CNRS - TELECOM ParisTech at ImageCLEF 2013 Scalable Concept Image Annotation Task: Winning Annotations with Context Dependent SVMs

Hichem SAHBI

CNRS TELECOM ParisTech,
46 rue Barrault, 75013 Paris, France
hichem.sahbi@telecom-paristech.fr

Abstract. *In this paper, we describe the participation of CNRS TELECOM ParisTech in the ImageCLEF 2013 Scalable Concept Image Annotation challenge. This edition promotes the use of many contextual cues attached to visual contents. Image collections are supplied with visual features as well as tags taken from different sources (web pages, etc.).*
Our framework is based on training support vector machines (SVMs) using a class of kernels referred to as context dependent. These kernels are designed by minimizing objective functions mixing visual features and their contextual cues resulting from surrounding tags. The results clearly corroborate the complementarity of tags and visual features and the effectiveness of these context dependent SVMs for image annotation.

Keywords: Context-Dependent Kernels, Support Vector Machines, Image Annotation

1 Introduction

Conventionally, visual information search requires a preliminary step known as image annotation. The latter is a major challenge (see for instance [14, 33, 31, 23, 24, 17, 5, 8]) and consists in assigning list of keywords (a.k.a concepts) to given visual content. These concepts may either correspond to physical entities (pedestrians, cars, etc.) or to high level aspects resulting from the interaction of many entities into scenes (races, fights, etc.). In both cases, image annotation is challenging due to the perplexity when assigning concepts to scenes especially when the number of possible concepts is taken from a large vocabulary and when analyzing highly semantic contents.

Existing annotation methods (see for instance [5, 17]) are usually *content-based*; they first model image observations using low level features (color, texture, shape, etc.), treat each concept as an independent class, and then train the

corresponding concept-specific classifier to identify images belonging to that concept using a variety of machine learning and inference techniques such as latent Dirichlet allocation [2], Markov models [17, 23], probabilistic latent semantic analysis [21] and support vector machines (SVMs) [10, 30]. These learning machines are used to model correspondences between concepts and low level features and make it possible to assign concepts to new images.

The above annotation methods heavily rely on their visual content for image annotation [26]. Due to the semantic gap, they are unable to fully explore the semantic information inside images; this comes from the statistical inconsistency of low level features with respect to the learned concepts and also complexity of scenes. Another class of annotation methods, referred to as *context-based*, has emerged that takes advantage of extra information (such as contextual cues in social networks [7]) in order to better capture the correlations between images and their semantic concepts. Early methods started to emerge for text documents in social networks [41] and now recent work is handling visual content annotation, in different contexts; such as the approach of [18, 11] that uses visual links as context in social networks, in order to propagate image tags and the method of [34] that uses friendship connections and conditional random fields in order to improve the performance of photo annotation. Other works consider distances between tags using Flickr [38], or context informations taken from personal calendars [9], GPS locations [15], visual appearances [4, 19] and multiple cues [40] in order to improve annotation.

In this paper, we describe the participation of "CNRS-TELECOM ParisTech" at ImageCLEF 2013 Scalable Concept Image Annotation Task [36]. Our proposed solution is based on the design of similarity functions that compare images, using context-dependent kernels. The latter are designed using multiple visual features as well as multiple contextual (text) informations provided in this task. When plugged into SVMs, for image classification and annotation, these kernels turned out to be very effective.
The rest of this paper is organized as follows; in Section 2, we describe motivation and proposed method at a glance. In Section 3, we describe our participation and different runs submitted to this task as well as our results and comparison against other participants' runs. Finally, we conclude the paper in Section 4.

2 Motivation and Proposed Method at a Glance

Among image annotation methods mentioned earlier, those based on machine learning and particularly kernel methods (such as SVMs) are particularly successful but their success remains highly dependent on the choice of kernels. The latter, defined as symmetric and positive semi-definite functions [35, 32], should reserve large values to very similar content and vice-versa. Usual kernels, either holistic [16, 22] or alignment-based [12, 1, 13, 3, 20, 37, 6, 25], consider similarities as decreasing functions of distances between patterns or proportional to the qual-

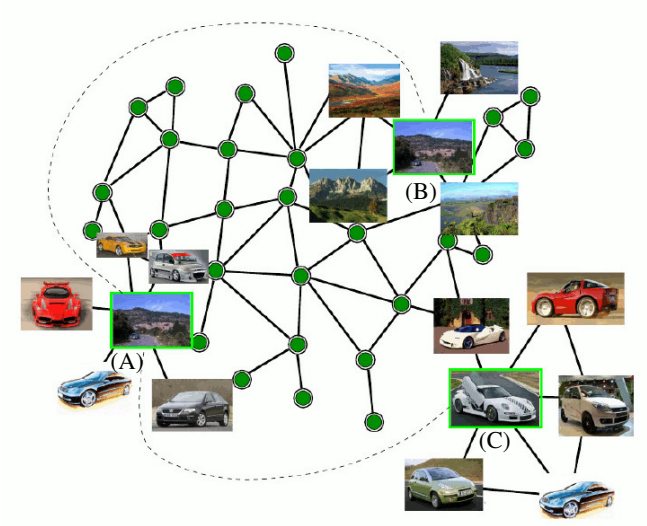

Fig. 1. This figure shows examples of images (taken from Flickr) and their social tag links. Even though images (A), (B) are visually identical, they should be declared as dissimilar as their contexts are different (i.e., they belong to two different communities: "cars" and "landscape" respectively) while images (A), (C) should be declared as strongly related or similar, as they have similar contexts (i.e., they belong to similar communities: "cars").

ity of aligning primitives inside patterns. In both cases, kernels rely only on the intrinsic properties of patterns without taking into account their contextual cues.

We are interested, in this work, in the integration of context in kernels in order to further enhance their discrimination power, for image annotation, while ensuring their positive definiteness and also their efficiency. The guiding principle relies on a basic assertion: kernels should not depend only on intrinsic aspects of images (as images with the same semantic may have different visual and textual features), but also on different sources of knowledge including context. The designed family of kernels, takes high values not only when images share the same content but also the same context. The context of an image is defined as the set of images sharing links (eg. tags) and exhibiting better semantic descriptions, compared to both pure visual and tag based descriptions. The issue of combining context and visual content for image annotation and search has been investigated in previous related work (see for instance [9, 4, 40, 39, 29, 28, 30, 27] and work discussed earlier); the novel part of this work aims to integrate context (from the ImageCLEF 2013 collection), in kernel design for classification and annotation, and plug these kernels in support vector machines in order to take benefit from their well established generalization power [35].

In this work, we use a novel class of kernels (referred to as explicit and context-dependent) for image annotation [27] (see also [29, 28]). An image database is modeled as a graph where nodes are pictures and edges correspond to the shared tagged links. The proposed kernel design method is based on the optimization of an objective function mixing a fidelity term, a context criterion and a regularization term. The fidelity term, takes into account the visual content of images, so highly visually similar contents encourage high kernel values. The context criterion, considers the local graph structure and allows us to further enhance the relevance of our designed kernel, by diffusing and restoring the similarity *iff* pairs of images are *also* surrounded by highly similar images that should also recursively share the same context. The regularization term controls the smoothness of the learned kernel and makes it possible to obtain a closed form solution. Solving this minimization problem results into a recursive similarity function (with an explicit kernel map) that converges to a positive semi-definite fixed-point.
Again, our proposed method goes beyond the naive use of low level features and usual context free kernels (established as the standard baseline in image annotation) in order to design a family of kernels applicable to annotation and suitable to integrate the "contextual" information taken from tagged links in interconnected datasets. In the proposed context-dependent kernel, two images (even with different visual content and even sharing different tags) will be declared as similar if they share the same visual context (see also Fig. 1). This is usually useful as tags in interconnected data may be noisy and misspelled. Furthermore, the intrinsic visual content of images might not always be relevant especially for concepts exhibiting large variation of the underlying visual aspects.

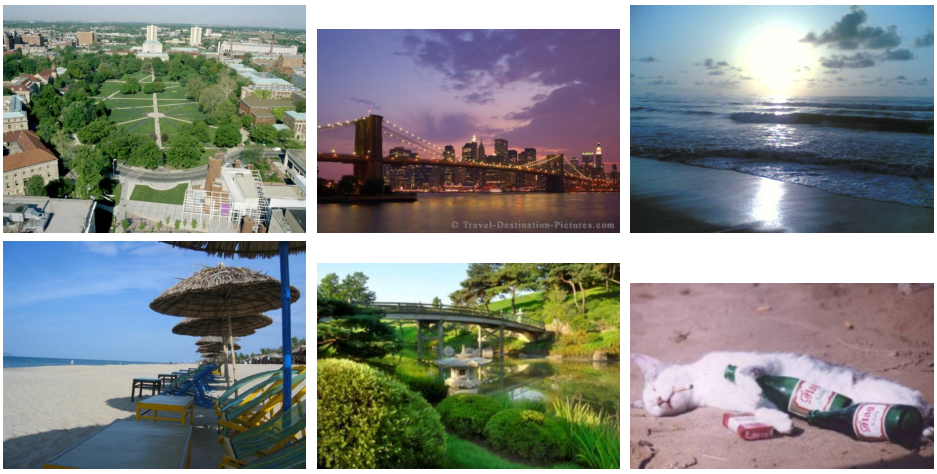

Fig. 2. Figures in the top are taken from the dev set and correspond to the concepts "aerial", "bridge" and "cloud" respectively. Figures in the bottom are taken from the test set.

3 ImageCLEF 2013 Evaluation

The targeted task is image annotation also known as "concept detection"; given a picture of a database, the goal is to predict which concepts (classes) are present into that picture.

3.1 ImageCLEF 2013 Collection

The annotation task, of this year, concentrated on developing annotation algorithms that rely only on data obtained automatically from the web. A very large amount of images was gathered from the web by the organizers, and using associated web pages, tags were also obtained. As tags are noisy (i.e., the degree of relationship between images and the surrounding tags varies greatly), we use some preprocessing in order to assign tags to images.

Dev set. this set is labeled and consists in 1,000 images belonging to 95 categories including "aerial", "bridges", "clouds", etc. Sample of images belonging to the dev set is shown in Fig. 2, top.

Test set. as the objective, of this year task, is to develop algorithms that can easily change or scale the list of concepts used for image annotation, an unlabeled test set was provided and includes 2,000 images belonging to 116 categories; 21 of them are not available in the dev set and are considered as out of list concepts. These concepts include "bottle", "butterfly", "chair", etc. Sample of images belonging to the out of list concepts is shown in Fig. 2, bottom.

Training set label generation. a larger set including 250,000 images was provided with meta-data but without labels. The meta-data, associated to a given image, includes a list of keywords used in order to retrieve that image, in the web, with different search engines.
For a given concept (among the 116 concepts), we extract a training set, by collecting among the 250k images those which include that concept, in their meta-data. As keywords associated to a given concept may appear in different forms, we applied some morphological expansions in order to increase the recall when searching for training images belonging to a given concept.

Context matrix generation. we design a left stochastic adjacency matrix (denoted $\mathbf{P}$) between images with each entry proportional to the number of shared keywords in the meta-data of the underlying images. We use this adjacency matrix in order to build our context dependent kernels as discussed in section 3.3.

3.2 ImageCLEF 2013 Visual Features

We used only the visual features provided in this imageCLEF task including GIST, Color Histograms, SIFT, C-SIFT, RGB-SIFT and OPPONENT-SIFT.

For all the SIFT-based descriptors, a bag-of-words representation is provided. Even though provided, images were not used in order to extract any extra features.

3.3 CNRS-TELECOM ParisTech Runs and Comparison

All our submitted runs (discussed below) are based on SVM training. Again the goal is to achieve image annotation also known as concept detection. For this purpose, we trained "one-versus-all" SVM classifiers for each concept; we use many random folds (taken from training data) for multiple SVM training and we use these SVMs in order to predict the concepts on the dev and test sets. We repeat this training process, for each concept, through different random folds from the training set and we take the average scores of the underlying SVM classifiers. This makes classification results less sensitive to the sampling of the training set.

For all the submitted runs (see runs 1 - 6 below), the only difference resides in the used kernels. We plug the latter into SVMs in order to achieve concept detection. Performances are evaluated using the mean F-measures (at concept and sample levels) as well as the mean average precisions. Details about these measures are given in the ImageCLEF 2013 web page[1].

Run 1. for this run, we build 7 gram matrices[2] associated to the visual features mentioned earlier. Then, we linearly combine those matrices into a single one. Notice that this combination does not result from multiple kernel learning but just a convex combination of kernels with uniform weights. We plug the resulting kernel into SVMs for training and testing. A given test image is assigned to a given concept, iff the underlying SVM score is $\geq \tau$ (with $\tau = 0.5$ in practice).

Run 2. the setting of this run is exactly the same as run 1 except that the cut-off threshold τ is set to 1.

Run 3. the linear combination of kernel matrices (denoted $\mathbf{K}^{(0)}$) obtained in runs 1 and 2 is used as an initialization to the context dependent kernel (CDK) defined as $\mathbf{K}^{(t+1)} = \mathbf{K}^{(0)} + \gamma\ \mathbf{P}\mathbf{K}^{(t)}\mathbf{P}'$, with $\gamma \geq 0$ (see [27]). The latter is computed iteratively (in two iterations) using the adjacency matrix $\mathbf{P}$ introduced earlier. Once designed, we plug CDK into SVMs for training and testing. A given test image is again assigned to a given concept, iff the underlying SVM score is $\geq \tau$ (with $\tau = 0.5$ in practice)

Run 4. the setting of this run is exactly the same as run 3 except that the cut-off threshold τ is set to 1.

[1] http://imageclef.org/2013/photo/annotation
[2] Based on histogram intersection kernel.

Run 5. before computing the convex combination of kernels (as done in runs 3, 4), we first evaluate for each kernel matrix (associated to a given visual feature) its underlying CDK ($\mathbf{K}^{(t+1)} = \mathbf{K}^{(0)} + \gamma\ \mathbf{P}\mathbf{K}^{(t)}\mathbf{P}'$ with $\mathbf{K}^{(0)}$ being the linear kernel matrix). Then, we apply histogram intersection kernel to these CDKs and we linearly combine the resulting kernels with uniform weights. Again, the number of iterations in CDK is set to 2. Once designed, we plug the final kernel matrix into SVMs, for training and testing. A given test image is again assigned to a given concept, iff the underlying SVM score is $\geq \tau$ (with $\tau = 0.5$ in practice)

Run 6. the setting of this run is exactly the same as run 5 except that the cut-off threshold τ is set to 1.

Diagrams in Figs. 3, 4 and 5, show the mean F-measures and mean average precisions of our runs and their comparisons with respect to different participants' runs. From all these results it is clear that our best runs (runs 6 and 4) outperform the others for almost all the evaluation measures.

4 Conclusion

We discussed in this paper, the participation of "CNRS-TELECOM ParisTech" in ImageCLEF 2013 Scalable Concept Image Annotation Task. Our submissions include pure visual runs based on linear combination of elementary histogram intersection kernels, as well as combined visual/textual runs, that consider the context of images through context dependent kernels. The latter turned out to be the most effective and achieved the best performance among 57 participants' runs.

Acknowledgments

This work was supported in part by a grant from the Research Agency ANR (Agence Nationale de la Recherche) under the MLVIS project and a grant from DIGITEO under the RELIR project.

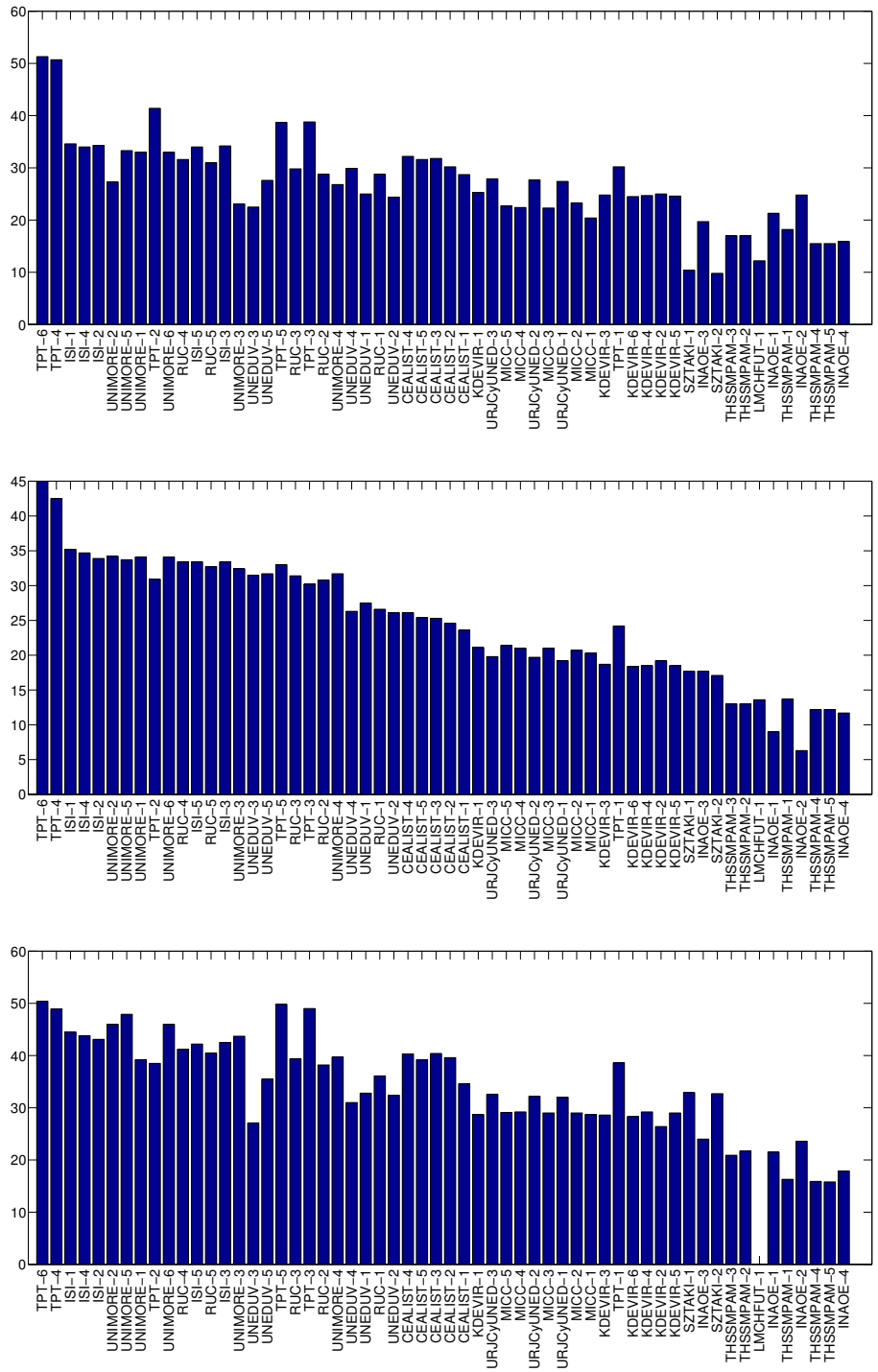

Fig. 3. These diagrams show a comparison (as released by the ImageCLEF 2013 organizers in http://imageclef.org/2013/photo/annotation) of our runs (denoted **TPT-***) and other participants' runs on the dev set. Acronyms stand for **ISI:** Tokyo U., **UNIMORE:** U. of Modena and Reggio Emilia, **RUC:** Renmin U. of China, **UNEDUV:** National U. of Distance Education at Spain, **CEALIST:** CEA, France, **KDEVIR:** Toyohashi U. of Technology in Japan, **URJCyUNED:** King Juan Carlos U. in Spain, **MICC:** Florence U. in Italy, **SZTAKI:** Hungarian Academy of Sciences, **INAOE:** National Institute of Astrophysics, Optics and Electronics in Mexico, **THSSMPAM:** Tsinghua U., Beijing, China, **LMCHFUT:** Hefei University of Technology, China. **Top diagram:** mean F-measures for samples, **middle:** mean F-measures for concepts, and **Bottom:** mean average precision.

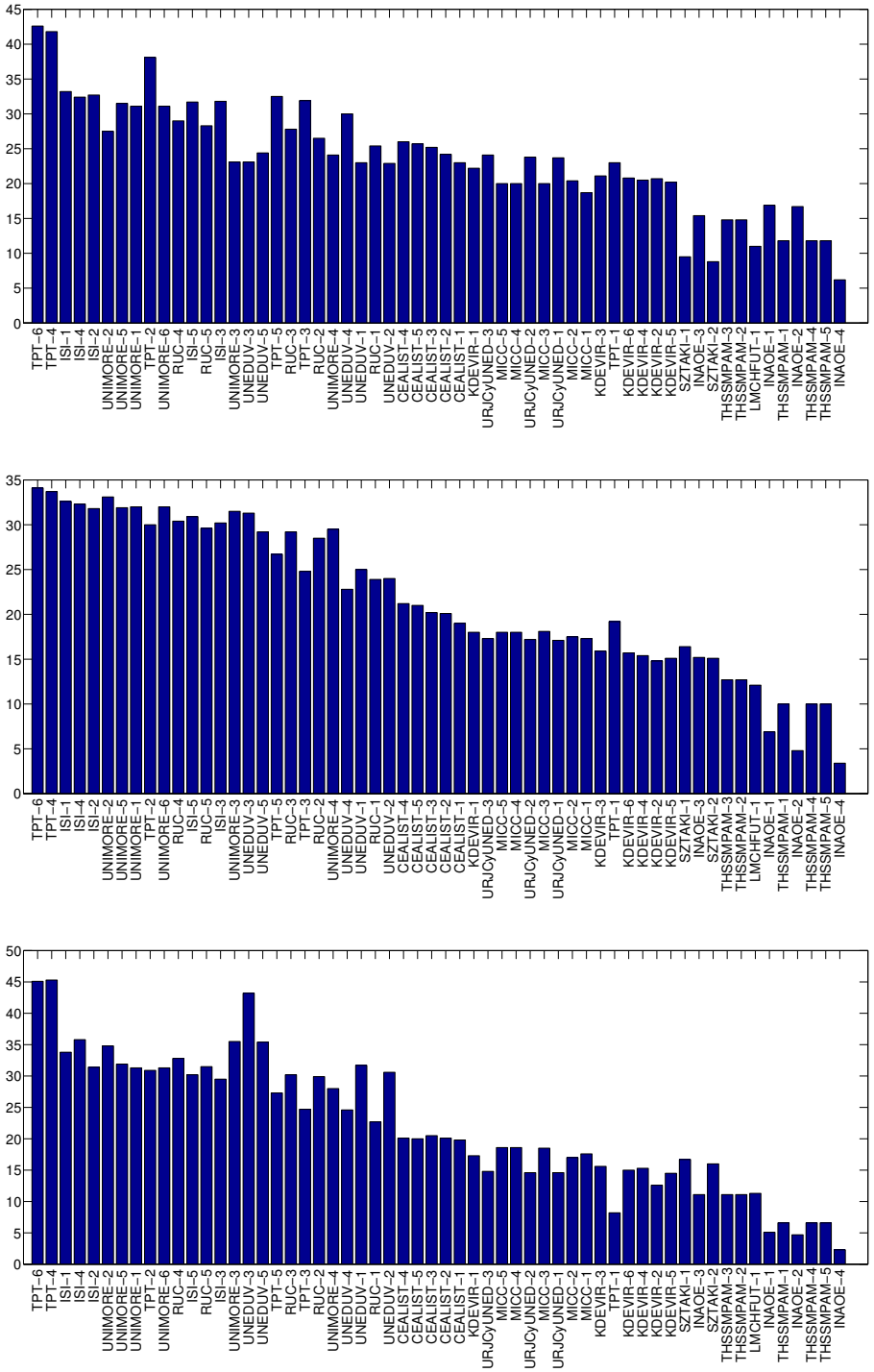

Fig. 4. These diagrams show a comparison (as released by the ImageCLEF 2013 organizers in http://imageclef.org/2013/photo/annotation) of our runs (denoted **TPT-***) and other participants' runs on the test set. Acronyms stand for **ISI:** Tokyo U., **UNIMORE:** U. of Modena and Reggio Emilia, **RUC:** Renmin U. of China, **UNEDUV:** National U. of Distance Education at Spain, **CEALIST:** CEA, France, **KDEVIR:** Toyohashi U. of Technology in Japan, **URJCyUNED:** King Juan Carlos U. in Spain, **MICC:** Florence U. in Italy, **SZTAKI:** Hungarian Academy of Sciences, **INAOE:** National Institute of Astrophysics, Optics and Electronics in Mexico, **THSSMPAM:** Tsinghua U., Beijing, China, **LMCHFUT:** Hefei University of Technology, China. **Top diagram:** mean F-measures for samples, **middle:** mean F-measures for concepts, and **Bottom:** mean F-measures for concepts unseen in the dev set.

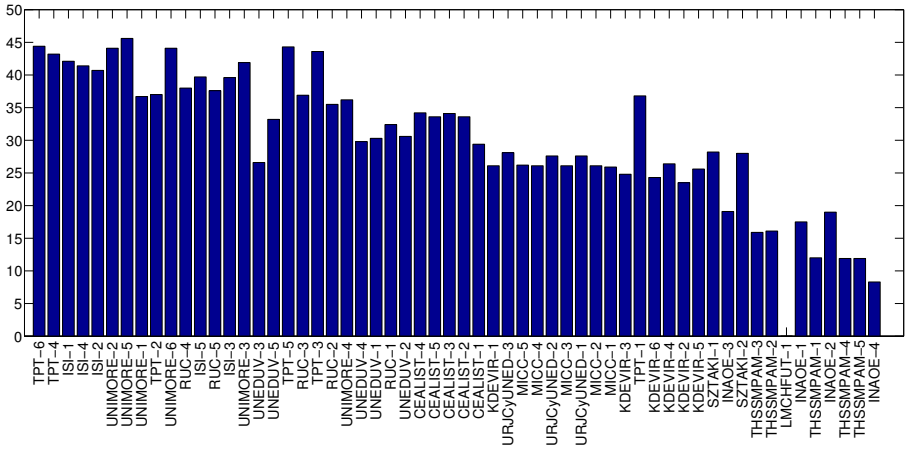

Fig. 5. This diagram shows a comparison (as released by the ImageCLEF 2013 organizers in http://imageclef.org/2013/photo/annotation) of the mean average precision of our runs (denoted **TPT-***) and other participants' runs on the test set. Acronyms stand for **ISI:** Tokyo U., **UNIMORE:** U. of Modena and Reggio Emilia, **RUC:** Renmin U. of China, **UNEDUV:** National U. of Distance Education at Spain, **CEALIST:** CEA, France, **KDEVIR:** Toyohashi U. of Technology in Japan, **URJCyUNED:** King Juan Carlos U. in Spain, **MICC:** Florence U. in Italy, **SZTAKI:** Hungarian Academy of Sciences, **INAOE:** National Institute of Astrophysics, Optics and Electronics in Mexico, **THSSMPAM:** Tsinghua U., Beijing, China. **LMCHFUT:** Hefei University of Technology, China.

References

1. C. Bahlmann, B. Haasdonk, and H. Burkhardt. On-line handwriting recognition with support vector machines, a kernel approach. *IWFHR*, pages 49–54, 2002.
2. K. Barnard, P.Duygululu, D. Forsyth, D. Blei, and M. Jordan. Matching words and pictures. *The Journal of Machine Learning Research*, 2003.
3. S. Boughorbel, J. Tarel, and N. Boujemaa. The intermediate matching kernel for image local features. *IEEE International Joint Conference on Neural Networks*, 2005.
4. L. Cao, J. Luo, and T. Huang. Annotating photo collection by label propagation according to multiple similarity cues. *ACM Multimedia*, 2008.
5. G. Carneiro and N. Vasconcelos. Formulating semantic image annotation as a supervised learning problem. *In: Proc. of CVPR*, 2005.
6. M. Cuturi. Fast global alignment kernels. In *Proceedings of the International Conference on Machine Learning*, 2011.
7. M. Davis, S. King, N. Good, and R. Sarvas. From context to content: leveraging context to infer media metadata. *In: Proceedings of 12th Annual ACM International Conference on Multimedia, MM 2004, Brave New Topics Session on From Context to Content: Leveraging Contextual Metadata to infer Multimedia Content in New York, ACM Press, 188-195*, 2004.
8. P. Duygulu, K. Barnard, J. deFreitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. *In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2353, pp. 97-112. Springer, Heidelberg*, 2002.

9. A. Gallagher, C. Neustaedter, L. Cao, J. Luo, and T. Chen. Image annotation using personal calendars as context. *ACM Multimedia*, 2008.
10. Y. Gao, J. Fan, X. Xue, and R. Jain. Automatic image annotation by incorporating feature hierarchy and boosting to scale up svm classifiers. *in Proc. of ACM MULTIMEDIA*, 2006.
11. Y. Gao, M. Wang, Z.-J. Zha, J. Shen, X. Li, and X. Wu. Visual-textual joint relevance learning for tag-based social image search. *IEEE Trans. Image Processing*, 22, 2013.
12. T. Gartner. A survey of kernels for structured data. *Multi Relational Data Mining*, 5(1):49–58, 2003.
13. K. Grauman and T. Darrell. The pyramid match kernel: Efficient learning with sets of features. *Journal of Machine Learning Research (JMLR)*, 8:725–760, 2007.
14. X. He, R. Zimel, and M. Carreira. Multiscale conditional random fields for image labeling. *In CVPR*, 2004.
15. D. Joshi, A. Gallagher, J. Yu, and J. Luo. Inferring photographic location using geotagged web images. *Multimedia Tools and Applications*, 56(1):131–153, 2012.
16. R. Kondor and T. Jebara. A kernel between sets of vectors. *In proceedings of the 20th International conference on Machine Learning*, 2003.
17. J. Li and J. Z. Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Trans. on PAMI*, 25(9):1075–1088, 2003.
18. X. Li, C. Snoek, and M. Worring. Learning tag relevance by neighbor voting for social image retrieval. *In MIR conference*, 2008.
19. Y. Liu, D. Xu, I.-H. Tsang, and J. Luo. Textual query of personal photos facilitated by large-scale web data. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(5):1022–1036, 2011.
20. S. Lyu. Mercer kernels for object recognition with local features. *In the proceedings of the IEEE Computer Vision and Pattern Recognition*, 2005.
21. F. Monay and D. GaticaPerez. Plsa-based image autoannotation: Constraining the latent space. *in Proc. of ACM International Conference on Multimedia*, 2004.
22. P. Moreno, P. Ho, and N. Vasconcelos. A kullback-leibler divergence based kernel for svm classfication in multimedia applications. *In Neural Information Processing Systems*, 2003.
23. G. Moser and B. Serpico. Combining support vector machines and markov random fields in an integrated framework for contextual image classification. *In TGRS*, 2012.
24. S. Nowak and M. Huiskes. New strategies for image annotation: Overview of the photo annotation task at imageclef 2010. *in The Working Notes of CLEF 2010*, 2010.
25. J. Qiu, M. Hue, A. Ben-Hur, J.-P. Vert, and W. S. Noble. A structural alignment kernel for protein structures. *Bioinformatics*, 23(9):1090–1098, 2007.
26. D. Ritendra, D. Joshi, J. Li, and J. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 2008.
27. H. Sahbi. Explicit context-aware kernel map learning for image annotation. *The 9th International Conference on Computer Vision systems*, 2013.
28. H. Sahbi, J.-Y. Audibert, and R. Keriven. Context-dependent kernels for object classification. *In Pattern Analysis and Machine Intelligence (PAMI)*, 4(33), 2011.
29. H. Sahbi, J.-Y. Audibert, J. Rabarisoa, and R. Keriven. Context-dependent kernel design for object matching and recognition. *In the proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

30. H. Sahbi and X. Li. Context based support vector machines for interconnected image annotation (the saburo tsuji best regular paper award). *In the Asian Conference on Computer Vision (ACCV)*, 2010.
31. D. Semenovich and A. Sowmya. Geometry aware local kernels for object recognition. *In ACCV*, 2010.
32. J. Shawe-Taylor and N. Cristianini. Support vector machines and other kernel-based learning methods. *Cambridge University Press*, 2000.
33. A. Singhal, L. Jiebo, and Z. Weiyu. Probabilistic spatial context models for scene content understanding. *In CVPR*, 2003.
34. Z. Stone, T. Zickler, and T. Darrell. Auto-tagging facebook: Social network context improves photo annotation. *in IVW*, 2008.
35. V. N. Vapnik. Statistical learning theory. *A Wiley-Interscience Publication*, 1998.
36. M. Villegas, R. Paredes, and B. Thomee. Overview of the imageclef 2013 scalable concept image annotation subtask. *CLEF 2013 working notes, Valencia, Spain*, 2013.
37. C. Wallraven, B. Caputo, and A. Graf. Recognition with local features: the kernel recipe. *ICCV*, pages 257–264, 2003.
38. L. Wu, X. Hua, N. Y. nd W. Ma, and S. Li. Flickr distance. *In: Proc. of ACM MULTIMEDIA*, 2008.
39. L. Yang, B. Geng, A. Hanjalic, and X.-S. Hua. A unified context model for web image retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, 8(3):28, 2012.
40. Y.-H. Yang, P-T.Wu, C.-W. Lee, K.-H. Lin, W. Hsu, and H. Chen. Contextseer: Context search and recommendation at query time for shared consumer photos. *ACM Multimedia*, 2008.
41. D. Zhou, J. Bian, S. Zheng, H. Zha, and C. Giles. Exploring social annotations for information retrieval. *in the WWW conference, Beijing, China*, 2008.