A Dual-CNN Model for Multi-label Classification by Leveraging Co-occurrence Dependencies Between Labels

Peng-Fei Zhang[1], Hao-Yi Wu[2], and Xin-Shun Xu[1(✉)]

[1] School of Computer Science and Technology, Shandong University, Jinan, China
xuxinshun@sdu.edu.cn

[2] Faculty of Science, Engineering and Technology, University of Tasmania, Hobart, Australia

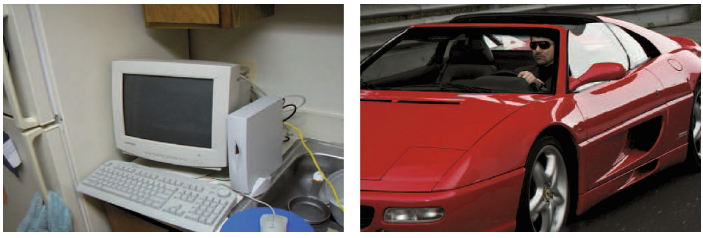
Abstract. In recent years, deep convolutional neural network (CNN) has demonstrated its great power in image classification. In real world, there are many images contain abundant contents so that they have multiple labels. Moreover, there are correlations between labels. Traditional deep methods for such data rarely take into account such correlations. In this paper, we propose a dual-CNN model, i.e., Dual-CNN model for Multi-Label classification (Dual-CNN-ML), which can make full use of the dependencies of labels to enhance classification performance. Specifically, we first obtain co-occurrence dependency matrix from training datasets; then, we merge the co-occurrence dependency matrix and image representation together; finally, we use the new representation to predict labels of samples. Extensive experiments on public benchmark datasets demonstrate that the proposed method obtains satisfying results and outperforms several state-of-the-art methods.

Keywords: CNN · Multi-label · Classification
Label co-occurrence dependencies

1 Introduction

The past few years have witnessed the great success of Deep Neural Network (DNN) in computer vision, e.g., recognition [15,16] and classification [1,19,21], etc. Many works have demonstrated that, given sufficient training data, CNN can achieve great performance in vision tasks [4,8,12]. For example, for image classification task, DNN models can obtain much better results than non-deep models. In early years, most deep models only consider the scenario that each sample contains one label [9,16,25]. However, as shown in Fig. 1, in real world, one image may be associated with multiple semantic labels, because an image normally abounds with rich semantic information, such as objects, parts, scenes, actions, and their interactions or attributes [19]. Thus, multi-label image classification is a more general and practical problem compared to single-label image

B. Zeng et al. (Eds.): PCM 2017, LNCS 10735, pp. 315–324, 2018.
https://doi.org/10.1007/978-3-319-77380-3_30



(a) keyboard, mouse, monitor (b) car, person

Fig. 1. Samples with multiple labels.

classification. As a result, multi-label problem has attracted much attention in machine learning and other communities. Correspondingly, many methods have been proposed to address this challenging problem [5,13,20]. In addition, there are correlations between labels. For example, as shown in Fig. 1(a), the mouse is very small and hard to be recognized. However, if a model could consider that mouse and keyboard often appear together, it could predict the mouse label with high probability. Similarly, the sample problem exists in Fig. 1(b).

Many works have shown that exploiting label correlations is helpful for multi-label image classification [6,11,17,18,22–24]. For example, in [6], the authors use Markov random fields to capture the correlations between labels. The method in [24] constructs co-occurrence matrix by means of harmonic mean on empirical conditional probabilities. The method in [14] trains a set of classifier to predict a label based on previous predicted labels and image representation. These method try to make full use of the correlations in labels; however, they only consider pairwise labels and ignore the complex structure of labels. That is to say, they cannot model complex high-order correlations between labels.

To deal with multi-label image classification, a typical approach in deep neural network is to convert multi-label image classification into multiple single-label image classification problems, so that this problem can be easily handled with existing single-label image classification methods. For example, Hypotheses-CNN-Pooling (HCP) [20] deep network extracts and selects a set of candidate object windows and then feeds them into CNN. Finally, it combines the representations as the final representation. Apparently, such work treats labels independently, and ignores the correlations in labels as shown in Fig. 1.

In order to model high-order label co-occurrence dependencies, in this paper, we utilize CNN to model label co-occurrence dependencies. The framework of the proposed model is shown in Fig. 2, which mainly contains two sub-networks: The image CNN accepts images as input to generate image embeddings; the matrix CNN takes label co-occurrence matrix as input to get matrix embedding. Then the two embeddings are imported into fusion layer to get the joint embedding. At last, by dealing with the joint embedding, we can predict labels for multi-label images. The matrix CNN can not only capture local feature which is the relation between labels, but also obtain global feature (correlations in all labels).

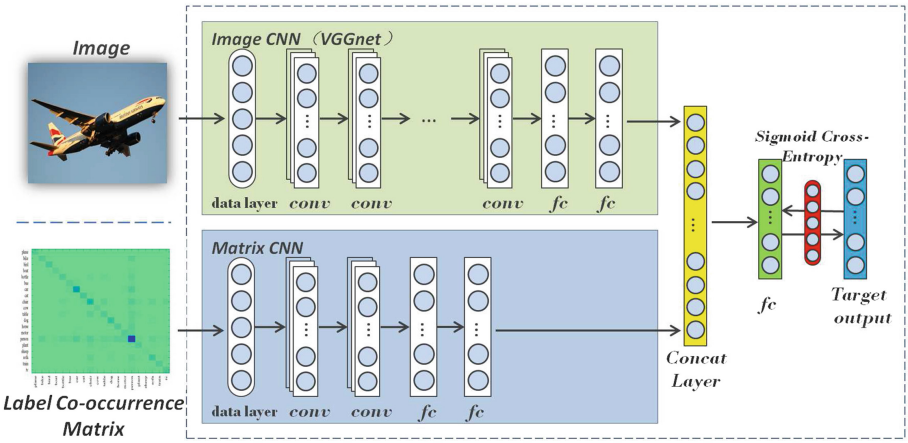


Fig. 2. Overview of the proposed model for multi-label classification task.

By this way, the proposed method can model the high-order label co-occurrence dependency. Moreover, the proposed model can be trained in end-to-end way.

Compared to previous methods, our main contributions are as follows:

- We propose an efficient end-to-end CNN framework to tackle multi-label image classification problem, and experiments on public datasets prove that our model can achieve better performance than previous works.
- By applying CNN on label co-occurrence matrix, our framework can efficiently model the high-order label co-occurrence dependencies and the complex relevance with images, which will be jointly learned to generate labels.
- We conduct extensive experiments on two benchmark multi-label datasets, e.g., Pascal VOC 2007 and VOC 2012. The results demonstrate that the proposed method can achieve competitive results compared with several state-of-the-art methods.

2 Proposed Model

The framework of our proposed model for multi-label image classification is shown in Fig. 2. The architecture contains three components: image CNN, matrix CNN and the convolution fusion and prediction layers. Specifically, the image CNN takes image as input and then extracts the image representation; the matrix CNN models label co-occurrence dependencies; at last the convolution fusion layers fuse the image and co-occurrence matrix representations together and obtain the new representation. At the same time, the joint representation is input to prediction layer to generate the labels. These three parts are connected tightly and trained together. The details of each part are given in the following subsections.

2.1 Image CNN

Compared with hand-crafted features, features extracted from the deep networks are more abstract, exact and meaningful. Thus, in this paper, we use deep neural network to extract representations for images. To do this, we employ the network in [16]. Suppose that $\mathcal{X} = \{x_i\}_{i=1}^n$ is a multi-label training set and x_i is the i-th sample, them the image representation can be denoted as:

$$u_i = \phi\left(w_t\left(F\left(x_i\right)\right) + b_t\right), \tag{1}$$

where w_t is the parameter for feature map, F represents the process of the network [16], ϕ is a nonlinear activation function, b_t is the bias. In this part, we slightly modify the network by chopping out the last layer, and add a nonlinear activation function behind it. For each image, the final representation is a 1000-dimension vector. By doing this, the benefits are that the nonlinear activation function can enhance the nonlinear property of model, on the other hand lower dimension representation can reduce the number of parameters that will be trained in fusion layer.

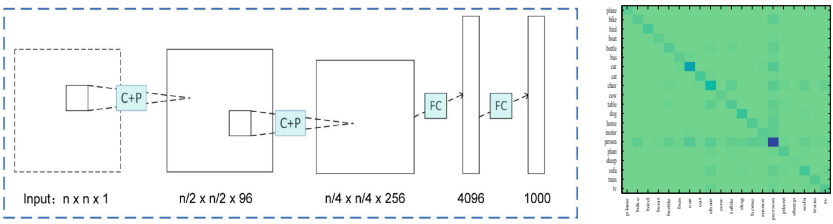
2.2 Matrix CNN

In CNN, the local receptive field of the convolution unit can perceive local information of the input and transport information to higher layers; then, higher layers further process these feature maps and obtain the global representation. Correspondingly, by means of these stacked convolution layers, we can capture the complex structure and content contained in an input. To model the high-order label co-occurrence dependencies, here we introduce a CNN into our framework to process the label co-occurrence matrix. By doing this, CNN can capture the local feature and the global feature of the input so that we could obtain the abundant contents about label co-occurrence dependencies. This process is shown in Fig. 3(a). Note that, in Fig. 3(a), C is the convolutional layer, P is the pooling layer, FC is the fully connected layer. Totally, it contain four layers: two convolutional layers and two fully connected layers. For a label co-occurrence matrix M, the convolutional process for nonlinear feature map of the l-th layer is:

$$\mathcal{T}^l = \phi\left(w_m^l \mathcal{T}^{l-1} + b_m^l\right), \tag{2}$$

where w_m^l denotes the parameters of feature map of the l-th layer, ϕ is the nonlinear activation function, $\mathcal{T}^{l-1}$ is the output of the $(l-1)$-th layer, b_m^l is the corresponding bias.

Note that our model is not the first work that use the label co-occurrence matrix. The method in [24] constructs co-occurrence matrix by means of harmonic mean on empirical conditional probabilities. Different from that in [24], our label co-occurrence matrix records the times of label co-occurrence. The reason is that we want to maintain the original information in labels so that CNN can capture such original information. For a training set, we count the times of



(a) Matrix CNN learning label co-occurrence dependencies.

(b) Label co-occurrence matrix.

Fig. 3. Matrix CNN learning label co-occurrence dependencies and corresponding label co-occurrence matrix extracted from Pascal VOC 2007.

appearance of pairwise labels as the elements of the label co-occurrence matrix, if the value of a element is too big, we divide matrix by a certain definite value.

To show that the above method can obtain the co-occurrence dependencies, we show the label co-occurrence matrix extracted from Pascal VOC 2017 by using the process above, which is shown in Fig. 3(b). In this figure, the color is deeper, the labels appear together more frequently.

2.3 Fusion and Prediction Layer

After obtaining the image and co-occurrence matrix presentations, we concatenate these two representations and use the fully connected layer to produce the new representation. We denote the new representation as $\mathcal{V} = \{v_i\}_{i=1}^{n}$, $v_i = \{v_i^k\}_{k=1}^{K}$, K is the number of category.

$$c_i = u_i \parallel \mathcal{T}, \tag{3}$$

$$v_i = w_f * c_i + b_f, \tag{4}$$

where u_i is the i-th image representation and $\mathcal{T}$ is the matrix representation, w_f and b_f are the parameters of feature map, c_i is the input of the fully connected layer.

Thereafter, we use the sigmoid cross entropy loss as the loss function for our multi-label classification task, which is formulated as:

$$\mathcal{E}_i = \sum_{k=1}^{K} \left[p_i^k \log \hat{p}_i^k + \left(1 - p_i^k\right) \log \left(1 - \hat{p}_i^k\right) \right], \tag{5}$$

where

$$\hat{p}_i^k = 1/\left(1 + exp\left(-v_i^k\right)\right), p_i^k \in [0, 1], \tag{6}$$

where $\mathcal{E}_i$ is the loss for i-th sample, $\hat{p}_i^k$ is the probability of that v_i belongs to the k-th category, p_i^k is the target probability.

Table 1. Results on Pascal VOC 2007

	Plane	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table
INRIA [7]	77.2	69.3	56.2	66.6	45.5	68.1	83.4	53.6	58.3	51.1	62.2
NUS [1]	82.5	79.6	64.8	73.4	54.2	75.0	77.5	79.2	46.2	62.7	41.4
AGS [2]	82.2	83.0	58.4	76.1	56.4	77.5	88.8	69.1	62.2	61.8	64.2
CNN-SVM [15]	88.5	81.0	83.5	82.0	42.0	72.5	85.3	81.6	59.9	58.5	66.5
I-FT [20]	91.4	84.7	87.5	81.8	40.2	73.0	86.4	84.8	51.8	63.9	67.9
HCP-1000C [20]	95.1	90.1	92.8	89.9	51.5	80.0	**91.7**	91.6	57.7	77.8	70.9
CNN-RNN [19]	96.7	83.1	94.2	**92.8**	61.2	82.1	89.1	94.2	**64.2**	83.6	70.0
Dual-CNN-ML	**96.9**	**94.1**	**96.5**	90.4	**71.7**	**90.1**	79.4	**96.1**	59.1	**84.4**	**73.4**
	Dog	Horse	Motor	Person	Plant	Sheep	Sofa	Train	Tv	MAP	
INRIA[7]	45.2	78.4	69.7	86.1	52.4	54.4	54.3	75.8	62.1	63.5	
NUS[1]	74.6	85.0	76.8	91.1	53.9	61.0	67.5	83.6	70.6	70.5	
AGS[2]	51.3	85.4	80.2	91.1	48.1	61.7	67.7	86.3	70.9	71.1	
CNN-SVM[15]	77.8	81.8	78.8	90.2	54.8	71.1	62.6	87.4	71.8	73.9	
I-FT [20]	82.7	84.0	76.9	90.4	51.5	79.9	54.1	89.5	65.8	74.4	
HCP-1000C[20]	89.3	89.3	85.2	93.0	64.0	85.7	62.7	94.4	78.3	81.5	
CNN-RNN [19]	**92.4**	**91.7**	84.2	**93.7**	59.8	**93.2**	75.3	**99.7**	78.6	84.0	
Dual-CNN-ML	84.3	89.7	**91.5**	90.6	**81.6**	89.2	**78.5**	82.6	**83.8**	**85.2**	

3 Experiments

In this section, we test our proposed method on two benchmark multi-label image datasets, e.g., Pascal VOC 2007 and VOC 2012 [3]. We employ the VGG-16 [16] network pretrained on ImageNet dataset as the image CNN. The dimension of image and matrix representations is set to 1000. The evaluation metric is the mean average precision (mAP).

On VOC 2007, we compare the proposed model with seven state-of-the-art methods including INRIA [7], NUS [1], AGS [2], CNN-SVM [15], I-FT [20], HCP-1000C [20], and CNN-RNN [19]. INRIA [7] employs the traditional feature extraction-coding-pooling pipeline. NUS [1] trains a codebook for descriptors from VOC. AGS [2] learns a layered representation by transferring VOC data to subcategories. CNN-SVM [15] is based on the OverFeat features and applies SVM as the classifier. I-FT [20] applies the squared loss function for multi-label classification. HCP-1000C [20] employs hypotheses extraction approach to get input hypotheses of the given images and then utilizes the cross-hypothesis max-pooling operation to produce the final prediction. In [19], the authors propose a unified CNN-RNN model which uses the RNN to model the label co-occurrence dependencies by learning the joint image-label embedding.

On VOC 2012, the proposed model is compared with six models including I-FT [20], PRE-1000C and PRE-1512 [13], HCP-1000C [20], LeCun et al. [10], and NUS-PSL [20]. NUS-PSL [20] combines the ambiguity-guided Mixture Model(AMM)[1] and the ambiguity Guided Subcategory (AGS [2]) mining approach. PRE-1000C and PRE-1512 [13] transfer CNN parameters pre-trained on

Table 2. Image classification result on pascal VOC 2012

	Plane	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table
I-FT [20]	94.6	74.3	87.8	80.2	50.1	82.0	73.7	90.1	60.6	69.9	62.7
PRE-1000C [13]	93.5	78.4	87.7	80.9	57.3	85.0	81.6	89.4	66.9	73.8	62.0
LeCun-ICML [10]	96.0	77.1	88.4	85.5	55.8	85.8	78.6	91.2	65.0	74.4	67.7
HCP-1000C [20]	**97.7**	83.0	93.2	**87.2**	59.6	88.2	81.9	**94.7**	66.9	81.6	68.0
NUS-PSL [20]	97.3	84.2	80.8	85.3	60.8	89.9	**86.8**	89.3	**75.4**	77.8	75.1
PRE-1512 [13]	94.6	82.9	88.2	84.1	60.3	89.0	84.4	90.7	72.1	**86.8**	69.0
Dual-CNN-ML	92.2	**91.1**	**94.0**	**87.2**	**79.7**	**96.0**	80.7	90.9	68.8	72.9	**69.1**
	Dog	Horse	Motor	Person	Plant	Sheep	Sofa	Train	Tv	MAP	
I-FT [20]	86.9	78.7	81.4	90.5	45.9	77.5	49.3	88.5	69.2	74.7	
PRE-1000C [13]	89.5	83.2	87.6	95.8	61.4	79.0	54.3	88.0	78.3	78.7	
LeCun-ICML [10]	87.8	86.0	85.1	90.9	52.2	83.6	61.1	91.8	76.1	79.0	
HCP-1000C [20]	**93.0**	88.2	87.7	92.7	59.0	85.1	55.4	93.0	77.2	81.7	
NUS-PSL [20]	83.0	87.5	**90.1**	95.0	57.8	79.2	**73.4**	94.5	80.7	82.2	
PRE-1512 [13]	92.1	**93.4**	88.6	**96.1**	64.3	86.6	62.3	91.1	79.8	82.8	
Dual-CNN-ML	92.8	73.3	83.3	91.9	**90.3**	**91.4**	73.0	**95.0**	**85.9**	**85.0**	

the ImageNet dataset with 1000 categories and the augmented one with 1512 categories to other visual recognition tasks with limited training data, respectively.

The results on VOC 2007 are shown in Table 1. From this table, we can see that when taking the label co-occurrence dependencies into consideration, CNN-RNN [19] and Dual-CNN-ML can achieve better performance on most cases than other approaches that do not consider the label co-occurrence dependencies. This confirms that the label co-occurrence dependencies are helpful for multi-label classification. In addition, Dual-CNN-ML obtains better performance than CNN-RNN. It is worthy to note that Dual-CNN-ML performs poorly on some labels. The possible reason is that Dual-CNN-ML pays more attention to some labels with high dependencies, such as "bottle" and "table"; therefore, it may ignore some labels that have little dependencies with other labels, e.g., "train".

The results on VOC 2012 are shown in Table 2. From this table, we can observe that Dual-CNN-ML obtains better results on ten labels than other methods. Moreover, it has the best overall performance, i.e., the best MAP. Note that PRE-1000C [13], PRE-1512 [13] and HCP-1000C [20] obtain the best results on several labels. The main reason is that they use additional samples for training, from which they learn more information.

From Tables 1 and 2, we can find that our proposed method–Dual-CNN-ML can work well for the multi-label classification task.

In addition, to get an intuitive observation of the results of Dual-CNN-ML, we further show some image classification results and its ground-truth in Fig. 4. From this figure, we can observe that Dual-CNN-ML can correctly predict most labels of these instances.



Ground-truth: chair, person, tvmonitor	**Ground-truth**: chair, person, bottleplant, sofa	**Ground-truth**: dog, sofa
Predictions: chair, person, tvmonitor	**Predictions**: chair, person, sofa	**Predictions**: dog, sofa

Fig. 4. Example prediction results and ground-truth on Pascal VOC datasets.

4 Conclusion

In this paper, we propose a novel CNN based model for multi-label image classification task, i.e., Dual-CNN-ML, which makes use of two CNN models (image CNN and matrix CNN) to extract features and label co-occurrence dependencies, respectively. Specifically, the image CNN takes image as input and then extracts the image representation; the matrix CNN models label co-occurrence dependencies. In addition, the convolution fusion layers fuse the image and co-occurrence matrix representations together and obtain the new representation. Finally, the joint representation is input to prediction layer to generate the labels. Extensive experiments on benchmark datasets demonstrate that our method can achieve competitive results compared with several state-of-the-art approaches.

Acknowledgement. This work was partially supported by National Natural Science Foundation of China (61173068, 61573212), Program for New Century Excellent Talents in University of the Ministry of Education, Key Research and Development Program of Shandong Province (2016GGX101044).

References

1. Chen, Q., Song, Z., Dong, J., Huang, Z., Hua, Y., Yan, S.: Contextualizing object detection and classification. IEEE Trans. Pattern Anal. Mach. Intell. **37**(1), 13–27 (2015)
2. Dong, J., Xia, W., Chen, Q.. Feng, J., Huang, Z., Yan, S.: Subcategory-aware object classification. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 827–834 (2013)
3. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (VOC) challenge. Int. J. Comput. Vis. **111**(1), 117–176 (2010)
4. Gong, Y., Jia, Y., Leung, T., Toshev, A., Ioffe, S.: Deep convolutional ranking for multilabel image annotation. arXiv (2013)

5. Guillaumin, M., Mensink, T., Verbeek, J., Schmid, C.: Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In: Proceedings of IEEE International Conference on Computer Vision, pp. 309–316 (2009)
6. Guo, Y., Gu, S.: Multi-label classification using conditional dependency networks. In: Proceedings of International Joint Conference on Artificial Intelligence, pp. 1300–1305 (2011)
7. Harzallah, H., Jurie, F., Schmid, C.: Combining efficient object localization and image classification. In: Proceedings of IEEE International Conference on Computer Vision, pp. 237–244 (2009)
8. Kan, M., Shan, S., Chen, X.: Multi-view deep network for cross-view classification. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 4847–4855 (2016)
9. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Proceedings of Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
10. LeCun, Y., Ranzato, M.: Deep learning tutorial. In: Proceedings of International Conference on Machine Learning (2013)
11. Li, X., Zhao, F., Guo, Y.: Multi-label image classification with a probabilistic label enhancement model. In: Proceedings of International Conference on Uncertainty in Artificial Intelligence, pp. 430–439 (2014)
12. Murthy, V.N., Singh, V., Chen, T., Manmatha, R., Comaniciu, D.: Deep decision network for multi-class image classification. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 2240–2248 (2016)
13. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Learning and transferring mid-level image representations using convolutional neural networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1717–1724 (2014)
14. Read, J., Pfahringer, B., Holmes, G., Frank, E.: Classifier chains for multi-label classification. Mach. Learn. **85**, 333–359 (2011)
15. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y.: Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv (2013)
16. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv (2014)
17. Song, X., Jiang, S., Herranz, L.: Multi-scale multi-feature context modeling for scene recognition in the semantic manifold. IEEE Trans. Image Process. **26**(6), 2721–2735 (2017)
18. Song, X., Jiang, S., Herranz, L., Kong, Y., Zheng, K.: Category co-occurrence modeling for large scale scene recognition. Pattern Recogn. **59**, 98–111 (2016)
19. Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., Xu, W.: CNN-RNN: A unified framework for multi-label image classification. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 2285–2294 (2016)
20. Wei, Y., Xia, W., Huang, J., Ni, B., Dong, J., Zhao, Y., Yan, S.: CNN: Single-label to multi-label. arXiv (2014)
21. Wu, J., Yu, Y., Huang, C., Yu, K.: Deep multiple instance learning for image classification and auto-annotation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 3460–3469 (2015)
22. Xu, X.-S., Jiang, Y., Peng, L., Xue, X., Zhou, Z.-H.: Ensemble approach based on conditional random field for multi-label image and video annotation. In: Proceedings of ACM International Conference on Multimedia, pp. 1377–1380 (2011)

23. Xu, X.-S., Jiang, Y., Xue, X., Zhou, Z.-H.: Semi-supervised multi-instance multi-label learning approach for video annotation task. In: Proceedings of ACM International Conference on Multimedia, pp. 737–740 (2012)
24. Xue, X., Zhang, W., Zhang, J., Wu, B., Fan, J., Lu, Y.: Correlative multi-label multi-instance image annotation. In: Proceedings of IEEE International Conference on Computer Vision, pp. 651–658 (2011)
25. Zhou, F., Lin, Y.: Fine-grained image classification by exploring bipartite-graph labels. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1124–1133 (2016)