RESEARCH ARTICLE OPEN ACCESS

Automatic Image Annotation Using Modified Multi-label Dictionary Learning

SREEDHANYA.S[1], CHHAYA.S.PAWAR[2]

[1](Computer Engineering, Datta Meghe College of Engineering, Mumbai University, Airoli, India)

[2] (Computer Engineering, Datta Meghe College of Engineering, Mumbai University, Airoli, India)

Abstract:

Automatic image annotation has attracted lots of research interest, and effective method for image annotation. Find effectively the correlation among labels and images is a critical task for multi-label learning. Most of the existing multi-label learning methods exploit the label correlation only in the output label space, leaving the connection between label and features of images untouched. In image annotation, a semi supervised learning which incorporates a large amount of unlabeled data along with a small amount of labelled data, is regarded as an effective tool to reduce the burden of manual annotation. But some unlabeled data in semi-supervised models contain distance that negatively affects the training stage. Outliers in the method can be over-fitting problem especially when a small amount of training data is used. In this paper, proposing an automatic image annotation method called modified MLDL with hierarchical sparse coding for solving these problems. This method prevents the over-fitting associated with the semi-supervised based approach by using sparse representation to maximizing the correlation between the data. Apply a Tree Conditional Random Field to construct the Hierarchical structure of an image. The result will be multi-label set prediction of a query image and semantic retrieval of images. Experiment results using LabelMe datasets and Caltech datasets confirms the effectiveness of this method.

Keywords — **Automatic image annotation, MLDL, Sparse representation, Semi-supervised learning (Semi CCA), Hierarchical representation.**

I. INTRODUCTION

With the large growth of web images, image annotation, this is beneficial to information management. Given an image, the goal of image annotation is to analyse its visual content and assign labels to it. Automatic image annotation is a promising research topic and is still an important issue in multimedia and computer vision fields, which has attracted much researcher's interest.

The objective of image annotation is to automatically annotate an image with appropriate labels, which reflect visual content in the image. Automatic image annotation is a key step towards semantic keyword based image retrieval, which is considered to be easy way for retrieving images on the web. It plays an important role in bridging the semantic gap between low-level features used to represent images and high-level semantic labels used to describe image content. With the increasing number of images in social network and on the sharing websites (Facebook, Flicker, and YouTube, etc.), there is a huge demand for automatic image annotation.

In this paper, conduct the multi-label dictionary learning using Hierarchical sparse coding, for enhancing the feature extraction capability and the performance. Most of the standard methods are based on a supervised labelling approach in order to achieve an exact classification. However, it has been pointed out that with this approach the training cost is extremely high because an enormous amount of training data must be labelled manually. To reduce the issue, a semi-supervised approach is used in this paper. The semi-supervised approach that inputs a large amount of non-labelled data for the training and not so much labelled data. So, it helps to improve the training accuracy without using a lot of labelled data.

II. PROBLEM DEFINITION

Today, due to the increasing growth of digital images and the need of managing and retrieving

them image annotation has become an important field in research. The aim of image annotation is to denoting the meanings and concepts with the images. But the manual annotation is become impossible, costly, and time consuming, so need to introduce automate the annotation process. So that the information and features extracted from the images do not always reflect the Image content and the semantic gap as the lack of coincidence between the information as the main challenge of automatic systems.

A. Main Objectives

Specifically the aim is:

- To bridging the semantic gap (maximizing the correlation) between low level features of the images and high level Semantic concepts to automatically annotate an image with appropriate keywords, such as multiple labels, which reflect visual content in the image.
- To enhance the relationship between labels and visual features.
- To improve the performance, speed and accuracy with reduce the computational cost.

From this section, the main challenge is to suppress the outliers between the visual data and the semantic concept of the image. The main aim of the project is to enhance the relationship between them and improve the performance of the system.

III. LITERATURE SURVEY

Automatic image annotation has been the subject of an intense level of research over the past decade, and a diverse set of methods have been presented to tackle the problem. Most of automatic image annotation methods can be broadly divided into four categories: (i) generative models, (ii) nearest neighbour models, (iii) discriminative models, and (iv) sparse coding models. In related work, we briefly present a review on existing image annotation methods.

A. S. Moran and V. Lavrenko, "Sparse kernel learning for image annotation," in Proceedings of International Conference on Multimedia Retrieval, ACM, 2014 [1]

Generative models mainly consist of mixture models and topic models. Mixture models usually define a joint distribution over image visual features and labels. To annotate a new image, mixture models compute the conditional probability of each label given the visual features of the image. A fixed number of labels with the highest probability are used for annotation.

In which, sparse kernel learning framework into the continuous relevance model and greedily selects an optimal combination of kernels. Here address this gap by formulating a sparse kernel learning framework for the CRM, dubbed the SKL-CRM that greedily selects an optimal combination of kernels. The kernel learning framework rapidly converges to an annotation accuracy that substantially outperforms a host of state-of-the-art annotation models. They made two surprising conclusions: firstly, if the kernels are chosen correctly, only a very small number of features are required so to achieve superior performance over models that utilize a full suite of feature types; and secondly, the standard default selection of kernels commonly used in the literature is sub-optimal, and it is much better to adapt the kernel choice based on the feature type and image dataset.

Advantages

- Comparatively less running time

Limitations

- The generative data may not be optimal for image annotation task.
- Many parameters using in these models, and the parameter estimation process is usually computationally expensive.

B. "Multi-label Dictionary Learning for Image Annotation "IEEE Trans Image Process 2016 March 31, Volume: 25, Issue 6, Yuan Jing, Fei Wu, Zhiqiang Li, Ruimin Hu.[3]

In this paper, describing multi-label learning approach, the method used is Multi-Label Dictionary Learning with label consistency regularization and partial-identical label embedding (MLDL), which conducts multi-label dictionary learning and partial-identical label embedding simultaneously. In the input feature space,

incorporate the dictionary learning technique into multi-label learning and design the label consistency regularization term to learn better representation of features. In the output label space, design the partial-identical label embedding, in which samples with the exactly same label set can cluster together, and samples with partial-identical label sets can collaboratively represent each other.

Advantages

- Showing better performance than existing methods.
- Higher the relationship between labels and visual features.
- Obtain effective for semantic retrieval and image annotation.

Limitations

- Lack of correlation
- More computation time needed

C. M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid, "TagProp: Discriminative metric learning in nearest neighbor models for image auto-annotation," in IEEE International Conference on Computer Vision, 2009.[2]

TagProp is a nearest neighbor model. TagProp is in short for Tag Propagation, a new nearest neighbor type model that predicts tags by taking a weighted combination of the tag absence/presence among neighbors. First, the weights for neighbors are either determined based on the neighbor rank or its distance, and set automatically by maximizing the annotations in a set of training images. With rank based weights the k-th neighbor always receives a fixed weight, whereas distance based weights decay exponentially with the distance. Combine a collection of image similarity metrics that cover different aspects of image content, such as local shape descriptors or global color histograms.

Advantages

- Tag prediction model is conceptually simple.
- Annotate images by ranking the tags for a given image, or can do keyword based retrieval by ranking images for a given tag.

Limitations

- Lack of correlation between image and label feature
- When the number of training examples is limited, the nearest neighbor-based models may fail.
- The similarity between training samples and query sample is computed only according to visual features of images.

D. "Multi-Instance Multi-Label Learning for Image Classification with Large Vocabularies" Oksana Yakhnenko Google Inc New York, NY, USA, Vasant Honavar Iowa State University Ames, IA, USA, 2014[4]

In this paper introducing a learning algorithm for Multiple Instance Multiple Label learning (MIML).The algorithm trains a set of discriminative multiple instance classifiers and models the correlations among labels by finding a low rank weight matrix thus forcing the classifiers to share weights. MIML learning is a generalization of supervised learning. Model is a discriminative model trained to maximize the probability of the labels presenting the image and minimize the probability of the labels absent from the image.

Advantages

- Easy-to-implement MIML learning algorithm that can be used to train MIML classifiers for image annotation on large datasets and in settings where the vocabulary of possible labels is large.
- MIML algorithms, is scalable to setting with large number of images and large vocabulary of possible labels.

Limitations

- Compare to semi supervised learning method, supervised learning method has low performance.
- With this approach the training cost is extremely high because an enormous amount of training data must be labeled manually.

TABLE I
STUDY OF EXISTING METHODS

Sl. no	Comparison of existing methods		
	Title	Advantages	Disadvantages

1	Sparse kernel learning for image annotation	Comparatively less running time	The generative data may not be optimal for image annotation task; Parameter estimation process computationally expensive.
2	Tag-Prop	Tag prediction model is conceptually simple. Annotate images by ranking the tags for a given image	Lack of correlation between image and label feature, When the number of training examples is limited, the nearest neighbor-based models may fail.
3	Mlml using descriptive method	Easy-to-implement MIML learning algorithm , large number of images and large vocabulary of possible labels.	Low performance, Training cost is extremely high.
4	Multi-Label Dictionary Learning with label consistency regularization and partial-identical label embedding (MLDL)	Annotate images with more correct labels, Annotate different images using the same label with relatively proper weights, Effective for semantic retrieval	Time complexity, Lack of accurate annotation

In Table II show a study on some notable and recent research work on automatic image annotation.

IV. PROPOSED SYSTEM

In this section, describes the proposed approach of Modified multi-label dictionary learning (MLDL) using Hierarchical sparse coding [5]. In the training stage, firstly describes the feature vector creation using the datasets. Then, describes the dictionary learning with Hierarchical sparse coding. The Hierarchical model used here is Tree conditional random field model (TCRF), this model is robust to scale variance**.** In the Testing stage, extract the feature value and then using the trained dictionary, calculating the score with the database dictionary score and maximum value selected from that. The performance of proposed method evaluated and compared with existing methods. The experiment results using two datasets confirm the performance of the proposed method.

A. Feature Vector Creation

Sparse representation based methods have interesting image annotation results, and the dictionary used for sparse coding plays a main role in it. The dictionary can be constructed by directly using the original training samples, whereas the original samples have much redundancy and noise that are affecting to prediction. To obtain an effective representation samples, we need to adaptively learn dictionaries.

The main feature descriptors used here are, SSIM, GIST, LBP, HOG, SIFT and Color descriptors. Histogram of oriented gradients (HOG) is the feature descriptors used for the purpose of object detection. This method using overlapping local contrast normalization for improved accuracy and it computed on a denze grid of uniformly spaced cells (local descriptors). GIST summarizes the gradient information (scales and orientations) for different parts of an image, which provides a rough description of the scene. Scale-Invariant Feature Transform in short SIFT, it detects and uses a larger number of features from the images, which reduces the contribution of the errors caused by local variation. Extracting uniform local binary pattern (LBP) from a greyscale image, it is a visual descriptors. The LBP features encode local Pattern information or texture information, which can use for tasks such as classification, detection, and recognition. Self similarity descriptors (SSIM), it is a local descriptor finding the structural similarity of the images. Divide each image into certain blocks to finding the similarity among them.

B. Hierarchical Sparse Coding

In consideration of the drawbacks in a supervised approach, propose an automatic image annotation method where an effective semi-supervised method, semi-CCA and sparse representation collaboratively suppress the outliers in image. The first step is to generate subspaces that maximize the correlation between image features and label features are

generated by semi-CCA, using a certain amount of labelled data and unlabeled data. Semi-CCA extends canonical correlation analysis (CCA), to avoid the over fitting of data when it has a few labelled training data. And apply a Regularized Orthogonal Matching Pursuit (ROMP), one of the available sparsing algorithms to suppress the outliers.

In base paper (MLDL), using only the sparse algorithm where the low-level features are extracted from the training images, such as the color feature and the texture feature. But, there is a problem in these methods in that they are not robust to the scale variance due to their inability to discriminate the features extracted from the local regions. Therefore, proposing a hierarchical representation in this paper. This hierarchical structured model, which is called TCRF, is robust to the scale variance.

Fig.1 shows the flow of the proposed method. In the training stage, first divide all training images including labelled and unlabeled data into hierarchical subregions using SWA (Segmentation by Weighted Aggregation) method. For each region in each layer, extract features and apply sparse-representation in semi-supervised learning to estimate class label confidence. Then, train the TCRF using the confidence and the label data in all layers to estimate the probability of co-occurrences within classes in a hierarchical structure. In the test stage as the same way, first apply SWA to obtain hierarchical subregions, and then project the regions into sub-space in each layer, and finally estimate the class label for each region using TCRF**.**

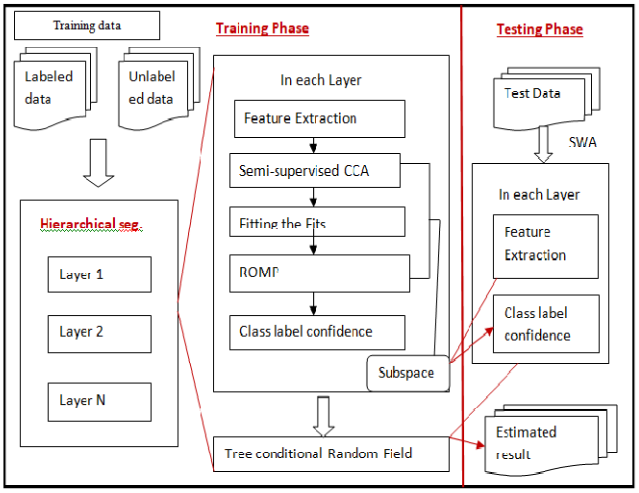

Fig. 1: System flowchart of proposed method

C. Training Dictionary

In Training stage, using multilayer sparse coding a dictionary is constructed. It is the concatenation of both sparse code feature descriptors (high level) and descriptors extracted from the trained image (low level). The main aim of this step is to producing a High dimensional feature set using hierarchical sparse coding. In this process the representation is transformed from low dimensional. But the entire image is converted to high dimensional feature descriptors.

The main objective of this work is to encode powerful local descriptor such as SIFT using a deep hierarchy for image categorization. Here feature coding is done using hierarchical sparse coding with dictionary learning. The output of this step will be a high dimensional feature vector and a learned dictionary.

D. Sparse Representation

A semi-supervised approach is used which uses unlabeled data instead of some of labelled data. But there still exist outliers in the unlabeled data. So here are introducing a new method that generates subspaces using the semi-supervised approach called Semi-supervised Canonical Correlation Analysis. This will suppresses such outliers using Regularized Orthogonal Matching Pursuit (ROMP) sparse algorithm. The Fig.2 shows the sparse representation.

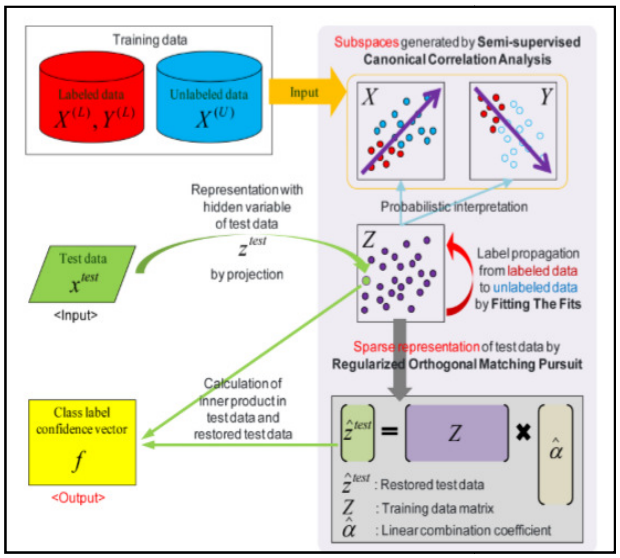

Fig.2: Flow of outlier suppression using sparse representation.

*1) **Semi - CCA:*** The semi-CCA is an extended version of Canonical Correlation Analysis (CCA) so that it substitutes unlabeled data for some labelled data. Both methods find the subspace that maximizes a correlation between two different types of features. In this paper, the correlation between an image and the appropriate labels are obtained.

Let $\{X^L, Y^L, X^U\}$ be the training datasets. Where X^L and Y^L are the labelled data and X^U represent unlabeled data. Then X and Y represent image feature and label feature. The aim of Semi-CCA or CCA is to find the optimum subspace that maximizes a correlation between projected x and y.

Semi-CCA can be expressed as the combination of two factors: CCA with labelled data and PCA with all data including unlabeled data. Therefore, the semi-CCA formulation is also obtained by combination of the two problems. The image feature and the label feature are connected via latent variables Z in the subspace.

*2) **Sparse co-efficient finding:*** In this step, suppress the effect of outliers that unintentionally appear when there is a large amount of unlabeled data, by employing sparse representation. If a sufficient amount of training data is prepared, an input image in the subspace can be represented as a linear combination of the training data. Our aim is to find sparse coefficients associated with each training data. Those entries are mostly zero, except for a few elements. This can be formulated as a minimizing problem with respect to a coefficient vector α and a training data matrix Z, Z test= Z α

But it is computationally difficult to find the optimum vector, because the coefficient vector α is in differentiable. So apply one of the popular greedy algorithms, Regularized Orthogonal Matching Pursuit (ROMP) to solve this problem. At the end, the test data can be restored by multiplying a training data and the obtained vector $\hat{\alpha}$ Then obtained $\hat{Z}test=Z\hat{\alpha}$

Where Z is the label matrix of the training samples, and $\hat{Z}test$ is the label set of the query sample. The top five labels with the largest values in $\hat{Z}test$ are considered as the annotations of the query sample. Therefore, multi-label classification can be realized by calculating all the confidences (C is the number of label classes). Then we use the best class label in all classes by taking the maximum. Then obtain the class label for one sub-pixel in a certain layer. After obtaining all the class labels for each sub-pixel in each layer, finally estimate the class label for each region, considering the hierarchical method and Tree Conditional Random Field (TCRF).Then we train TCRF using the confidence and label data in all layers to estimate the probability of co-occurrence with in class in hierarchical structure.

The features extracted are Gabor feature, position then area of a region. Then these features are concatenated to the feature vector obtained in the first step. Here we obtain a label feature vector. Which is a binary vector which indicates whether each label assigned in the subregion or not**.** Once the coefficient in the matrix Alpha obtained, the output of ranking function can be calculated using the training and testing data. The output of the ranking function gives ranking values for the corresponding image in the testing data.

*3) **Algorithm: Hierarchical sparse-coding:***

1. Input

The $\{\mathbf{X^L}, \mathbf{Y^L}, \mathbf{X^U}\}$ be the training datasets. Where X^L and Y^L are the labelled data and X^U represent unlabeled data. Then X and Y represent image feature and label feature, learned dictionary D, query image Z.

2. Subspace generation and sparse representation

Where $\mathbf{X^L} = \{\mathbf{X_n}\}\mathbf{N_{n=1}}$ and $\mathbf{Y^L} = \{\mathbf{Y_n}\}\mathbf{N_{n=1}}$ are N labeled data and $\mathbf{X^U} = \{\mathbf{X_m}\}\mathbf{M_{m=1}}$ is M unlabeled data. X and Y indicate image feature and label feature respectively.

$$r(w_x, w_y) = \frac{w_x^T S_{xy}^{(L)} w_y}{\sqrt{w_x^T S_{xx}^{(L)} w_x}\sqrt{w_y^T S_{yy}^{(L)} w_y}}$$

Here w_x and w_y are the projection vectors to the subspace from the original feature space x and y respectively. $\mathbf{S^L_{xx}}$ indicates each variance - covariance Matrix within the labelled data.

3. Solving sparse co-efficient and Image annotation

The coding coefficient vector α of the query image Z over D can be obtained

Z test= Z α

$$\min_{\alpha} ||\alpha||_{\epsilon} \quad s.t. \quad \mathbf{z}^{\text{test}} = \sum_{n=1}^{N+M} \alpha_n \mathbf{z}_n = \mathbf{Z}\boldsymbol{\alpha}$$

At the end, the test data can be restored by multiplying a training data and the obtained vector.

$$\hat{Z}test=Z\hat{\alpha}$$

Where Z is the label matrix of the training samples, and $\hat{Z}test$ is the label set of the query sample. The top five labels with the largest values in the result are considered as the annotations of the query sample.

4. Obtain Class label confidence vector

$$f_c = \frac{\mathbf{z}^{test^T}\hat{\mathbf{z}}_c^{test}}{||\mathbf{z}^{test}||_2||\hat{\mathbf{z}}^{test}||_2} = \frac{\mathbf{z}^{test^T}\mathbf{Z}_c\hat{\boldsymbol{\alpha}}_c}{||\mathbf{z}^{test}||_2||\hat{\mathbf{z}}^{test}||_2}$$

Here Zc is a training data matrix that only contains the data given the label c, and $\hat{\alpha}$ is a coefficient vector associated with the training data in Zc. The restoration ratio implies a confidence of the class fc. Therefore, multi-label classification can be obtained by calculating all the confidences (C is the number of label classes). However, we use the best class label in all classes by taking the maximum from that. Once the coefficient in the matrix Alpha obtained, the output of ranking function can be calculated using the training and testing data. The output of the ranking function gives ranking values for the corresponding image in the testing data.

E. Advantages of Proposed Method

- The experimental results using two datasets showed the effectiveness of our proposed method satisfactorily.
- Using sparse-representation-based approach, increased the accuracy, reduced time complexity, taking full advantage of semi-supervised learning.
- Applied hierarchical representation, obtained better results than when the non-hierarchical structured approach is used in base paper.
- Obtain both query images and annotation of images.
- The method is effective for semantic retrieval.

V. EXPERIMENTAL EVALUATION

Here describes the datasets used in this project. Then, evaluation and result of the proposed method that combines the sparse representation and hierarchical representation is in next.

A. Results and Discussion

For the image annotation experiments, used a LabelMe image data set and Caltech image data set. In which total 96 images has taken for the experiment and out of that 60 images for training and 36 images for testing. In experiment, the training and test are conducted using features in each subregion.

We compute the annotation Precision and Recall for testing set. Precision is the number of correctly annotated images divided by the total number of images annotated with the particular label (correctly or not), and it can be defined as Precision = A/ B. Recall is the ratio of the number of images correctly annotated with a given label and the number of images having the specific label in the ground-truth labels. Recall can be defined as Recall = A / C.

TABLE II: PERFORMANCE COMPARISON

Sl. no	Method	Precision	Recall
1	Tag-Prop	0.503472	0.402778
2	MIML	0.552771	0.44216
3	MLDL	0.518973	0.415179
4	Proposed Method	0.568824	0.457059

Table II shows the performance of existing methods and proposed method.

B. Image Annotation Using Proposed Method

In this section, shows the automatic image annotation results of the proposed method. Fig.3 shows the experimental result of the proposed automatic image annotation. Fig.4 shows the semantic retrieval of the images and in Fig.5 representing the running time among the existing method with proposed method.

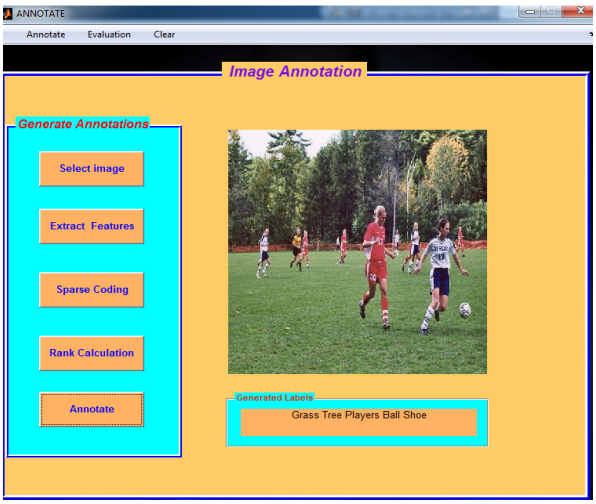

Fig. 2 Image annotation of one image

Fig. 4: Semantic retrieval of images using single label.

Method	Tag-Prop	MLDL	MIML	Proposed
Training	15m	160m	120m	15m
Testing	0.1m	.4m	.5m	.01m

Fig.5: Running time for each method

In semantic retrieval (Query search), evaluate the semantic retrieval performance. We can first use annotation methods, to assign top five annotations to testing images. For a given query label, then rank the images annotated with that label according to their annotation scores. In this project consider the three options are given: search by image, by single label and by multiple labels. The total labels are 30, out of total we can select any labels and common keywords with images will be output. The result is shown in Fig. 4.

V1. CONCLUSIONS

In this paper, proposed an Automatic image annotation approach, Modified MLDL using Hierarchical sparse coding, result in maximum correlation between image and label feature. The method combines semi-supervised approach, sparse representation, and hierarchical representation. Semi-supervised learning has the advantage to capture a true data distribution if we are giving only a small amount of labeled training data. We can Suppresses outliers, increase the accuracy in terms of using sparse representation. The experimental results using two datasets showed the effectiveness of the proposed method.

REFERENCES

1. *S. Moran and V. Lavrenko, "Sparse kernel learning for image annotation," in Proceedings of International Conference on Multimedia Retrieval, ACM, 2014.*

2. *M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid,"TagProp: Discriminative metric learning in nearest neighbor models for image auto-annotation," in IEEE International Conference on Computer Vision, 2009.*

3. *"Multi-label Dictionary Learning for Image Annotation" Xiao-Yuan Jing, Fei Wu, Zhiqiang Li, Ruimin Hu, Senior Member, IEEE, and David Zhang, Fellow, IEEE,2015.*

4. *"Multi-Instance Multi-Label Learning for Image Classification with Large Vocabularies"Oksana Yakhnenko Google Inc New York, NY, USA Vasant Honavar Iowa State University, USA,2014*

5. *"Hierarchical sparse representation for object recognition, Transaction on machine learning", volume 2, issue1,Toru Nakashika, Takeshi Okumura, Tetsuya Takiguchi.*

6. *"Efficient Multi-label Ranking for Multi-classLearning: Application to Object Recognition", Serhat S. Bucak, Pavan Kumar Mallapragada, Rong Jin and Anil K. Jain Michigan State University East Lansing, MI 48824, USA.*