KDEVIR at ImageCLEF 2014 Scalable Concept Image Annotation Task: Ontology-based Automatic Image Annotation

Ismat Ara Reshma[1], Md Zia Ullah[2], and Masaki Aono[†]

Department of Computer Science and Engineering,
Toyohashi University of Technology,
1-1 Hibarigaoka, Tempaku-Cho, Toyohashi, 441-8580, Aichi, Japan,
{reshma[1],arif[2]}@kde.cs.tut.ac.jp, aono@tut.jp[†]

Abstract. In this paper, we describe our participation in the ImageCLEF 2014 Scalable Concept Image Annotation task. In this participation, we propose a novel approach of automatic image annotation by using ontology at several steps of supervised learning. In this regard, we construct tree-like ontology for each annotating concept of images using WordNet and Wikipedia as primary source of knowledge. The constructed ontologies are used throughout the proposed framework including several phases of training and testing of one-vs-all SVMs classifier. Experimental results clearly demonstrate the effectiveness of the proposed framework.

Keywords: Concept Detection, Classification, Image Annotation, Ontology

1 Introduction

Due to the explosive growth of digital technologies, collections of images are increasing tremendously in every moment. The ever growing size of the image collections has evolved the necessity of image retrieval (IR) systems; however, the task of IR from a large volume of images is formidable since binary stream data is often hard to decode, and we have very limited semantic contextual information about the image content.

To enable the user for searching images using semantic meaning, automatically annotating images with some concepts or keywords using machine learning is a popular technique. During last two decades, there are a large number of researches being lunched using state-of-the-art machine learning techniques [1–4] (e.g. SVMs, Logistic Regression). In such efforts, most often each image is assumed to have only one class label. However, this is not necessarily true for real world applications, as an image might be associated with multiple semantic tags. Therefore, it is a practical and important problem to accurately assign multiple labels to one image. To alleviate above problem i.e. to annotate each image with multiple labels, a number of research have been carried out; among

them adopting probabilistic tools such as the Bayesian methods is popular [5–7]. More review can be found in [8, 9]. However, accuracy of such approach depends on expensive human labeled training data.

Fortunately, some initiatives have been taken to reduce the reliability on manually labeled image data [10–13] by using cheaply gathered web data. Although the "semantic gaps" between low-level visual features and high-level semantics still remain and accuracy is not improved remarkably.

In order to reduce the dependencies of human-labeled image data, ImageCLEF [14] has been organizing the photo annotation and retrieval task for the last several years, where training data is a large collection of Web images without ground truth labels. Despite the proposed methods in this task shown encouraging performance on a large scale dataset, unfortunately none of them utilizes the sematic relations among annotating concepts. In this paper, we describe the participation of KDEVIR at ImageCLEF 2014 Scalable Concept Image Annotation Task [15], where, we proposed a novel approach, ontology based supervised learning that exploits both low-level visual features and high-level semantic information of images during training and testing. The evaluation results reveal the effectiveness of proposed framework.

The rest of the paper is organized as follows: **Section 2** describes the proposed framework. **Section 3** describes our submitted runs to this task as well as comparison results with other participants' runs. Finally, Concluded remarks and some future directions of our work are described in **Section 4**.

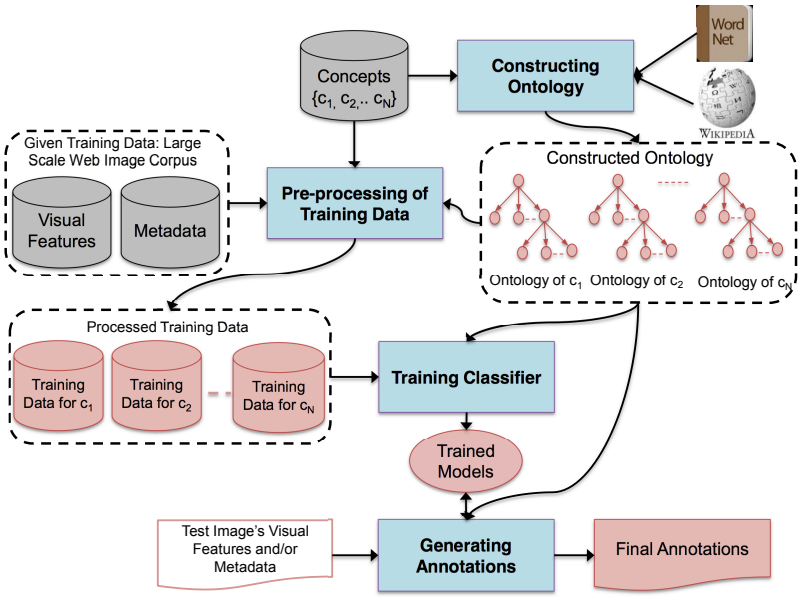


Fig. 1: Proposed Framework

2 Proposed Framework

In this section, we describe our method for annotating images with a list of semantic concepts. We divide our method into four steps: 1) Constructing Ontology, 2) Pre-processing of Training Data, 3) Training Classifier, and 4) Generating Annotations. An overview of our proposed framework is depicted in Fig. 1.

2.1 Constructing Ontology

Ontologies are the structural frameworks for organizing information about the world or some part of it. In computer science and information science, ontology is defined as an explicit, formal specification of a shared conceptualization [16, 17] and it formally represents knowledge as a set of concepts within a domain, and the relationships between those concepts. To utilize these relationships in image annotation, we construct ontology for each concept of a predefined list of concept used to annotate images.

In real world, an image might contain multiple objects (aka concepts) in a single frame, where concepts are inter-related and maintain a natural way of being co-appearance. We use these hypotheses to construct ontologies for concepts. In this regard, we utilize WordNet [18] and Wikipedia as primary sources of knowledge. However, WordNet and Wikipedia themselves have some limitations which cause obstacles to construct ontology using its. For example, WordNet considers very small number of conceptual relations and very few cross-POS (Parts of Speech) pointers among words; on the other hand, Wikipedia contains wide range of semantic information, however, is not structured as WordNet and prone to contain noises, as of being free to edit for all expert and non-expert contributor. As, both of the sources have some limitations, during knowledge extraction we choose those parts of both sources which are less prone to noise and semantically more confident. Thus, take the advantage of both structured representation of WordNet and wide diversity of semantic relations of Wikipedia.

Let C be a set of concepts. We will construct a tree-like [19] ontology for each concept $c_c \in C$. In order to build ontologies, first of all, we select some types of relations including: 1) taxonomical R_t, 2) bionomical R_b, 3) food habitual R_{fh}, and 4) weak hierarchical, R_{wh}. The first and fourth types of relations define relations among any types of concepts, where second and third types define relations among concepts which are biological living things. The relations are extracted empirically according to our observations on WordNet and hundreds of Wikipedia articles. According to the semantic confidence, the order of relation types is: $R_t > R_b > R_{fh} > R_{wh}$. For each type of relations, we extract a set of relations as listed below:

- $R_t = \{inHypernymPathOf,\ superClassOf\}$
- $R_b = \{habitat,\ inhabit,\ liveIn,\ foundOn,\ foundIn,\ locateAt,\ nativeTo\}$
- $R_{fh} = \{liveOn,\ feedOn\}$
- $R_{wh} = \{kindOf,\ typeOf,\ representationOf,\ methodOf,\ appearedAt,\ appearedIn,\ ableToProduce\}$

Finally, we apply some *"if-then"* type inference rules to add an edge from a parent-concept to a child-concept by leveraging the above relations as illustrated in Fig. 2. In addition, for some concepts, especially adjectives (e.g. indoor, outdoor), which have neither much lexical information in WordNet, nor any Wikipedia articles, we manually determine the relations to other concepts.

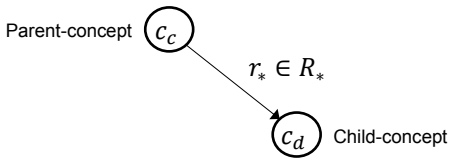


Fig. 2: Connecting concepts c_c and $c_d \in C$ according to relation, $r_* \in R_*$

2.2 Pre-processing of Training Data

For a given list of concepts, we select the most weighted images for each concept from the noisy training images by exploiting their metadata (details about metadata are given in [15]) and pre-constructed concept ontologies. In this regards, first of all, we detect the nouns and adjectives from metadata using WordNet followed by singularizing with Pling Stemmer[1]. Secondly, detected terms from metadata: Web text (scofeat), keywords, and URLs are weighted by BM25 [20], mean reciprocal rank (MRR), and a constant weight,$\vartheta \in (0, 1)$ respectively, which is followed by detecting concepts from the weighted sources on appearance basis. Thus, we have three lists of possible weighted concepts from three different sources of metadata for each image.

We take the inverted index of image-wise weighted concepts, thus generate the concept-wise weighted images. To aggregate the images for a concept from three sources, we normalize the weight of images, and linearly combine the normalized BM25 (nBM25) weight, normalized MRR (nMRR), and constant weight ϑ to generate the final weight of images. From the resultant aggregated list of images, top-m images are primarily selected for each concept.

Finally, in order to increase the recall, we merge the primarily selected training images of each concept with its predecessor concepts of highest semantic confident (i.e. predecessors connected by $r_t \in R_t$) by leveraging our concept ontologies. Thus, we enhance training images per-concept as well as number of annotated concepts per-image.

2.3 Training Classifier

This subsection is partially inspired by the winner of ImageCLEF 2013 [21].

[1] http://www.mpi-inf.mpg.de/yago-naga/javatools/index.html

Image annotation is a multi-class multi-label classification problem; current state-of-the-art classifiers are not able to solve this problem in their usual format. Towards this problem, we propose a novel technique of using ontologies during different phases of learning a classifier. In this regard, we choose Support Vector Machines (SVMs) as a classifier for its robustness of generalization. We subdivide the whole problem into several sub-problems according to the number of concepts, i.e. train SVMs for each concept separately, since using a large dataset at a time is not rational in terms of memory and time.

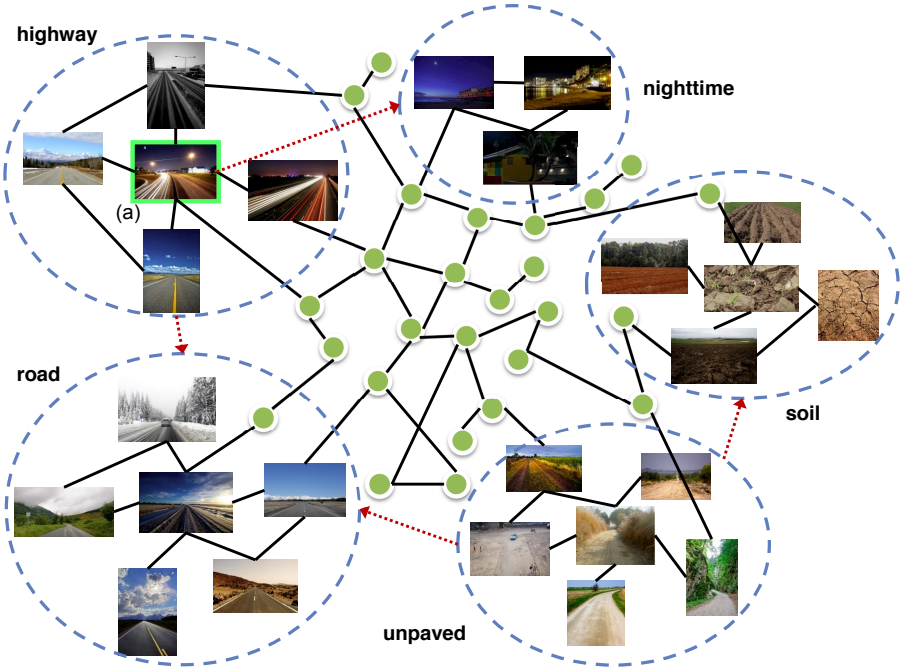


Fig. 3: This figure illustrates examples of images (taken from Flickr) and their social tag links. Areas surrounded by dashed line represents five different communities, *"highway"*, *"road"*, *"unpaved"*, *"soil"*, and *"nighttime"*. Here, black solid-lines represent contextual connections, where, red dashed-lines represent semantic connections emerged from ontological information. Even though image (a) is in *"highway"* community, it should also belong to *"nighttime"* community as its one of the tag is *"moon"*, which is semantically related to nighttime. At the same way, all the images of *"unpaved"* should also belong to *"road"* and *"soil"* as unpaved is a characteristic of a kind of road and unpaved-road contains soil. Likewise, *"highway"* community also belongs to *"road"* community, as highway is a kind of road.

Another problem is that, along with the different parameters, the classification accuracy of SVMs depends on the positive and negative examples which are used to train the classifier. It is obvious that if classifiers are trained with wrong examples, the prediction will be wrong. However, selecting appropriate training example is formidable without any semantic clues. For example, if we train a classifier about *"soil"* without taking into account semantic inter-links with other concepts, one might choose only the *"soil"* community of Fig. 3 as positive examples, and the remaining are as negative examples. However, it might result in wrongly trained model, since semantically *"unpaved"* contains soil, which should not be in negative example. To handle this issue, we use our pre-constructed concept ontologies. We randomly select with replacement n-folds positive image examples for each concept from its image list and negative examples from image lists of other concepts which are not its successor of strong or weak semantic confident in its ontology.

From the n-folds positive and negative examples, we train n probabilistic one-vs-all SVM models for each concept, where $n \in [1, 10]$. We use LIBSVM [22] to learn the SVM models. As kernel, two hybrid kernels are plugged in, instead of using the default choice linear kernel or Gaussian kernel, since image classification is a nonlinear problem and distribution of image data is unknown. We choose histogram intersection kernel (HIK) [23] as primary kernel which is further used to generate two other hybrid kernels. The HIK is defined as:

$$k_{HI}(h^{(a)}, h^{(b)}) = \sum_{q=1}^{l} min(h_q^{(a)}, h_q^{(b)}) \tag{1}$$

where $h^{(a)}$ and $h^{(b)}$ are two normalized histograms of l bins; in context of image data, two feature vectors of l dimensions.

One of the hybrid kernels is convex combination of HIKs (CCHIK) generated from low level visual features of image defined as:

$$K^{(0)} = \frac{1}{|F|} \sum_{s=1}^{|F|} K_{HI}(f_s) \tag{2}$$

where $K_{HI}(f_s)$ is a HIK matrix, computed from feature vector type $f_s \in F$; F is a set of visual feature types (details about used visual features are given in [15]); and $|F|$ is the number of elements in F. In this task $|F| = 7$.

Another hybrid kernel is context dependent kernel (CDK) [21, 24], defined as:

$$K^{(t+1)} = K^{(0)} + \gamma P K^{(t)} P' \tag{3}$$

where $K^{(0)}$ is the CCHIK kernel, P is the left stochastic adjacency matrix between images with each entry proportional to the number of shared labels, and $\gamma \geq 0$. Unlike the original CDK, here, we consider semantic links emerged from ontological information along with contextual links (as shown in Fig. 3). These kernels are plugged into the SVMs for training and testing.

2.4 Generating Annotations

The trained models generated in the previous subsection are used to predict annotations. Given a test image with its visual features (which are similar types of training images' visual features) and URLs as metadata, the system finds out the concepts from URLs on appearance basis as did before for URLs of training images. The visual features and detected concepts are used to calculate kernel values mentioned in previous subsection, which are in turn used for predicting annotations. For the given test image, if a model of particular concept responds positively, the image is considered as voted by current model i.e. the corresponding concept is primarily selected for annotation. At the same time, the tracks of predicted probability and vote are kept. This process is repeated for all learned models of all concepts. The concept-wise predicted probabilities and votes are accumulated for n-models. In second level selection, empirical thresholds for accumulated probabilities and votes are used to select more relevant annotations. In third level, we take top-k weighted concepts, and finally, the test image is annotated with the selected concepts along with their predecessor concepts in concept ontologies.

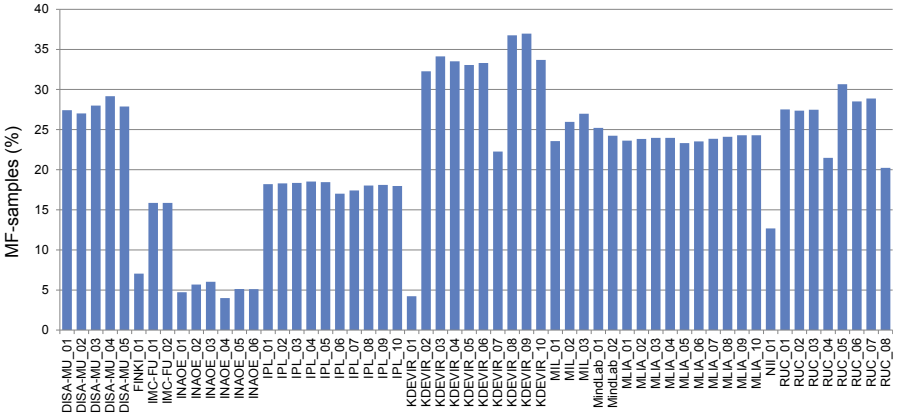
3 KDEVIR Runs and Comparative Results

We submitted total ten runs, which are differ from each other in terms of: use of ontology or not, if used, then in terms of used relation types of different semantic confident during final stage of generating annotation; used kernel (e.g. CCHIK, CDK); number of primarily selected training images, m; and number of trained models for each concept, n. The configurations of all runs are given in Table 1, where, runs are arranged according to their original name to ease the flow of description. All the parameters used in our proposed framework were set empirically to obtain optimal F-measure based on sample (MF-samples) of corresponding run on development set. Details about all the performance measures are given in [15].

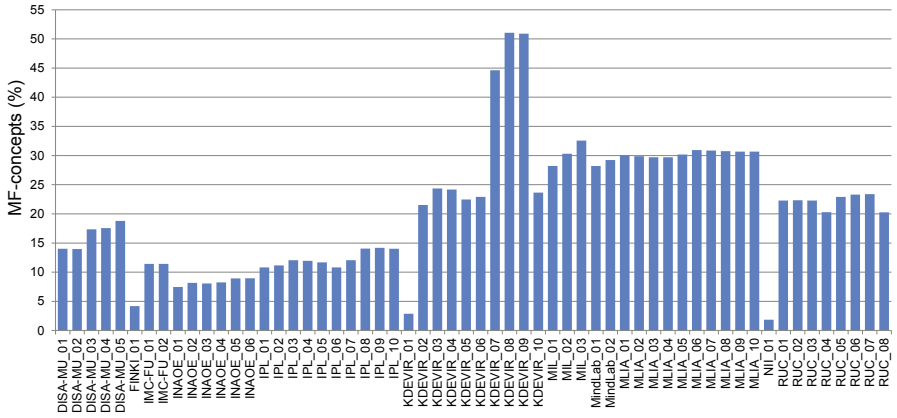
In Fig. 4, and 5, comparisons of our runs (denoted KDEVIR-*) and other participants' runs are illustrated. It reveals the most effectiveness of our proposed approach over other participants' runs. Among the our submitted runs, in Run 1 and 7, we did not exploit semantic information from ontology to compare the effectiveness of our proposed ontology-based approach over ontology-free ordinary one-vs-all SVMs setting with CCHIK and CDK respectively. The comparison results depict that proposed approach tremendously outperform the ordinary one-vs-all SVMs setting.

4 Conclusion

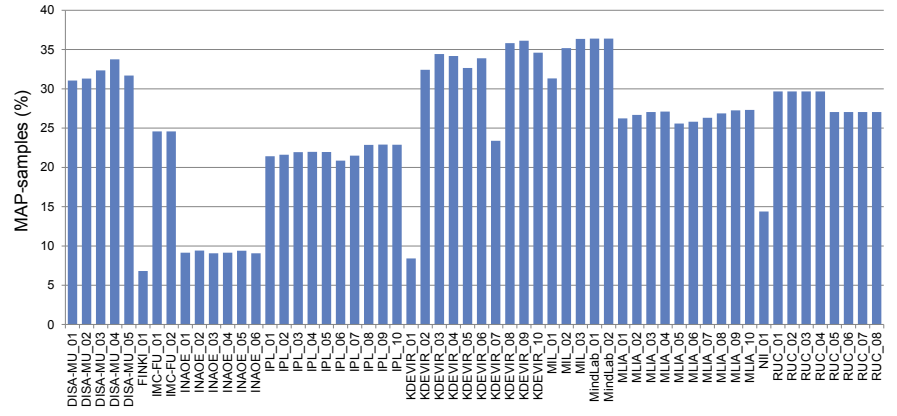
In this paper, we described the participation of KDEVIR at ImageCLEF 2014 Scalable Concept Image Annotation task, where, we proposed a novel approach



(a)

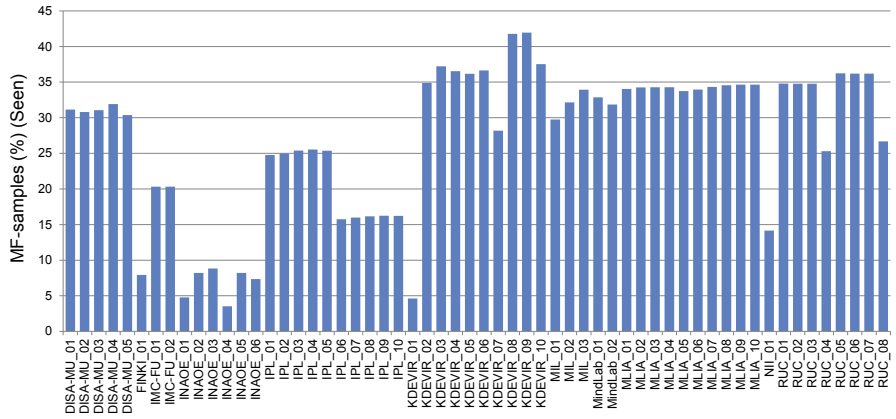


(b)

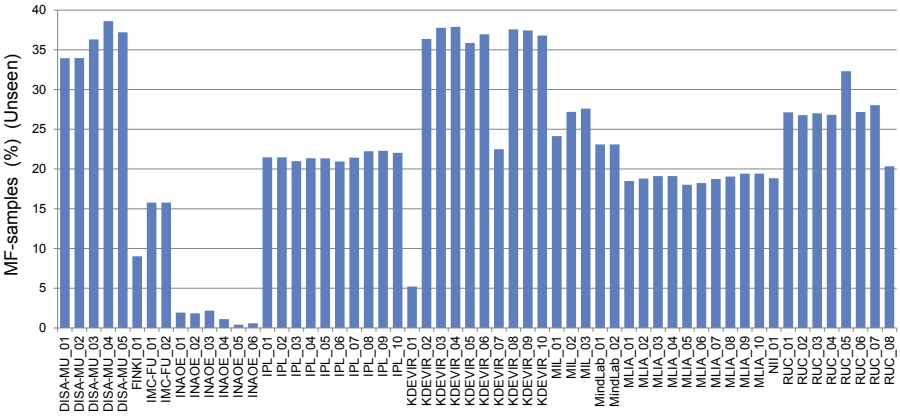


(c)

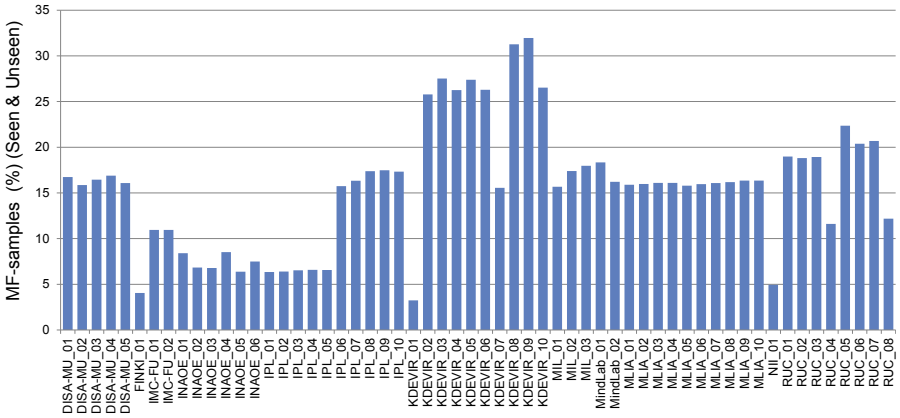
Fig. 4: These figures illustrate a comparison (as released by the ImageCLEF 2014 organizers in http://www.imageclef.org/2014/annotation/results) of our runs (denoted **KDEVIR-***) and other participants' runs on the test set. Acronyms stand for **RUC**: Renmin U. of China, **DISA-MU**: Masaryk U. in Czech Republic, **MIL**: Tokyo U., **MindLab**: National U. of Colombia, **MLIA**: Kyushu U. in Japan, **IPL**: Athens U. of Economics and Business, **IMC-FU**: Fudan U. in China, **NII**: National Institute of Informatics in Japan, **FINKI**: Ss.Cyril and Methodius U. in Macedonia, **INAOE**: National Institute of Astrophysics, Optics and Electronics in Mexico. (a) mean F-measures for samples (MF-samples), (b) mean F-measures for concepts (MF-concepts), and (c) mean average precision for samples (MAP-samples)



(a)



(b)



(c)

Fig. 5: These figures illustrate a comparison (as released by the ImageCLEF 2014 organizers in http://www.imageclef.org/2014/annotation/results) of our runs (denoted **KDEVIR-***) and other participants' runs on three different subsets of the test set in terms of mean F-measures for samples (MF-samples). Acronyms stand for **RUC**: Renmin U. of China, **DISA-MU**: Masaryk U. in Czech Republic, **MIL**: Tokyo U., **MindLab**: National U. of Colombia, **MLIA**: Kyushu U. in Japan, **IPL**: Athens U. of Economics and Business, **IMC-FU**: Fudan U. in China, **NII**: National Institute of Informatics in Japan, **FINKI**: Ss.Cyril and Methodius U. in Macedonia, **INAOE**: National Institute of Astrophysics, Optics and Electronics in Mexico. (a) for the subset of test set seen during development, (b) for the subset of test set unseen during development, and (c) for the subset of test set, which contains both seen and unseen samples during development

Table 1: Configurations of submitted runs. The run pairs **Run** {**1**, **2**} and **Run** {**7**, **8**} were conducted to show the effectiveness of using ontology (proposed method) for CCHIK and CDK respectively; while, the run pairs **Run** {**5**, **6**} and **Run** {**8**, **9**} were conducted to show the effect of considering most semantic confident relation type R_t over all relation types (R_t, R_b, R_{fh}, R_{wh}) for CCHIK and CDK respectively. **Run 3** and **4** were conducted to show the effect of selecting different number of training images during primary selection. We aggregated the decision of models from **Run 3** and **6** at **Run 10**.

Run	**Ontology?**	**Relation type**	**Kernel**	m	n
Run 1 (KDEVIR-01)	**No**	-	CCHIK	800	1
Run 2 (KDEVIR-02)	**Yes**	R_t	Ditto	800	1
Run 3 (KDEVIR-05)	Yes	Ditto	Ditto	**2000**	4
Run 4 (KDEVIR-06)	Yes	Ditto	Ditto	**800**	4
Run 5 (KDEVIR-04)	Yes	**All**	Ditto	800	6
Run 6 (KDEVIR-03)	Yes	$\boldsymbol{R_t}$	Ditto	800	6
Run 7 (KDEVIR-07)	**No**	-	CDK	800	1
Run 8 (KDEVIR-08)	**Yes**	**All**	Ditto	800	1
Run 9 (KDEVIR-09)	Yes	$\boldsymbol{R_t}$	Ditto	800	1
Run 10 (KDEVIR-10)	Yes	Ditto	CCHIK	2000, 800	10

for annotating images using ontologies at several phases of supervised learning from large scale noisy training data.

The evaluation result reveals that our proposed approach achieved the most effective and best performance among 58 submitted runs in terms of MF-samples and MF-concepts. Moreover, according to the MAP-samples it produced comparable result, although we did not prioritize the annotated concepts came from semantic relation (i.e. we assigned the same weights of originally predicted concepts to their corresponding semantically emerged concepts in annotation of a particular image). In future, we will consider fuzzy relations among concepts in ontologies to facilitate more robust ranking of annotation, thus increase the MAP, and incorporate distributed framework to ensure scalability.

Acknowledgement

This research was partially supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Grant-in-Aid (B) 26280038.

References

1. Dumont, M., Marée, R., Wehenkel, L., Geurts, P.: Fast multi-class image annotation with random windows and multiple output randomized trees. In: Proc. International Conference on Computer Vision Theory and Applications (VISAPP) Volume. Volume 2. (2009) 196–203

2. Alham, N.K., Li, M., Liu, Y.: Parallelizing multiclass support vector machines for scalable image annotation. Neural Computing and Applications **24**(2) (2014) 367–381
3. Qi, X., Han, Y.: Incorporating multiple svms for automatic image annotation. Pattern Recognition **40**(2) (2007) 728–741
4. Park, S.B., Lee, J.W., Kim, S.K.: Content-based image classification using a neural network. Pattern Recognition Letters **25**(3) (2004) 287–300
5. Rui, S., Jin, W., Chua, T.S.: A novel approach to auto image annotation based on pairwise constrained clustering and semi-naïve bayesian model. In: Multimedia Modelling Conference, 2005. MMM 2005. Proceedings of the 11th International, IEEE (2005) 322–327
6. Yang, C., Dong, M., Fotouhi, F.: Image content annotation using bayesian framework and complement components analysis. In: Image Processing, 2005. ICIP 2005. IEEE International Conference on. Volume 1., IEEE (2005) I–1193
7. Jeon, J., Lavrenko, V., Manmatha, R.: Automatic image annotation and retrieval using crossmedia relevance models. (2003)
8. Gong, Y., Jia, Y., Leung, T., Toshev, A., Ioffe, S.: Deep convolutional ranking for multilabel image annotation. arXiv preprint arXiv:1312.4894 (2013)
9. Zhang, D., Islam, M.M., Lu, G.: A review on automatic image annotation techniques. Pattern Recognition **45**(1) (2012) 346–362
10. Cai, D., He, X., Li, Z., Ma, W.Y., Wen, J.R.: Hierarchical clustering of www image search results using visual, textual and link information. In: Proceedings of the 12th annual ACM international conference on Multimedia, ACM (2004) 952–959
11. Gupta, M.R., Bengio, S., Weston, J.: Training highly multiclass classifiers. Journal of Machine Learning Research **15** (2014) 1–48
12. Weston, J., Bengio, S., Usunier, N.: Large scale image annotation: learning to rank with joint word-image embeddings. Machine learning **81**(1) (2010) 21–35
13. Wang, X.J., Zhang, L., Jing, F., Ma, W.Y.: Annosearch: Image auto-annotation by search. In: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on. Volume 2., IEEE (2006) 1483–1490
14. Caputo, B., Müller, H., Martinez-Gomez, J., Villegas, M., Acar, B., Patricia, N., Marvasti, N., Üsküdarlı, S., Paredes, R., Cazorla, M., Garcia-Varea, I., Morell, V.: ImageCLEF 2014: Overview and analysis of the results. In: CLEF proceedings. Lecture Notes in Computer Science. Springer Berlin Heidelberg (2014)
15. Villegas, M., Paredes, R.: Overview of the ImageCLEF 2014 Scalable Concept Image Annotation Task. In: CLEF 2014 Evaluation Labs and Workshop, Online Working Notes. (2014)
16. Gruber, T.R.: Toward principles for the design of ontologies used for knowledge sharing? International journal of human-computer studies **43**(5) (1995) 907–928
17. Studer, R., Benjamins, V.R., Fensel, D.: Knowledge engineering: principles and methods. Data & knowledge engineering **25**(1) (1998) 161–197
18. Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM **38**(11) (1995) 39–41
19. Wei, W., Gulla, J.A.: Sentiment learning on product reviews via sentiment ontology tree. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics (2010) 404–413
20. Robertson, S.E., Walker, S., Beaulieu, M., Willett, P.: Okapi at trec-7: automatic ad hoc, filtering, vlc and interactive track. Nist Special Publication SP (1999) 253–264

21. Sahbi, H.: Cnrs-telecom paristech at imageclef 2013 scalable concept image annotation task: Winning annotations with context dependent svms. In: CLEF 2013 Evaluation Labs and Workshop, Online Working Notes. Valencia, Spain (September 23-26 2013) Overview of the ImageCLEF. (2013)
22. Chang, C.C., Lin, C.J.: Libsvm: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) **2**(3) (2011) 27
23. Swain, M.J., Ballard, D.H.: Color indexing. International journal of computer vision **7**(1) (1991) 11–32
24. Sahbi, H., Audibert, J.Y., Keriven, R.: Context-dependent kernels for object classification. Pattern Analysis and Machine Intelligence, IEEE Transactions on **33**(4) (2011) 699–708