LATENT SEMANTIC INTERMEDIATE REPRESENTATION FOR MEDICAL IMAGE RETRIEVAL

Jianhan Mei[1] *Meng Yang*[1] *Di Deng*[2] *Rong Deng*[3]

SHENZHEN University[1] XIAMEN University[2] Waseda University[3]

ABSTRACT

The intermediate representation method has been adopted to bridge the semantic gap. However, the semantic information can not be well extracted in existing intermediate representation. In this paper, a latent semantic intermediate representation (LSIR) is proposed to extract the common latent semantic concepts. Especially from training data, we jointly learn its collective matrix factorization (CMF) and intermediate representation, where latent semantic concepts are extracted via CMF with the guidance of supervised intermediate representation. The latent semantic level which is added between the source data and the intermediate representation can help to select useful concepts. The latent semantic concepts are more expressive than original concepts so that the representation based on latent semantic concepts can be more precise. For medical image retrieval application, we test our algorithm on the ImageCLEF dataset. Thorough experimental results show the effectiveness of the proposed scheme.

Index Terms— Image retrieval, intermediate representation, collective matrix factorization

1. INTRODUCTION

Medical image is playing a more important role in clinical diagnosis. In [1], interactive search-assisted decision support (ISADS) which helps doctors to make better decisions by utilizing retrieval technology is introduced. Specially in medical image application, such systems face the problem of retrieving semantic similar medical images from large repositories. Recent years, research on content-based image retrieval (CBIR) provides novel options for solving such problem.

Research on image retrieval has long been faced with the challenge of semantic gap between low-level features and semantic content description of images [2] [3] [4]. In medical application, the semantic similarity is not simply defined in terms of superficial image characteristics, but in a medically relevant sense [1]. Various approaches have been proposed to bridge the semantic gap in decade. With the wide application of local features, bag-of-visual-words (BoVW) method has become a classical visual description technique. Based on BoVW framework, Fisher vector (FV) [5] and its simplified non-probabilistic version, vector of locally aggregated descriptors (VLAD) [6], are proposed, which combines the strengths of generative and discriminative approaches. Aforementioned representation methods remains aliased information between classes which are not relevant semantically. As showing in Fig. 1, the semantic gap performs as the similarity of representation vectors from different categories and the difference of representation vectors from same category. To construct better representation, [7] [8] build new feature spaces via descriptor learning. In [1] [9] [10] [11] [12] [13], the distance metric learning algorithms are studied, which optimizes the distance metric of proximity measure to improve image similarity search in CBIR system.

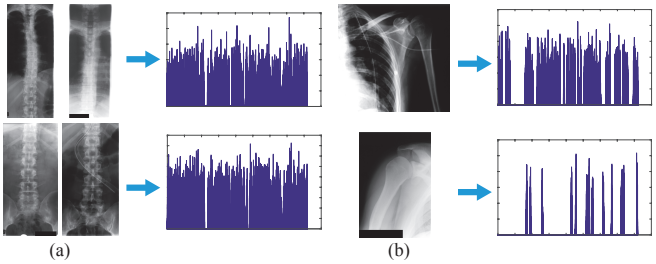

Fig. 1. (a) Samples from different categories have similar aggregating representation vectors. (b) Samples from same category have different aggregating representation vectors.

To find the semantically relevant representation of images, a linear dimensionality reduction is used to solve multi-label classification problem [14]. Following it, a classifier-specific intermediate representation method, semantic analysis via intermediate representation (SAIR), is proposed in [15]. Inspired by [16] [17] which uses collective matrix factorization for multi-view hash, we propose a new intermediate representation algorithm which learns latent semantic feature via collective matrix factorization for medical image retrieval. The proposed method has following characteristics:

1) Using collective matrix factorization to extract latent semantic concepts, the proposed algorithm follows an assumption that the interlinked data should have the same latent semantic representation.

2) The intermediate representation is built using the latent semantic concepts. The latent semantic concepts extraction and the intermediate representation are integrated into a joint

framework.

3) The intermediate representation level establishes the connection between the source data and the class tags. To some extent, it bridges the semantic gap.

For simplicity, we name our algorithm latent semantic intermediate representation (LSIR). For image retrieval in medical application, we test our algorithm on the X-Ray dataset from ImageCLEF and verify its high performance.

2. RELATED WORK

Over the last decade, subspace learning and distance metrics learning has been successfully used in many computer visual tasks such as classification, retrieval and object detection. We focus on the intermediate representation methods and would compare the kernel similarity learning method [9] later.

2.1. Intermediate Representation

In [15], the concepts-based representation is introduced. For video semantic understanding, a variety of semantic concept detectors have been built. Videos can be represented by the results of the concept classifiers. In [3] [4], c classifiers are trained from c classes. And each of the source data is represented using these c classifiers as an intermediate representation. However, methods in [3] [4] train the classifiers and refined representation independently. An integrated framework which learns a refined representation and classifiers jointly is proposed in [15]. However, the existing methods are almost based on specific concepts. For specific object, only some of the concepts are discriminative while others are comparatively useless or even noisy.

2.2. Problem Formulation of Intermediate Representation

For image retrieval application, suppose that there are n images which have been represented by using an aggregating representation method into d_1-dimensional vectors X=$\{x_1,...,x_n\}$. To code the vectors, a projection function $f(x_i)$ will be learnt referring to each vector's label y_i. And the codes in d_2 dimension can be represented as Q=$\{q_1,...,q_n\}$=$\{f(x_1),...,f(x_n)\}$. Finally, a similarity measure function $d(q_1, q_2)$ needs to be constructed and the final result is returned by ranking the value of the function.

Considering the linear case, the problem becomes to find a linear projection which connect the data to the label. And it can be also considered as a classical classification framework.

$$\min_{W} \|WX - Y\|_F^2 + \beta \|W\|_F^2 , \tag{1}$$

where, W denotes the linear projection matrix. And β is a tradeoff parameter. The final representation vector can be get from:

$$Q = WX, \tag{2}$$

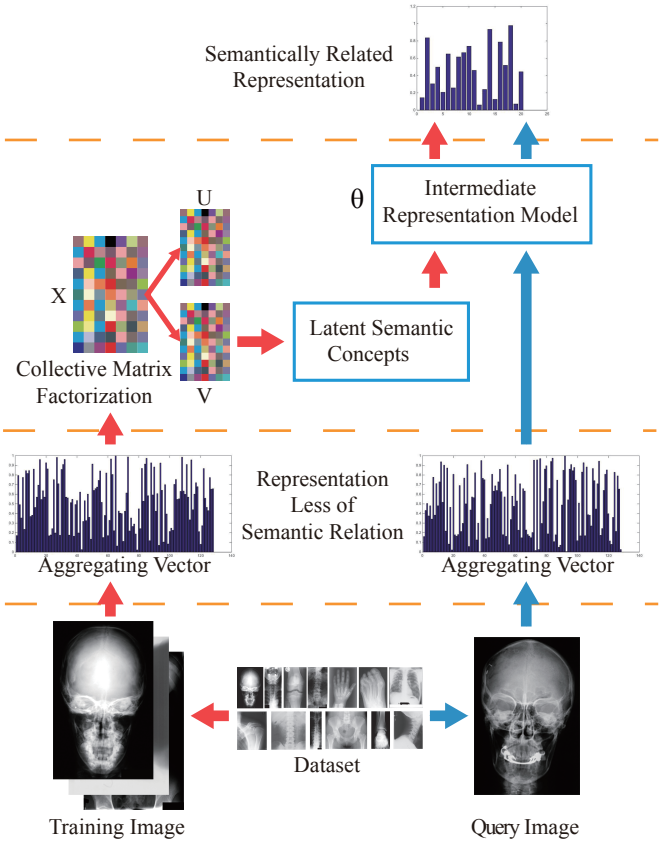

Fig. 2. Overview of the proposed latent semantic intermediate representation method. The red flow denotes the training step and the blue denotes the query.

For retrieval task, each low level representation vector of images is mapped to label space. And the final representation vector will be used as the query vector.

To build the intermediate representation, a projection function can be inserted between the data and the final representation. In [15], the representation model in linear situation is structured as fellow:

$$\begin{aligned} &\min_{W,\theta} \|W\theta X - Y\|_F^2 + \beta \|W\|_F^2 , \\ &s.t. \theta^T\theta = I \end{aligned} \tag{3}$$

where, θ are regarded as linear classifiers [15]. In [18], constraint $\theta^T\theta = I$ is added for avoiding arbitrary scaling of the intermediate representation and preserving as much information as possible.

3. INTERMEDIATE REPRESENTATION BASED ON LATENT SEMANTIC CONCEPTS

In [17], the collective matrix factorization (CMF) is used to learn latent semantic feature from source datasets. Here, we learn common latent semantic concepts and intermediate rep-

resentation jointly. The overview of the proposed method is presented in Fig. 2.

3.1. Intermediate Representation Based on Latent Semantic Feature

Following the assumption that the interlinked data should have the same latent semantic representation, the latent semantic concepts can be learnt from the source data by matrix factorization:

$$X = UV, \tag{4}$$

where, $U \in \mathbb{R}^{d\times k}$ and $V \in \mathbb{R}^{k\times n}$, and k is the number of latent factors. In V, each column vector $v_t (1 \leq t \leq n)$ is a latent semantic representation of the source data.

The factor U and V can be regarded as basis and coefficients respectively in the new feature subspace. The coefficients will be used to build the intermediate representation model. For preserving information from the source data and associating the semantic concepts with the labels, a reconstruction term will be used.

Similar with [17], the balanced constraint is added to maximum information on X. In this paper, it is used to construct the intermediate representation from the latent semantic concepts. Combining with (3), we build the overall objective function for latent semantic intermediate representation learning and it is given as follow:

$$\min_{W,\theta,V,U} G(W,\theta,V,U), \tag{5}$$

where,

$$\begin{aligned} G = &\|W\theta X - Y\|_F^2 + \|X - UV\|_F^2 + \alpha \|\theta X - V\|_F^2 \\ &+\beta(\|W\|_F^2 + \|\theta\|_F^2 + \|V\|_F^2 + \|U\|_F^2), \end{aligned} \tag{6}$$

where, α and β are tradeoff parameters. And the regularization term $\|\cdot\|_F^2$ is added to avoid overfitting. Term $\|X - UV\|_F^2$ is for matrix factorization and $\|\theta X - V\|_F^2$ is the reconstruction term.

And the final representation of the source data is given by:

$$Q = W\theta X, \tag{7}$$

where, Q are the final vectors of the source data. And the result of the retrieval can be given by ranking the distance among these vectors.

3.2. Analysis

In (6), samples are mapped to the label space via a linear projection matrix W. Before the final representation, the intermediate representation θ is added to represent the source data using semantic concepts. The source data is decomposed into basis and coefficients of the latent semantic concepts space. As illuminated in Fig. 2, coefficients V are used to build the intermediate representation model. After CMF, the source data X and the intermediate representation model θ are jointly used to reconstruct the coefficients V. The CMF step and the reconstruction step are executed iteratively so that the extracted semantic concepts are related to the labels and the intermediate representation model can preserve as much information of the source data as possible.

3.3. Solution

The optimization problem (5) is non-convex with four matrix variables W, θ, V, U. But it is convex with respect to any one of the four matrix variables while fixing the others [17]. The optimization problem (5) can be solved by solving following equations:

Fix θ, V, U, let $\frac{\partial G}{\partial W} = 0$, then obtain:

$$W = Y(\theta X)^T(\theta X(\theta X)^T + \beta I)^{-1}, \tag{8}$$

where, I is the identity matrix. Analogously, V and U can be updated by:

$$V = (U^T U + (\alpha + \beta)I)^{-1}(\alpha\theta X + U^T X), \tag{9}$$

and

$$U = XV^T(VV^T + \beta I)^{-1}. \tag{10}$$

Specially, fix W, V, U, let $\frac{\partial G}{\partial \theta} = 0$, then obtain:

$$\begin{aligned} &\beta(W^T W + \alpha I)^{-1}\theta + \theta X X^T \\ &= (W^T W + \alpha I)^{-1}(W^T Y X^T + \alpha V X^T), \end{aligned} \tag{11}$$

where, Eq. (15) is a standard Sylvester equation:

$$A\theta + \theta B = C, \tag{12}$$

where,

$$\begin{aligned} A &= \beta(W^T W + \alpha I)^{-1} \\ B &= XX^T \\ C &= (W^T W + \alpha I)^{-1}(W^T Y X^T + \alpha V X^T), \end{aligned} \tag{13}$$

The update steps for four variables should be repeated until convergency.

4. EXPERIMENT

Since our algorithm is based on aforementioned aggregation methods, experiments are conducted to evaluate the performance of these methods first. Then kernel similarity learning [9] and SAIR [15] are compared with our method. All the training methods are under same condition that uses same percentage of the training samples.

4.1. Datasets and Experiment Setup

There are two datasets which collect anonymous radiographs in our experiment. Both of them are provided by the ImageCLEF conference. One is from [9], which contains 20 categories with 7157 images. The other is from IRMA[1], which

[1] https://ganymed.imib.rwth-aachen.de/irma/

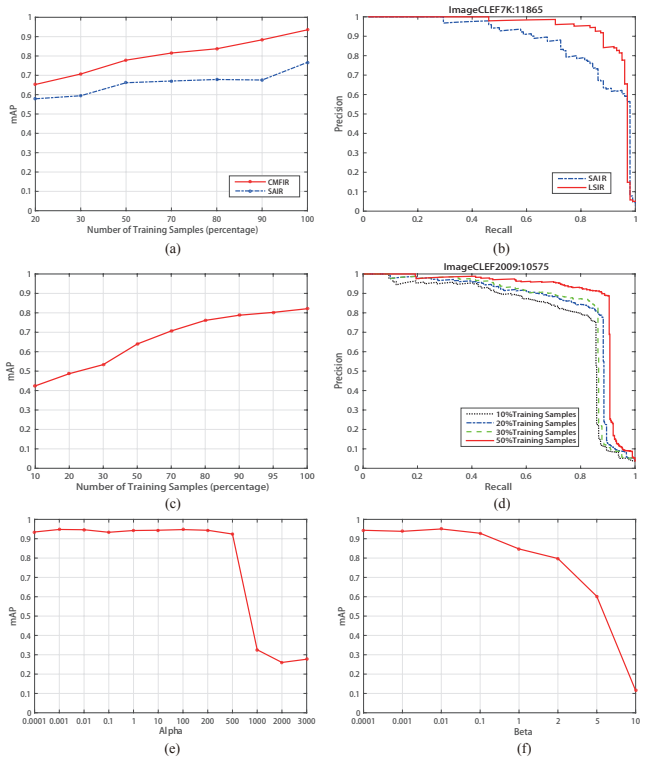

Fig. 3. (a) Performance with respect to the number of training samples. (b) P-R Curve of 11865 in ImageCLEF7K. (c) Performance on ImageCLEF2009. (d) P-R Curve of 10575 in ImageCLEF2009. (e) (f) Parameter sensitivity.

is classified according to the IRMA code. For dataset from IRMA, not all the annotation code is used in this paper. We classify the images according to first two of the anatomy code. After discarding and merging some classes, 21 categories are given from retained 12188 images. Finally, the comprehensive performance of the retrieval algorithm is evaluated by computing the mean average precision (mAP) [19]. For efficiency, the 16384 dimension VLAD is used to represent each image by aggregating SIFT features.

4.2. Performance Analysis

As illustrated in TABLE I, the three aforementioned aggregating representation methods are tested first. Based on the aggregating representation, intermediate representation is added to improve its performance. Here, we compare SAIR [15] and LSIR. Results show that intermediate representation significantly improves the performance of aggregating representation. The mAP score is raised by nearly 40%. Comparing with SAIR, LSIR outperforms it by 10%. Also, the kernel similarity learning method which is proposed in [9] is tested. Under same condition that using 50% training samples, the mAP score of LSIR outperforms it by nearly 10%.

To study how the number of training samples affects the LSIR performance, we conduct an experiment which explores the relation between the percentage of training samples and the mAP score. As showing Fig. 3 (a), the mAP score raises when using more training samples. LSIR may have a bigger slope, but it also performs well when using less than 30% of the training samples. And in Fig. 3 (b), the P-R Curve of target 11865 is presented under the 50% training samples condition. Also, our algorithm is tested on the ImageCLEF2009 dataset. And the P-R curve is presented using different percentage of training samples. Comparing with traditional intermediate representation methods, the proposed representation needs less training samples to maintain high precision. The latent semantic concepts are the further abstraction of the original concepts, which can be regarded as the selection and merging of the original concepts. By contrast, the method which is based on latent semantic concepts is more robust with respect to the quality of the training samples.

Table 1. Performance Comparison on ImageCLEF7K

Method	Condition	mAP
BoVW	Vocabulary Size: 50K	0.317
FV	Vocabulary Size: 128 L2-normalization	0.269
VLAD	Vocabulary Size: 128 Intra-normalization	0.291
Kernel Similarity Learning	[9]	0.668
SAIR	50% Training Samples	0.661
LSIR	50% Training Samples	**0.778**

4.3. Parameter Sensitivity

In object function (5), there are two tradeoff parameters α and β. To evaluate how the parameters affect the retrieval performance, we conduct the parameter sensitivity experiment. Fig. 3 (e) (f) show the affection from α and β respectively. For the ImageCLEF7K dataset, the best mAP score is obtained when both of the two parameters are small. In our experiments, the parameters both set as 0.001.

5. CONCLUSION

In this paper, we proposed a novel model of latent semantic intermediate representation for medical image retrieval. Combining with collective matrix factorization, the intermediate representation which is based on latent semantic concepts is proposed. The proposed model has extracted a reliable latent semantic concepts by jointly doing collective matrix factorization and intermediate representation. The experimental results on ImageCLEF showing by mAP score present high accuracy of our algorithm.

6. REFERENCES

[1] L. Yang, R. Jin, L. Mummert, R. Sukthankar, A. Goode, B. Zheng, S. C. H. Hoi, and M. Satyanarayanan, "A boosting framework for visuality-preserving distance metric learning and its application to medical image retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 30 – 44, 2010.

[2] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 1349 – 1380, 2000.

[3] A. G. Hauptmann, R. Yan, W.-H. Lin, M. G. Christel, and H. D. Wactlar, "Can high-level concepts fill the semantic gap in video retrieval? a case study with broadcast news," *IEEE Transactions on Multimedia*, vol. 9, pp. 958 – 966, 2007.

[4] A. G. Hauptmann, M. G. Christel, and R. Yan, "Video retrieval based on semantic concepts," *Proceedings of the IEEE*, vol. 96, pp. 602 – 622, 2008.

[5] F. Perronnin, Y. Liu, J. Sanchez, and H. Poirier, "Large-scale image retrieval with compressed fisher vectors," in *CVPR*, 2010.

[6] H. Jegou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid, "Aggregating local image descriptors into compact codes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, pp. 1704 – 1716, 2011.

[7] J. Philbin, M. Isard, J. Sivic, and A. Zisserman, "Descriptor learning for efficient retrieval," in *ECCV*, 2010.

[8] K. Simonyan, A. Vedaldi, and A. Zisserman, "Learning local feature descriptors using convex optimisation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, pp. 1573 – 1585, 2014.

[9] H. Xia, S. C. H. Hoi, R. Jin, and P. Zhao, "Online multiple kernel similarity learning for visual search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, pp. 536 – 549, 2012.

[10] S. C. Hoi, W. Liu, M. R. Lyu, and W.-Y. Ma, "Learning distance metrics with contextual constraints for image retrieval," in *CVPR*, 2006.

[11] L.Si, R. Jin, S. C. Hoi, and M. R. Lyu, "Collaborative image retrieval via regularized metric learning," *Multimedia Systems*, pp. 34 – 44, 2006.

[12] J. E. Lee, R. Jin, and A. K. Jain, "Rank-based distance metric learning: An application to image retrieval," in *CVPR*, 2008.

[13] S. C. Hoi, W. Liu, and S.-E. Chang, "Semi-supervised distance metric learning for collaborative image retrieval and clustering," *ACM Transactions on Multimedia Computing, Communications, and Applications*, pp. 18:1 – 18:26, 2010.

[14] S. Ji and J. Ye, "Linear dimensionality reduction for multi-label classification," in *IJCAI*, 2009.

[15] Z. Ma, Y. Yang, N. Sebe, K. Zheng, and A. G. Hauptmann, "Multimedia event detection using a classifier-specific intermediate representation," *IEEE Transactions on Multimedia*, vol. 15, pp. 1628 – 1637, 2013.

[16] J. Zhou, G. Ding, and Y. Guo, "Latent semantic sparse hashing for cross-modal similarity search," *Proceedings of the 37th international ACM SIGIR conference on Research and development in information retrieval*, pp. 415 – 424, 2014.

[17] G. Ding, Y. Guo, and J. Zhou, "Collective matrix factorization hashing for multimodal data," in *CVPR*, 2014.

[18] E. Kokiopoulou and Y. Saad, "Orthogonal neighborhood preserving projections: A projection-based dimensionality reduction technique," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, pp. 2143 – 2156, 2007.

[19] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *CVPR*, 2007.