Image Annotation using Multi-scale Hypergraph Heat Diffusion Framework

Venkatesh N. Murthy[1], Avinash Sharma[2], Visesh Chari[3], R Manmatha[1]
[1] {venk,manmatha}@cs.umass.edu
[1] School of Computer Science, University of Massachusetts, Amherst, MA, USA.
[2] asharma@iiit.ac.in [3] visesh@gmail.com
[2][3] Centre for Visual Information Technology, IIIT Hyderabad, Hyderabad, India.

ABSTRACT

The task of automatic image annotation involves assigning relevant multiple labels/tags to query images based on their visual content. One of the key challenge in multi-label image annotation task is the class imbalance problem where frequently occurring labels suppress the participation of rarely occurring labels. In this paper, we propose to exploit the multi-scale behavior in hypergraph heat diffusion framework for the automatic image annotation task. The proposed novel technique enables to model the higher order relationship among images in the feature space and provides a multi-scale label diffusion mechanism to address the class imbalance problem in the data.

General Terms

Algorithms, Experimentation, Measurement

Keywords

Image annotation, hypergraph, Deep learning, heat kernel

1. INTRODUCTION

The task of automatic image annotation involves assigning relevant labels/tags to query images based on their visual content [8]. However, it is difficult to extract variety of visual concepts embedded into an arbitrary image. The traditional techniques extract a set of low-level and/or high-level features on both training and testing images and subsequently apply fusion of various classifiers on these features to achieve automatic image annotation. In particular, nearest neighbor techniques with metric learning have reported best performance among all [17]. Nevertheless, these methods do not exploit the global structure of the underlying data domain well as they only focus on the local neighborhood structure (with a certain risk of overfitting).

One of the key challenge faced by existing techniques is the presence of severe class imbalance problem in real datasets.The class imbalance phenomenon is attributed to skewed distribution of labels in practical scenarios where a subset of over-represented labels (frequently occurring labels) dominate.It is common to exploit this dominance for achieving better performance (F-score) by focusing on retrieval of over-representend labels. Fortunately, the idea of reporting unique number of labels used in annotating test data (i.e., $N+$) keeps a good check on this trend. Thus, traditionally used datasets of moderate size are still relevant (challenging enough and unsaturated) to the task of automatic image annotation. One obvious solution to class imbalance problem is to give higher importance to under-representend labels (although by not loosing too much accuracy on over-represented labels).

In this paper, we propose to solve the automatic image annotation task using the novel multi-scale hypergraph heat diffusion framework. This would enable us to first capture the higher order similarity among multiple images in the feature space and subsequently exploit the topology of the underlying hypergraph. Such topological analysis enables us to perform simultaneous diffusion of the training labels at multiple scales in the transductive setup thereby addressing the key problem of class imbalance (by diffusing the under-represented labels at relatively large scale).

Firstly, we model the higher order feature similarity among images using the nearest-neighbour hypergraph modelling. Here, we use Convolutional Neural Networks (CNN) features that are proven to be more powerful for recognition kind of vision tasks as compared to the traditional array of low and high level features [13]. *Secondly,* we compute the spectrum of the associated hypergraph Laplacian matrix and use it to derive the hypergraph heat-kernel matrix. *Thirdly,* we diffuse the training image labels using the heat-kernel matrix at multiple scales and infer the test labels. *Finally,* we show the empirical validation of the proposed technique and demonstrate superior performance in comparison to many of the existing techniques that use multiple features.

2. LITERATURE SURVEY

Substantial research efforts have been made to solve the inherently difficult task of automatic image annotation in the past decade. At one hand, generative methods based on topic and mixture model exist where each image sample is modeled as a mixture of topics over visual features and labels [21, 2]. On the other hand, discriminative methods try to learn label specific discriminative models [19, 18]. Some of the recent methods proposed to combine both generative and discriminative [10].

Interestingly, recent data-driven techniques that are based

on the combination of either generative or discriminative method coupled with nearest-neighbor (NN) approach have reported better results. These methods find visually similar training images for a given test image and transfer labels from those images. JEC [8] proposed to consider equal contributions from different features (mean of distances) while transferring annotations from NN to the test image. Tag-Prop [4] proposed to combine large number of features via metric learning along with label-specific models in the NN setup. 2PKNN [17] exploits the semantic neighborhoods besides metric learning and learns weights for combining different features. Recently proposed NMF-KNN approach in [6] fused multiple features using weighted multi-view non-negative matrix factorization in conjunction with NN approach. Overall, the NN based approaches mainly exploit the local structure of the underlying data domain. However, this data domain seems to be non-Euclidean due to multiple labels and visual concepts assigned per image.

The graph-based transductive learning have recently became popular in the semi-supervised learning community [23]. These methods facilitates: 1) a natural representation of underlying non-Euclidean data domains as graphs and 2) an intuitive notion of scale dependent diffusion of class labels over graph neighbourhood for classification or labelling tasks. However, the simple graph based techniques suffers from the problem of finding the best similarity kernel parameters in the feature space. Another classical limitation is that each edge can only represent a dyadic relationship, thereby losing the higher order relationships among images in the feature space.

Hypergraph learning [22] have been proposed to address shortcomings of the simple graph techniques. It models the higher-order relationships among more than two data points (images) by using the concept of *hyperedge*. An adaptive hypergraph learning method for transductive image classification was proposed in [20], where the hyperedges were generated by linking images and their nearest neighbors (for varying size of neighborhood). Nevertheless, this method was used for predicting single class label per test image and only considered a fix scale spectral representation of graphs.

3. PROPOSED METHOD

In this section, we provide details of key steps of the novel multi-scale hypergraph heat diffusion (HHD) framework.

3.1 Feature Extraction

Given an image, we resize it to 224×224 irrespective of their aspect ratio in order to make it compatible with pre-trained CNN (VGG-16), followed by extracting a 4096-dimensional feature vector using a pre-trained CNN on ILSVRC-2014 dataset by VGG team as described in [1][11]. ILSVRC-2014 consists of 1.2 million images which are manually annotated with the presence/absence of 1000 categories. The VGG network was designed to classify these categories. Features are computed by forward propagating a mean-subtracted 224×224 RGB image through eight convolutional layers and three fully connected layers using Caffe software [5].

Let $\mathbf{x}_i$ be the 4096-dimensional feature vector representing i^{th} image, the entire dataset consisting of n images (including both training and testing sets) can be represented as:

$$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n].$$

We can separately represent training set of images with $\mathbf{X}^{train}$ and test set of images with $\mathbf{X}^{test}$ such that:

$$\mathbf{X} = \{\mathbf{X}^{train} \cup \mathbf{X}^{test}\}.$$

3.2 Hypergraph Construction

Hypergraph modelling enables capturing more information by employing the hyperedges that can link multiple nodes. We adopted hypergraph construction from [20] where each image is modelled as a node. Each node has one (or multiple) corresponding hyperedge(s) which connects k NN nodes in the feature space (for varying values of k). Let,

$$\mathbf{\Pi} = [\mathbf{he}_1, \ldots, \mathbf{he}_p] \qquad (1)$$

be the incidence matrix of the NN hypergraph induced on the image feature set $\mathbf{X}$. Here, each hyperedge $\mathbf{he}_i = [he_i^1, \ldots, he_i^n]^T$ is an indicator vector of size n where each element $he_i^j = 1$ if node $\mathbf{x}_j$ participate in hyperedge $\mathbf{he}_i$ or else zero. Thus, multiple 1's suggest that respective nodes contribute to the same hyperedge. The total number of hyperedges i.e., p is multiple of n depending on if we induce one or multiple hyperedges per node.

3.3 Hypergraph Heat Diffusion (HHD) Framework

The heat-kernel is a symmetric, non-linear (exponential) family of kernel (analogous to Gaussian kernel) for non-Euclidean spaces represented as graphs and is used as a diffusion tool for multi-scale label or information diffusion on graphs [15]. In case of dyadic graphs, it is derived from the spectra (constituted by both eigenvalues & eigenvectors) of the graph Laplacian matrix [14]. Interestingly, the Laplacian for hypergraph was derived in [22] where it was shown to be analogous to simple graph Laplacian.

The Hypergraph Heat Diffusion (HHD) framework enables multi-scale (topological) analysis of hypergraphs and we propose to exploit this for addressing the class imbalance property in the image annotation task. Using the definition of hypergraph incidence matrix in Eq.1, the hypergraph Laplacian is subsequently defined as:

$$\mathbf{L} = \mathbf{I} - \left(\mathbf{D}_v^{-\frac{1}{2}} \mathbf{\Pi} \mathbf{W}_{he} \mathbf{D}_{he}^{-1} \mathbf{\Pi}^T \mathbf{D}_v^{-\frac{1}{2}} \right) \qquad (2)$$

where, $\mathbf{\Pi}$ is incidence matrix of the hypergraph, $\mathbf{D}_v$ is degree matrix of nodes (defined as $\mathbf{D}_v = diag(\sum \mathbf{\Pi})$), $\mathbf{D}_{he}$ is the degree matrix of the hyperedges defined as $\mathbf{D}_{he} = diag(\sum \mathbf{\Pi}^T)$ and $\mathbf{W}_{he}$ is the hyperedge weight matrix defined as $\mathbf{W}_{he} = diag(w_1, \ldots, w_p)$.

The $\mathbf{W}_{he}$ matrix can be used to enforce the relative significance of certain hyperedges over others by setting a larger values. The eigen-decomposition of the hypergraph Laplacian matrix $\mathbf{L}$ is written as: $\mathbf{L} = \mathbf{U} \Lambda \mathbf{U}^T$ where, $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_n]$ be the matrix formed by the eigenvectors of $\mathbf{L}$ matrix and, $\Lambda = diag(\lambda_1, \ldots, \lambda_n)$ be the diagonal eigenvalue matrix, together defines the Laplacian spectra.

The hypergraph heat diffusion framework can be derived using the associated Laplacian spectra. The $n \times n$ heat-kernel matrix for hypergraph can be computed as:

$$\mathbf{H}(t) = \mathbf{U} \exp(-\Lambda t) \mathbf{U}^T \qquad (3)$$

where, t act as the scale parameter which govern the heat diffusion. It is straightforward to show that the $\mathbf{H}(t)$ matrix is indeed a kernel matrix as it satisfies Mercer's kernel property, i.e., it is a real, positive semi-definite matrix with

	Feature	Corel-5K				ESP Game				IAPRTC-12			
Method	Visual	P	R	F	N+	P	R	F	N+	P	R	F	N+
JEC [8]	HC	27	32	29	139	22	25	23	224	28	29	29	250
CCD [12]	HC	36	41	38	159	36	24	29	232	44	29	35	251
KSVM-VT [18]	HC	32	42	36	179	33	32	33	259	47	29	36	268
MBRM [3]	HC	24	25	25	122	18	19	19	209	24	23	24	223
TagProp(σML) [4]	HC	33	42	37	160	39	27	32	239	46	35	40	266
FastTag [2]	HC	33	43	37	166	46	22	30	247	47	26	34	280
2PKNN+ML [17]	HC	44	46	45	191	53	27	35.7	252	54	37	44	278
SVM-DMBRM [10]	HC	36	48	41	197	55	25	34	259	56	29	38	283
NMF-KNN*	HC	22	32	26	150	29	23	26	268	-	-	-	-
JEC	VGG-16	31	32	32	141	26	22	24	234	28	21	24	237
TagProp (σ)	VGG-16	30	35	32	149	31	28	30	246	38	30	34	260
2PKNN	VGG-16	33	30	32	160	40	23	29	250	38	23	29	261
HHD	VGG-16	31	49	38	194	35	36	34	257	32	44	36	280

Table 1: Results comparison. P: Average Precision, R: Average Recall, N+: Number of distinct labels that are correctly assigned to at least one test image. HC: bunch of hand-crafted/engineered features, VGG-16: CNN or deep learning feature.

real and positive eigenvalues ($\exp(-\Lambda t)$ for $t \geq 0$) (see [22] for details).

3.4 Multi-scale Label Diffusion & Inference

An explicit control over the scale of diffusion allow one to diffuse over-represented and under-represented labels separately, thereby addressing the prevalent class imbalance scenario in real data. In case of HDD framework, the scale of diffusion is governed by the parameter t (see Eq. 3). The value of t can vary from 0 to ∞. A small scale diffusion (t closer to zero) would enforce the label diffusion in the smaller neighbourhood while a large scale diffusion (t closer to infinity) would enforce the label diffusion in a very large neighbourhood.

Let, the ground truth labels for both training and test labels can be represented as:

$$\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_m]$$

where each $\mathbf{y}_i$ is an n-dimensional indicator vector (0's and 1's) for the multi-label annotation setup with label vocabulary of size m. Let $\mathbf{Y}^{train} \subset \mathbf{Y}$ be the set of known labels (training set) and $\mathbf{Y}^{test}$ be the complementary set of unknown labels (testing set). Thus, a scale dependent label diffusion can be accomplished as:

$$\mathbf{Y}_t = \mathbf{H}(t)\mathbf{Y}. \tag{4}$$

Let $\mathbf{Y}^{OR} \subset \mathbf{Y}$ be the subset of labels from over-represented class and $\mathbf{Y}^{UR}$ be the complementary set of under-represented class labels. The multi-scale (ms) diffusion to address the class imbalance problem can be achieved by diffusing over-represented labels at t_{small} and under-represented labels at t_{large} as:

$$\mathbf{Y}^{OR}_{t_{small}} = \mathbf{H}(t_{small})\mathbf{Y}^{OR}, \tag{5}$$
$$\mathbf{Y}^{UR}_{t_{large}} = \mathbf{H}(t_{large})\mathbf{Y}^{UR}, \tag{6}$$

These diffused label matrices can further be combined to

$$\mathbf{Y}_{ms} = \mathbf{Y}^{OR}_{t_{small}} \bigcup \mathbf{Y}^{UR}_{t_{large}}$$

.

Finally, we select the subset of multi-scale diffused labels for test set images (i.e., $\mathbf{Y}^{test}_{ms} \subset \mathbf{Y}_{ms}$) apply multi-label inference by taking the q largest entries of each row of $\mathbf{Y}^{test}_{ms}$

for inferring q labels for each test image. However, before inferring test image labels, we propose to normalize $\mathbf{Y}^{test}_{ms}$ with $L1$-normalization using $\mathbf{Y}^{train}_{ms}$. This type of normalization further helps in addressing the class imbalance problem.

4. EXPERIMENTS & RESULTS

In order to have a fair comparison to the previous methods, we use the same train and test split provided in [4] for the three datasets. Each test image is annotated with a fixed number of five labels, this makes it almost impossible to get perfect precision and recall score on these three datasets because each image has highly varying number of labels associated with it.

4.1 Dataset and Evaluation

We evaluate on four standard publicly available image annotation datasets - Corel-5k, ESP-Game, IAPRTC-12 and MIRFLICKR-25K. These datasets contain a variety of images like natural scene, game, sketches, transportation vehicles, personal photos and so on, thus making it a challenging task. We follow the standard evaluation metrics as reported in most of the previous work [10, 9, 17, 8, 3, 7].

4.2 Results

We evaluate the performance of the proposed method using CNN features on image annotation task and compare it with some of the existing methods in the literature. In addition, we also provide the effectiveness of using CNN features as opposed to 15 engineered/hand-crafted (HC) features (including both local and global features) in some of the existing methods like 2PKNN, TagProp and JEC. Experimental results provided in Table 1 are obtained using the standard evaluation metric [10].

We did not fully exploit the formulation by setting $W_{he} = I$, i.e., giving equal importance to all hyperedges. This was done intentionally, because we wanted to report generalized results and therefore a tuning which can be regarded as overfitting to datasets was avoided. Instead of using all eigenvectors of the Laplacian matrix, we used only 10% of the smallest eigenvectors for constructing a low rank heat kernel matrix for computation efficiency reasons. The scale parameters were empirically chosen for each dataset.
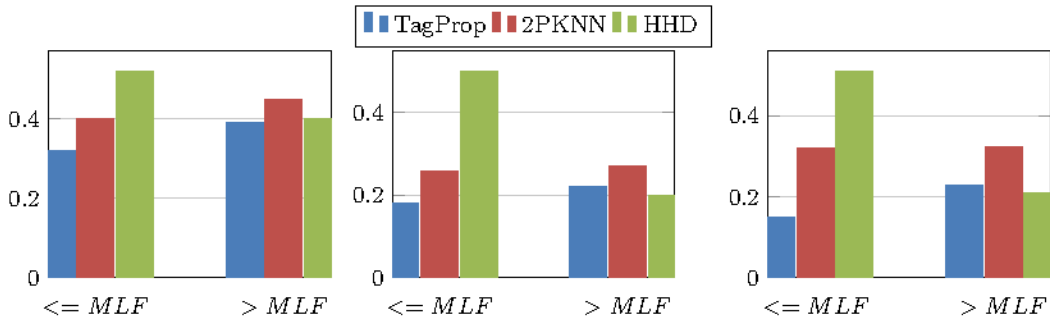
Figure 1: Annotation performance in terms of mean recall of labels in Corel 5k, ESP Game and IAPR TC12. On the Horizontal axis, the labels are grouped based on their frequency. The first bin corresponds to the 50% under represented labels and the second bin corresponds to the 50% over represented labels. *MLF*: Mean label frequency in the dataset.

For reporting the results of using CNN features in the existing methods, we implemented JEC method as described in [8] and for 2PKNN [17], we made use of the code provided by the authors (reported results are obtained using the following parameter values, $k = 4$ and $w = 1$). NMF-KNN* results are recomputed using the metric of computing mean recall and mean precision over all the words instead of just non-zero recall words (as originally reported in the paper [6]). Recomputed values are obtained as follows: $P_{new} = (P_{old} * (N+))/V_{size}$,, $R_{new} = (R_{old} * (N+))/V_{size}$ and $F_{new} = (2 * P_{new} * R_{new})/(P_{new} + R_{new})$.

Results on Corel-5k: From Table 1, we can clearly infer that the proposed method performs better than most of the existing methods when used with single CNN feature. Interestingly, the CNN features used in JEC, TagProp and 2PKNN methods performs better than some of the existing methods with HC features. However, their performance is poor in terms of both F and N+ measures when compared to our proposed method. This demonstrates that the improvement in the performance is attributed to our novel HHD framework and not just the feature alone.

We have also listed performance of existing methods on 15 HC features too though it does not amount to a fair comparison as these methods uses many features. For instance, our F and N+ measures are greater than the strong baseline TagProp(σML) method. Though our method underperformed slightly than the current state-of-the-art 2PKNN+ML method in terms of F measure, it is significantly better in terms of N+ measure (number of distinct labels used for annotation). Both 2PKNN+ML and TagProp(σML) methods are based on nearest neighbor approach employing metric learning to find an optimal combination of HC features. In terms of N+, it clearly indicates that our method provides a good generalization as opposed to metric learning techniques. For reporting the results using TagProp, we made use of the code provided by the authors [4].

Results on ESP-Game and IAPRTC-12: On both datasets, HHD provides a significant improvement over existing methods (JEC, 2PKNN) using a single CNN feature. In case of 15 HC features, we can observe that our HHD method performs almost similar to state-of-the-art 2PKNN method in terms of F measure and outperforms in N+ measure for ESP-Game. This demonstrates that our method effectively handles the imbalance data and the poor labeling problems. In case of IAPRTC-12, our method performance is competitive to TagProp(σML) and slightly lower compared to 2PKNN+ML in terms of F measure, but when it comes to N+ HHD method performs better than both the other methods. This suggests that hypergraph heat diffusion framework with the CNN feature is able to provide an effective solution for the image annotation task.

Results on MIRFLICKR: From Table 2, we can clearly see that proposed HHD method significantly outperforms both TagProp and SVM models in terms of **AP** measure. **AP** was computed as described in [16] per label. This clearly indicate that given a label our method is able to accurately retrieve most of the images which makes it suitable for real world applications and this is also indicates that the proposed method generalizes well compared to others.

Method	AP (Average Precision)
TagProp-Rank [16]	46.9
TagProp-Distance [16]	45.9
SVM [16]	52.0
HHD	**75.0**

Table 2: Experimental results of our proposed method with previously reported best scores on MIRFLICKR-25K.

4.2.1 Impact on Class Imbalance

Figure 1 demonstrates the performance of automatic annotation for both most frequent (>MLF) and least frequent labels ($\leq$MLF). Here, the performance is measured in terms of mean recall. For all three datasets, we can see that HHD method provides significant improvement for rare labels (under-represented label class)as compared to 2PKNN+ML & TagProp(σML) and with slight performance degradation for most frequent labels. This clearly indicates that multiscale diffusion in the hypergraph helps in striking the balance between retrieving the rare words and the most frequent words.

5. CONCLUSION & FUTURE WORK

The proposed novel image annotation technique exploits the multi-scale diffusion to address class imbalance problem and outperforms the majority of existing techniques through empirical evaluation on standard image annotation datasets.

As part of the future work, it will be interesting to explore how to exploit the label based semantic similarity among images in conjunction with visual similarity in the proposed HHD framework. One can also explore the adaptive techniques for finding optimal parameters in HHD framework.

6. REFERENCES

[1] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *British Machine Vision Conference*, 2014.

[2] M. Chen, A. X. Zheng, and K. Q. Weinberger. Fast image tagging. In *International Conference on Machine Learning*, pages 1274–1282, 2013.

[3] S. L. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, CVPR'04, pages 1002–1009, 2004.

[4] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *International Conference on Computer Vision*, pages 309–316, 2009.

[5] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678. ACM, 2014.

[6] M. M. Kalayeh, H. Idrees, and M. Shah. NMF-KNN: image annotation using weighted multi-view non-negative matrix factorization. In *Computer Vision and Pattern Recognition*, pages 184–191, 2014.

[7] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *in NIPS*. MIT Press, 2003.

[8] A. Makadia, V. Pavlovic, and S. Kumar. A new baseline for image annotation. In *European Conference on Computer VisionI*, pages 316–329, 2008.

[9] S. Moran and V. Lavrenko. A sparse kernel relevance model for automatic image annotation. *International Journal of Multimedia Information Retrieval*, 3(4):209–229, 2014.

[10] V. N. Murthy, E. F. Can, and R. Manmatha. A hybrid model for automatic image annotation. In *Proceedings of International Conference on Multimedia Retrieval*, pages 369:369–369:376, 2014.

[11] V. N. Murthy, S. Maji, and R. Manmatha. Automatic image annotation using deep learning representations. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 603–606.

[12] H. Nakayama. *Linear distance metric Learning for large-scale generic image recognition*. PhD thesis, The University of Tokyo, Japan, 2011.

[13] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf : An astounding baseline for recognition. In *Computer Vision and Patter Recognition*, pages –, 2014.

[14] A. Sharma. *Representation, Segmentation and Matching of 3D Visual Shapes using Graph Laplacian and Heat-Kernel*. PhD thesis, 2012.

[15] A. D. Szlam, M. Maggioni, and R. R. Coifman. Regularization on graphs with function-adapted diffusion processes. *Journal of Machine Learning Research*, 9:1711–1739, 2008.

[16] J. Verbeek, M. Guillaumin, T. Mensink, and C. Schmid. Image annotation with tagprop on the mirflickr set. In *Proceedings of the international conference on Multimedia information retrieval*, pages 537–546. ACM, 2010.

[17] Y. Verma and C. V. Jawahar. Image annotation using metric learning in semantic neighborhoods. In *Proceedings of the 12th European Conference on Computer Vision - Volume Part III*, ECCV'12, pages 836–849, Berlin, Heidelberg, 2012. Springer-Verlag.

[18] Y. Verma and C. V. Jawahar. Exploring SVM for image annotation in presence of confusing labels. In *British Machine Vision Conference*, 2013.

[19] X. Xu, A. Shimada, and R.-i. Taniguchi. Image annotation by learning label-specific distance metrics. In *Image Analysis and Processing âĂŞ ICIAP 2013*, volume 8156, pages 101–110, 2013.

[20] J. Yu, D. Tao, and M. Wang. Adaptive hypergraph learning and its application in image classification. *Transactions on Image Processing*, 21(7):3262–3272, 2012.

[21] S. Zhang, J. Huang, Y. Huang, Y. Yu, H. Li, and D. N. Metaxas. Automatic image annotation using group sparsity. In *CVPR*, pages 3312–3319, 2010.

[22] D. Zhou, J. Huang, and B. Schölkopf. Learning with hypergraphs: Clustering, classification, and embedding. In *Advances in Neural Information Processing Systems*, pages –, 2006.

[23] X. Zhu, A. B. Goldberg, R. Brachman, and T. Dietterich. *Introduction to Semi-Supervised Learning*. Morgan and Claypool Publishers, 2009.