Intelligent medical image grouping through interactive learning

Xuan Guo¹ · Qi Yu¹ · Rui Li¹ · Cecilia Ovesdotter Alm² · Cara Calvelli³ · Pengcheng Shi¹ · Anne Haake¹

Received: 16 June 2016 / Accepted: 11 August 2016 / Published online: 25 August 2016
© Springer International Publishing Switzerland 2016

Abstract Image grouping in knowledge-rich domains is challenging, since domain knowledge and human expertise are key to transform image pixels into meaningful content. Manually marking and annotating images is not only labor-intensive but also ineffective. Furthermore, most traditional machine learning approaches cannot bridge this gap for the absence of experts' input. We thus present an interactive machine learning paradigm that allows experts to become an integral part of the learning process. This paradigm is designed for automatically computing and quantifying interpretable grouping of dermatological images. In this way, 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. Performance evaluation shows that this paradigm effectively improves image grouping based on expert knowledge.

Keywords Dermatological images · Multimodal data · Image grouping · Visual analytics · Interactive machine learning

1 Introduction

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, it is challenging, because medical image interpretation usually requires domain knowledge that tends to be tacit. Therefore, to make expertise more explicit we propose an interactive machine learning paradigm that has experts in the loop to improve image grouping.

1.1 Contributions

The contributions of this paper 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 [2–6]. Original features are reduced to a lower-dimensional topic space, in which documents are grouped. One type of such system, including UTOPIAN [3] and iVisClustering [6], visualizes the topics, so that users can adjust the topic-term distribution at the term granularity. In contrast, our paradigm focuses

This paper is an extension version of the PAKDD2016 Long Presentation paper “An Expert-in-the-loop Paradigm for Learning Medical Image Grouping” [1].

✉ Xuan Guo
xxg3358@g.rit.edu

¹ B. Thomas Golisano College of Computing and Information Sciences, Rochester Institute of Technology, 20 Lomb Memorial Drive, Rochester, NY 14623, USA

² College of Liberal Arts, Rochester Institute of Technology, 92 Lomb Memorial Drive, 14623 Rochester, NY, USA

³ College of Health Sciences and Technology, Rochester Institute of Technology, 90 Lomb Memorial Drive, 14623 Rochester, NY, USA

Fig. 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

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 [4] and iVisClassifier [2], 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 [7–9]. The visualized data clusters can be easily interpreted. For knowledge discovery purposes, there are also applications in the domains such as geography [10], whose outcome is also straightforward. Our paradigm is presented in 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 to interactively tailor its suggestions [11], 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 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 outputs. 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.

1.2 Roadmap

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 Sect. 2 [12]. This initial image grouping is learned through a multimodal data fusion algorithm flexible to incorporate new images [13]. From here, the loop to improve image grouping begins (see Fig. 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 parses expert inputs as implicit constraints are described in Sect. 5. The incrementally reorganized image set is presented by the visualization techniques in Sect. 4. In this way, the computational evolution of an image grouping model, its visualization, and expert interactions form a loop to improve image grouping. We presented some preliminary result of this work in a conference [1]. This journal paper

Fig. 2 Image features filtered by an expert's eye gaze. (Image courtesy of Logical Images, Inc.). **a** A sample image, **b** SIFT features, **c** eye gaze, **d** filtered SIFT

includes extensions, such as additional ways of constructing neighboring matrix W , a new way of updating the underlying model, a more comprehensive evaluation and discussions on the system responsiveness to expert inputs, scalability and other implementation considerations. The interface design and the supported expert image manipulations are presented in Sect. 3. An expert-in-the-loop evaluation study is described in Sect. 6.

2 Paradigm initialization

The initial image grouping was learned from an offline collected expert dataset. To elicit expert data, 16 physicians were asked to inspect 48 medical images and describe the image content aloud toward a diagnosis, as if teaching a student who was seated nearby [14]. Their eye movements were recorded, as eye movement features highlight perceptually important image regions, which is especially useful in knowledge-rich domains [15]. In this paper, we use experts' eye fixation maps to filter image features (SIFT features [16]). See Fig. 2 for an example. A bag of visual words is created from the remaining image features, and each image is described by a histogram of the visual words. Physicians' verbal image descriptions were also recorded concurrently, as they provide insights into experts' diagnostic image understanding. Figure 3 shows a sample transcription. The medical concepts were extracted

from the transcriptions using MetaMap, a medical language processing resource [17, 18]. These concepts formed a high-dimensional feature space, in which each image is described by the occurrences of these medical concepts.

We choose a Laplacian sparse coding approach [19, 20] over latent semantic analysis (LSA) or latent Dirichlet allocation (LDA). LSA does not perform as well as Laplacian sparse coding to cluster images by object [21]. LDA is affected not only by an initial specification but also by the samples randomly generated at each iteration [3]. It does not support users to make incremental changes, due to the inconsistent results obtained from multiple runs.

To initialize an image grouping based on the features extracted from multiple modalities, we adopt a data fusion framework based on Laplacian sparse coding [13]. The objective function is presented in Eq. (1). Matrices $E \in \mathbb{R}^{n_e \times m}$ and $V \in \mathbb{R}^{n_v \times m}$ are eye gaze-filtered image features and verbal features, respectively (n_e being the number of visual words, n_v being the number of verbal features, and m being the number of images). This model provides flexibility to allow extra data modalities by adding terms like the first two in Eq. (1). The coefficient matrix $C \in \mathbb{R}^{k \times m}$ (k being the number of latent topics) stores the new image representations, each of which is a distribution of latent topics learned and stored in the basis matrices $P \in \mathbb{R}^{n_e \times k}$ and $Q \in \mathbb{R}^{n_v \times k}$. The matrices P and Q reveal the transformation from the original feature spaces to latent topics.

Fig. 3 A sample diagnostic narrative expressed by a dermatologist inspecting the medical image shown in Fig. 2a. (SIL) represents silent pause

(SIL) okay and so this is a classic in dermatology periungual warts uh (SIL) there are plaque like areas of uh thickened hyperkeratotic skin (SIL) uh they uh they show areas of hemorrhage kind of like the psoriasis does (SIL) but uh the uh distribution the thickness of the keratin (SIL) uh and this tendency to to (SIL) fissure centrally and the (SIL) thicker areas of the of the wart (SIL) uh is is kind of c- like characteristic (SIL) it's interesting that when you see this you almost always see (SIL) uh a person who has hangnails as well (SIL) and uh and that's because the virus gains entry through the defect in the hangnail (SIL) so periungual warts

Fig. 4 Image grouping interface (details of algorithms behind this interface are described in Sects. 4 and 5): Panel 1-a visualizes the image grouping before each round of expert image manipulation, and panel 1-b visualizes the resulting image grouping afterward. 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 Figs. 5, 6, 7, 8) 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; it can also be used to show experts the top related images for key term manipulations

$$\min_{P, Q, C \geq 0} \|E - PC\|_F^2 + \|V - QC\|_F^2 + \alpha \mathcal{G}(W, C) + \beta \mathcal{S}(C) \quad (1)$$

where $\mathcal{S}(\cdot)$ represents a sparsity constraint (l_1 -norm) and $\mathcal{G}(\cdot, \cdot)$ represents a graph-regularizer. These constraints form the Laplacian sparse coding that helps capture underlying semantics behind observations in both modalities [20, 22]. W is a neighboring matrix that indicates similarities between pairs of data instances. A multimodal variation of the *feature-sign search algorithm* is developed to selectively update some elements of each data instance to tackle the non-derivativeness of the l_1 -norm [22]. Since the sparse codes learned through general-purpose machine learning algorithms usually do not reflect ideal expert image understanding in a specific domain [23], we extend this framework with extra constraints from expert knowledge to improve semantic image representations.

3 Interface design

The initial image grouping purely based on offline collected expert data is first visualized in the *Older Image Organization* in Fig. 4 (panel 1-a) 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 Fig. 9). Second, experts can click the buttons in Fig. 4 (panel 3) to compare the image grouping obtained when using different subsets of features, such as only primary morphology terms¹

¹ Primary morphology terms (PRI) and eight other categories of terms were identified by two highly trained dermatologists as *thought units* in an annotation study to label the stages in diagnostic reasoning [14]. Used here in this interface, these thought units can disclose the influence of each category of terms on the medical image grouping.

Fig. 5 Image groupings generated using primary morphology terms (PRI)

(Fig. 5), with that using the whole feature set, see Fig. 4 (panel 1-a).

Experts have a few options to improve the image grouping in each round. First, they can directly drag images toward or apart from each other in Fig. 4 (panel 1-a). The system parses such expert inputs and incorporates them for updating the neighboring graph-based regularizer (see Sect. 5.1). Second, experts can select a topic from the listbox in Fig. 4 (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 (see Sect. 5.2). Likewise, experts can select a topic and indicate the term(s), so that the system updates the topic-term distribution matrices (see Sect. 5.3). After experts interact with the interface using any option, the image grouping in the previous round is copied to Fig. 4 (panel 1-a), and the improved one is shown in Fig. 4 (panel 1-b). In each round, both image groupings are visualized following the approaches discussed in Sect. 4.

4 Visualizing image groups

To comprehensively visualize the image grouping, our interface presents both a *graph view* shown in Fig. 4 (panel 1) and a *matrix view* shown in Fig. 4 (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 [24]. 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) [3]. 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 similarities² in

² 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.

Fig. 6 Image groupings generated using correct diagnosis terms (CD)

Fig. 7 Image groupings generated using body part terms (BOD)

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 Fig. 9. Different neighbors may share different features with the selected image. We visualize these shared features to enhance expert users' overall understanding of the image repository.

In order to help the expert users perceive how the underlying algorithm uses the concepts and features and how each subset of features contributes to the overall image grouping, we provide additional visualization functionality to allow them looking into various grouping results, e.g., based on primary morphology terms (PRI; Fig. 5), correct diagnosis terms (CD; Fig. 6), body part terms (BOD; Fig. 7), and image features filtered by eye-tracking features (ET; Fig. 8). The clusters result from *face*, *leg*, *foot*, etc. in Fig. 7 can be recognized easily. Clusters in Figs. 5, 6, 7, and 8 also make sense to experts. Since experts may have special preferences of features when grouping the images, such feature-specific visualizations are a foundation for interactive feature selection algorithms. More discussions are in Sect. 6.

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 Fig. 10 for a magnified matrix view. The matrix view provides a global indexing of pairwise image similarities in the learned representation.

Fig. 8 Image groupings generated using image features filtered by eye-tracking features (ET)

5 Expert knowledge constraints

There are mainly two approaches in prior studies allowing user interactions to help improve learning a model: document-level interactions [3,6], or topic/cluster-level interactions [2,4]. In our scenario, to improve medical image grouping, the documents are images. To develop this interface, we prefer image document-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 a multimodal expert dataset are not easily visualizable nor interpretable by physicians. Below are three functions in the

Fig. 9 An example visualization of an image and its high-dimensional close neighbors. The target image is shown in the *upper left quarter*, and its top 3 close neighbors in the learned topic space 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 perceptually important image features are also ranked, and the top ones are marked in both the target image and its neighbors. The colors of the markers differentiate the image pairs. (Images courtesy of Logical Images, Inc.)

Second closest neighboring image

Shared topics' top terms are:
linear, vitiligo, hyper,
allergen, four, aspect, etc.

Closest neighboring image

Shared topics' top terms are:
vasculitis, papule, linear,
vitiligo, hyper, allergen, etc.

Third closest neighboring image

Shared topics' top terms are:
vasculitis, linear, vitiligo,
hyper, allergen, four, etc.

Fig. 10 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.)

interface for receiving expert inputs and updating the model, both to support image-level interactions.

5.1 Constraint on neighboring matrix, W

Let the images in the original feature space be denoted as $\mathbf{x}_1, \dots, \mathbf{x}_m$. A nearest neighbor graph G with m vertices can be constructed. A heat kernel weighting can be used to compute the element W_{ij} in the neighboring matrix W of the graph G [25].

$$W_{ij} = \begin{cases} e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|}{\sigma}}, & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are close} \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

where W_{ij} equals 1, if $\mathbf{x}_i$ and $\mathbf{x}_j$ are identical.

Similarly, other graph-weighting strategies can be adopted, such as *0–1 weighting* in Eq. 3 and *histogram intersection kernel weighting* in Eq. 4, depending on the feature attributes. They can be used together to achieve superior clustering result [26].

$$W_{ij} = \begin{cases} 1, & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are close} \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

$$W_{ij} = \begin{cases} \sum_{d=1}^D \min(\mathbf{x}_{di}, \mathbf{x}_{dj}), & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are close} \\ & (d \text{ represents each feature dimension}) \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

The interface can encode expert image manipulations as a transformation of the neighboring matrix W . This transformation is determined by multiple factors, including previous image grouping and experts' interpretation of it. The transformation of W can be simplified as $\mathcal{F}(\cdot, \cdot)$ in Eq. (5) and be considered as a constraint set by experts to guide the learning process.

$$\min_{P, Q, C \geq 0} \|E - PC\|_F^2 + \|V - QC\|_F^2 + \alpha \mathcal{G}(\mathcal{F}(W, K), C) + \beta \mathcal{S}(C) \quad (5)$$

where K denotes the set of images selected by an expert in Fig. 4 (panel 1-a). In this paper, 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 W_{ij} 's ($i, j \in K, i \neq j$) to be 1. Likewise, W_{ij} 's ($i, j \in K, i \neq j$) are all set to be 0, if the expert thinks they should be grouped differently. This rule is designed to update the neighboring matrix W in Eq. (2). Once all W_{ij} 's specified by the expert are updated, the algorithm will trigger the further learning process for the image representation C and the visual and verbal topics P and Q with respect to the objective function in Eq. (5).

5.2 Constraint on topic-coefficient matrix, C

Experts can also improve the image grouping through the task illustrated in Fig. 4 (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 C according to the constraint in Eq. (6).

$$\min_{P, Q, C \geq 0} \|E - PC\|_F^2 + \|V - QC\|_F^2 + \alpha \mathcal{G}(W, C) + \beta \mathcal{S}(C) \quad (6)$$

s.t. $C_{ij} = 0, i \in T, \text{ and } j \in L(i)$

where T is the collection of selected topics and $L(i)$ represents the least relevant images for topic i . In this paper, the element C_{ij} will be set to 0, if image j is selected to be least relevant to topic i . Once all C_{ij} 's are updated, the algorithm begins to learn P , Q , and C further with respect to Eq. (6).

5.3 Constraint on topic-basis matrix, 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 [27].

$$\min_{P, Q, C \geq 0} \|E - PC\|_F^2 + \|V - QC\|_F^2 + \alpha \mathcal{G}(W, C) + \beta \mathcal{S}(C) + \gamma_1 \|Q - Q_r\| \quad (7)$$

where 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 Q_r is changed due to an expert interaction, the underlying verbal topics in Q are learned toward it.

For the update of neighboring matrix W , that of the topic-coefficient matrix C , and that of the topic-basis matrix 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 [Fig. 4 (panel 1)] for the next interaction [24].

6 Evaluation and discussions

To evaluate the effectiveness of the paradigm per expert's objectives, a domain expert (co-author) 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

Table 1 Image grouping performances of fully automated learning and our paradigm

		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

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). The best performance is highlighted in bold

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 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 [28], 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 1, Tables 2, 3, and 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 [29]. 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 [26]. 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 points us to future work involving improved use of eye movement data, such as using the high-level behavioral patterns [30]. Experimental outcomes show that Laplacian sparse coding does not always or substantially beat some of 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 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 [13].

The presentation of image groupings could also be based on experts' trade-off between various factors, such as the pri-

Table 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

The best performance is highlighted in bold

Table 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

The best performance is highlighted in bold

Table 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

The best performance is highlighted in bold

mary 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 [31]. 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 [32]. 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 Sect. 5.1 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 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. 7, we plan to allow updating eye movement-filtered image patterns through a similar term $\gamma_2 \|P - P_r\|$ 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 [33]. 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 real-time 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 [34]. Moreover, we may use a different learning framework for fast model updates than that for model initialization [35].

7 Conclusions

This paper 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.

Acknowledgements This work was partially supported by NIH Grant 1R21 LM010039-01A1 and NSF Grant IIS-0941452. We would like to thank the participating physicians, the reviewers, and Logical Images, Inc. for images. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the official views of the NIH or the NSF.

References

1. Guo, X., Yu, Q., Li, R., Alm, C.O., Calvelli, C., Shi, P., Haake, A.: An expert-in-the-loop paradigm for learning medical image grouping. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 477–488. Springer (2016)
2. Choo, J., Lee, H., Kihm, J., Park, H.: iVisClassifier: an interactive visual analytics system for classification based on supervised dimension reduction. In: IEEE Symposium on Visual Analytics Science and Technology (VAST), pp. 27–34. IEEE (2010)
3. Choo, J., Lee, C., Reddy, C.K., Park, H.: UTOPIAN: user-driven topic modeling based on interactive nonnegative matrix factorization. *IEEE Trans. Vis. Comput. Graph.* **19**(12), 1992–2001 (2013)
4. Crossno, P.J., Dunlavy, D.M., Sheard, T.M.: LSAView: a tool for visual exploration of latent semantic modeling. In: IEEE Symposium on Visual Analytics Science and Technology (VAST), pp. 83–90. IEEE (2009)
5. Hu, Y., Boyd-Graber, J., Satinoff, B., Smith, A.: Interactive topic modeling. *Mach. Learn.* **95**(3), 423–469 (2014)
6. Lee, H., Kihm, J., Choo, J., Stasko, J., Park, H.: iVisClustering: an interactive visual document clustering via topic modeling. In: Zhang, H.R., Chen, M.(eds.) *Computer Graphics Forum*, vol. 31, pp. 1155–1164. Wiley (2012)
7. Johansson, S., Johansson, J.: Interactive dimensionality reduction through user-defined combinations of quality metrics. *IEEE Trans. Vis. Comput. Graph.* **15**(6), 993–1000 (2009)
8. Yang, J., Hubball, D., Ward, M.O., Rundensteiner, E.A., Ribarsky, W.: Value and relation display: interactive visual exploration of

- large data sets with hundreds of dimensions. *IEEE Trans. Vis. Comput. Graph.* **13**(3), 494–507 (2007)
9. Zhou, C., Frankowski, D., Ludford, P., Shekhar, S., Terveen, L.: Discovering personal gazetteers: an interactive clustering approach. In: *Proceedings of the 12th Annual ACM International Workshop on Geographic Information Systems*, pp. 266–273. ACM (2004)
 10. Guo, D., Peuquet, D.J., Gahegan, M.: Iceage: interactive clustering and exploration of large and high-dimensional geodata. *GeoInformatica* **7**(3), 229–253 (2003)
 11. Amershi, S., Fogarty, J., Weld, D.: Regroup: interactive machine learning for on-demand group creation in social networks. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 21–30. ACM (2012)
 12. Vaidyanathan, P., Pelz, J., Li, R., Mulpuru, S., Wang, D., Shi, P., Calvelli, C., Haake, A.: Using human experts' gaze data to evaluate image processing algorithms. In: S. Hemami, T.N. Pappas (eds.) *10th IVMSP Workshop: Perception and Visual Signal Analysis*, pp. 129–134. IEEE (2011)
 13. Guo, X., Yu, Q., Li, R., Alm, C.O., Haake, A.R.: Fusing multimodal human expert data to uncover hidden semantics. In: *Proceedings of the 7th Workshop on Eye Gaze in Intelligent Human Machine Interaction: Eye-Gaze and Multimodality*, pp. 21–26. ACM (2014)
 14. McCoy, W., Alm, C.O., Calvelli, C., Li, R., Pelz, J.B., Shi, P., Haake, A.: Annotation schemes to encode domain knowledge in medical narratives. In: N. Ide, F. Xia (eds.) *Proceedings of the 6th Linguistic Annotation Workshop*, pp. 95–103. Association for Computational Linguistics, Stroudsburg (2012)
 15. Guo, X., Li, R., Alm, C., Yu, Q., Pelz, J., Shi, P., Haake, A.: 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 and Applications*, pp. 275–278. ACM (2014)
 16. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **60**(2), 91–110 (2004)
 17. Aronson, A.R.: Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program. In: *American Medical Informatics Association Annual Symposium Proceedings*, pp. 17–21. AMIA (2001)
 18. Guo, X., Yu, Q., Alm, C.O., Calvelli, C., Pelz, J.B., Shi, P., Haake, A.R.: From spoken narratives to domain knowledge: mining linguistic data for medical image understanding. *Artif. Intell. Med.* **62**(2), 79–90 (2014)
 19. Gao, S., Tsang, I.W., Chia, L.T., Zhao, P.: Local features are not lonely—Laplacian sparse coding for image classification. In: *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3555–3561. IEEE (2010)
 20. Zheng, M., Bu, J., Chen, C., Wang, C., Zhang, L., Qiu, G., Cai, D.: Graph regularized sparse coding for image representation. *IEEE Trans. Image Process.* **20**(5), 1327–1336 (2011)
 21. Cai, D., Bao, H., He, X.: Sparse concept coding for visual analysis. In: *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2905–2910. IEEE (2011)
 22. Lee, H., Battle, A., Raina, R., Ng, A.Y.: Efficient sparse coding algorithms. In: Schölkopf, B., Platt, J.C., Hofmann, T. (eds.) *Advances in Neural Information Processing Systems 19*, pp. 801–808. MIT Press, Cambridge, MA. <http://papers.nips.cc/paper/2979-efficient-sparse-coding-algorithms.pdf> (2007)
 23. Holzinger, A.: 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*, pp. 319–328. Springer (2013)
 24. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**(2579–2605), 85 (2008)
 25. Belkin, M., Niyogi, P.: Laplacian eigenmaps and spectral techniques for embedding and clustering. In: *NIPS*, vol. 14, pp. 585–591 (2001)
 26. Wang, J.J.Y., Bensmail, H., Gao, X.: Multiple graph regularized nonnegative matrix factorization. *Pattern Recognit.* **46**(10), 2840–2847 (2013)
 27. Choo, J., Lee, C., Reddy, C.K., Park, H.: Weakly supervised non-negative matrix factorization for user-driven clustering. *Data Min. Knowl. Discov.* **29**(6), 1598–1621 (2015)
 28. El-Hamdouchi, A., Willett, P.: Comparison of hierarchic agglomerative clustering methods for document retrieval. *Comput. J.* **32**(3), 220–227 (1989)
 29. Srivastava, N., Salakhutdinov, R.R.: Multimodal learning with deep boltzmann machines. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q., (eds.) *Advances in Neural Information Processing Systems 25*, pp. 2222–2230. Curran Associates, Inc., Red Hook, NY. <http://papers.nips.cc/paper/4683-multimodal-learning-with-deep-boltzmann-machines.pdf> (2012)
 30. Li, R., Shi, P., Haake, A.R.: Image understanding from experts' eyes by modeling perceptual skill of diagnostic reasoning processes. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2187–2194 (2013)
 31. Wang, J.Y., Almasri, I., Gao, X.: Adaptive graph regularized non-negative matrix factorization via feature selection. In: *2012 21st International Conference on Pattern Recognition (ICPR)*, pp. 963–966. IEEE (2012)
 32. Le, T., Lauw, H.W.: Semantic visualization with neighborhood graph regularization. *J. Artif. Intell. Res.* **55**, 1091–1133 (2016)
 33. Thomaz, A.L., Breazeal, C.: Transparency and socially guided machine learning. In: *5th International Conference on Development and Learning (ICDL)* (2006)
 34. Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online learning for matrix factorization and sparse coding. *J. Mach. Learn. Res.* **11**, 19–60 (2010)
 35. Fuchs, G., Stange, H., Samiei, A., Andrienko, G., Andrienko, N.: A semi-supervised method for topic extraction from micro postings. *IT* **57**(1), 49–56 (2015)