2011.

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

[176] Samuel Audet, Masatoshi Okutomi, and Masayuki Tanaka. Direct image alignment of projector-camera systems with planar surfaces. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 303–310. IEEE, 2010.

[177] Gerald Westheimer. Visual acuity and spatial modulation thresholds. In *Visual Psychophysics*, pages 170–187. Springer, 1972.

[178] Kathryn Womack, Cecilia Alm, Cara Calvelli, Jeff Pelz, Pengcheng Shi, and Anne Haake. Using linguistic analysis to characterize conceptual units of thought in spoken medical narratives. In *Proceedings of INTERSPEECH*, pages 3722–3726, 2013.

[179] David McG Squire, Wolfgang Müller, Henning Müller, and Jilali Raki. Content-based query of image databases, inspirations from text retrieval: Inverted files, frequency-based weights and relevance feedback. In *Proceedings of the* 11^{th} *Scandinavian Conference on Image Analysis*, pages 143–149, 1999.

[180] Pietro Michelucci. *Handbook of Human Computation*. Springer, 2013.

[181] Xuan Guo, Rui Li, Cecilia Ovesdotter Alm, Qi Yu, Jeff B Pelz, Pengcheng Shi, and Anne R Haake. Infusing perceptual expertise and domain knowledge into a human-centered image retrieval system: A prototype application.

In *Proceedings of the Symposium on Eye Tracking Research & Applications (ETRA)*, pages 275–278. ACM, 2014.

[182] Kristiina Jokinen. Turn taking, utterance density, and gaze patterns as cues to conversational activity. In *Proceedings of ICMI 2011 Workshop Multimodal Corpora for Machine Learning: Taking Stock and Road mapping the Future*, pages 31–36, 2011.

[183] Jarkko Salojärvi, Ilpo Kojo, Jaana Simola, and Samuel Kaski. Can relevance be inferred from eye movements in information retrieval. In *Proceedings of WSOM*, volume 3, pages 261–266, 2003.

[184] Joel A Tropp. Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE Transactions on Information Theory*, 52(3):1030–1051, 2006.

[185] Stefan Romberg, Rainer Lienhart, and Eva Hörster. Multimodal image retrieval. *International Journal of Multimedia Information Retrieval*, 1(1):31–44, 2012.

[186] David R Martin, Charless C Fowlkes, and Jitendra Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(5):530–549, 2004.

[187] Greg Mori. Guiding model search using segmentation. In 10^{th} *IEEE International Conference on Computer Vision (ICCV)*, volume 2, pages 1417–1423.

IEEE, 2005.

[188] Xiaofeng Ren and Jitendra Malik. Learning a classification model for segmentation. In *Proceedings of the* 9^{th} *IEEE International Conference on Computer Vision (ICCV)*, pages 10–17. IEEE, 2003.

[189] Richard S Sutton and Andrew G Barto. *Introduction to Reinforcement Learning*, volume 135. MIT Press Cambridge, 1998.

[190] Alex Endert, Patrick Fiaux, and Chris North. Semantic interaction for visual text analytics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 473–482. ACM, 2012.

[191] Christian Beecks, Tomáš Skopal, Klaus Schöffmann, and Thomas Seidl. Towards large-scale multimedia exploration. In *Proceedings of the* 5^{th} *International Workshop on Ranking in Databases (DBRank)*, pages 31–33, 2011.

[192] Christian Beecks. *Distance-based Similarity Models for Content-based Multimedia Retrieval*. PhD thesis, RWTH Aachen University, 2013.

[193] Tuan M.V. Le and Hady W. Lauw. Manifold learning for semantic visualization. *Journal of Artificial Intelligence Research*, 1:1–15, 2015.

[194] Andreas Holzinger. Human-computer interaction and knowledge discovery (HCI-KDD): What is the benefit of bringing those two fields to work together? In *Availability, Reliability, and Security in Information Systems and HCI*, pages 319–328. Springer, 2013.

[195] Hanseung Lee, Jaeyeon Kihm, Jaegul Choo, John Stasko, and Haesun Park. iVisClustering: An interactive visual document clustering via topic modeling. In *Computer Graphics Forum*, volume 31, pages 1155–1164. Wiley Online Library, 2012.

[196] Patricia J Crossno, Daniel M Dunlavy, and Timothy M Shead. LSAView: A tool for visual exploration of latent semantic modeling. In *IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pages 83–90. IEEE, 2009.

[197] Jaegul Choo, Hanseung Lee, Jaeyeon Kihm, and Haesun Park. iVisClassifier: An interactive visual analytics system for classification based on supervised dimension reduction. In *IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pages 27–34. IEEE, 2010.

[198] Jaegul Choo, Changhyun Lee, Chandan K Reddy, and Haesun Park. Weakly supervised nonnegative matrix factorization for user-driven clustering. *Data Mining and Knowledge Discovery*, 29(6):1598–1621, 2015.

[199] Abdelmoula El-Hamdouchi and Peter Willett. Comparison of hierarchic agglomerative clustering methods for document retrieval. *The Computer Journal*, 32(3):220–227, 1989.

[200] Jing-Yan Wang, Islam Almasri, and Xin Gao. Adaptive graph regularized nonnegative matrix factorization via feature selection. In 21^{st} *International Conference on Pattern Recognition (ICPR)*, pages 963–966. IEEE, 2012.

[201] Tuan Le and Hady W Lauw. Semantic visualization with neighborhood graph regularization. *Journal of Artificial Intelligence Research*, 55:1091–1133, 2016.

[202] Andrea L Thomaz and Cynthia Breazeal. Transparency and socially guided machine learning. In 5^{th} *International Conference on Development and Learning (ICDL)*, 2006.

[203] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online learning for matrix factorization and sparse coding. *The Journal of Machine Learning Research*, 11:19–60, 2010.

[204] Georg Fuchs, Hendrik Stange, Ahmad Samiei, Gennady Andrienko, and Natalia Andrienko. A semi-supervised method for topic extraction from micro postings. *IT-Information Technology*, 57(1):49–56, 2015.

[205] Yuening Hu, Jordan Boyd-Graber, Brianna Satinoff, and Alison Smith. Interactive topic modeling. *Machine Learning*, 95(3):423–469, 2014.

[206] Changqing Zhou, Dan Frankowski, Pamela Ludford, Shashi Shekhar, and Loren Terveen. Discovering personal gazetteers: An interactive clustering approach. In *Proceedings of the* 12^{th} *Annual ACM International Workshop on Geographic Information Systems*, pages 266–273. ACM, 2004.

[207] Sara Johansson and Jimmy Johansson. Interactive dimensionality reduction through user-defined combinations of quality metrics. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):993–1000, 2009.

[208] Jing Yang, Daniel Hubball, Matthew O Ward, Elke A Rundensteiner, and William Ribarsky. Value and relation display: Interactive visual exploration of large data sets with hundreds of dimensions. *IEEE Transactions on Visualization and Computer Graphics*, 13(3):494–507, 2007.

[209] Diansheng Guo, Donna J Peuquet, and Mark Gahegan. Iceage: Interactive clustering and exploration of large and high-dimensional geodata. *GeoInformatica*, 7(3):229–253, 2003.

[210] Saleema Amershi, James Fogarty, and Daniel Weld. Regroup: Interactive machine learning for on-demand group creation in social networks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 21–30. ACM, 2012.

Vita

Xuan Guo was born in Tianjin, China on Jan. 31^{st}, 1989, the son of Xin Guo and Jing Zhao. He received the Bachelor of Engineering degree in Software Engineering from Nankai University, Tianjin, China in 2011. He is expecting a PhD degree from Rochester Institute of Technology at the time of writing. His research interest includes machine learning and knowledge discovery, human computer interactions, and natural language processing. His current research includes multimodal data representation learning and interactive machine learning.

Temporary address: 20 Lomb Memorial Dr.
Rochester, New York 14623