Multi-modal Visual Concept Classification of Images via Markov Random Walk over Tags

Motoaki Kawanabe[‡†], Alexander Binder[†], Christina Müller[†] Wojciech Wojcikiewicz[†‡]
[‡]*Fraunhofer Institute FIRST, Kekuléstr. 7, 12489 Berlin, Germany*
[†]*Technical University of Berlin, Franklinstr. 28 / 29, 10587 Berlin, Germany*
nabe@first.fraunhofer.de alexander.binder@tu-berlin.de wojwoj@mail.tu-berlin.de

Abstract

Automatic annotation of images is a challenging task in computer vision because of "semantic gap" between high-level visual concepts and image appearances. Therefore, user tags attached to images can provide further information to bridge the gap, even though they are partially uninformative and misleading. In this work, we investigate multi-modal visual concept classification based on visual features and user tags via kernel-based classifiers. An issue here is how to construct kernels between sets of tags. We deploy Markov random walks on graphs of key tags to incorporate co-occurrence between them. This procedure acts as a smoothing of tag based features. Our experimental result on the ImageCLEF2010 PhotoAnnotation benchmark shows that our proposed method outperforms the baseline relying solely on visual information and a recently published state-of-the-art approach.

1. Introduction

Although much progress has been made during last decades, performance of automatic image annotation systems is far from ability of humans who can distinguish more than ten thousands visual concepts. One of the reasons is "semantic gap" between high-level semantic notions and low-level image features used in such systems. In Internet archives like Flickr, images are often accompanied with user tags which do not coincide with concepts of interest but can provide useful semantic information. In this paper, we consider multi-modal image classification based on visual features and user tags. In contrast to the previous work [3], we assume that user tags are also given for test images as the ImageCLEF2010 PhotoAnnotation task.

There have been considerable works on tags in the field of *searching images* without classification into concepts [6, 15, 13]. A classification approach using textual features directly was presented in [7] where each tag is assigned to be one dimension. Another recent approach [3] deployed essentially a linear tag kernel among frequent tags. In this paper, we will propose a smoothing technique based on Markov random walk on a relational graph between tags [9]. By combining a couple of similarity measures with different smoothness, we could outperform the baseline classifier only with visual features and the previous approach [3].

We are aware that there have been many mathematically appealing model-based approaches [4, 1, 17] which have been very successful in document analysis. However, as pointed out in preceding papers, properties of tag data are completely different from those of documents.

- Number of tags is very small compared to that of words in a document. We also encounter sparse feature representations. Furthermore, there exist substantial number of images without any tags.
- Tags are often very subjective and thus irrelevant or misleading (e.g. 'travel').
- Tags come from multiple languages (e.g. Spanish, German, Chinese).
- Tags can consists of multiple words without spacing (e.g. 'freeyourmind').

These facts make direct applications of e.g. the model-based techniques mentioned above difficult and give rise to new challenges. Our approach aims at dealing with the first issue.

The paper is organized as follows. In Section 2, we will first give a brief overview of the multi-modal classification method of visual concepts, and then will explain the representation of tags and the smoothing procedure via random walks. The experimental setup will be given in Section 3. It will include a highlight of certain challenging statistical properties of tags. Finally, our experimental results will be provided in Section 4, and some insights from them will further be discussed in Section 5.

2. Representations for user tags

In the last decades, support vector machines (SVMs) have been successfully applied to many practical problems in various fields including computer vision [11]. In image categorization, combining many kernels (similarity measures between images) constructed from various image descriptors has become a standard procedure. Furthermore, such learning procedures with multiple kernels also provide a generic framework for multi-modal image annotation [3], when kernels from each modality are prepared. Mathematically, the classifiers used in this approach can be expressed as

$$f(\mathbf{I}_{\text{test}}) = \sum_{i=1}^{n} y_i \alpha_i \left\{ \sum_{j=1}^{m} \beta_j k_j(\mathbf{I}_{\text{test}}, \mathbf{I}_i) \right\}, \qquad (1)$$

where y_i is the label of the image $\mathbf{I}_i$ and k_j is the j-th kernel between (some features of) two images. By multiple kernel learning (MKL), it is even possible to learn the *kernel weights* $\{\beta_j\}_{j=1}^{m}$ and the *sample weights* $\{\alpha_i\}$ simultaneously. However, for simplicity, we deployed uniform kernel weights and trained SVMs with the averaged kernels, which achieved comparable results to MKL. The detailed information about our base kernels will be described in Section 3.3 and 3.4. In the remaining of this section, we will explain our main contribution, a smoothing procedure with Markov random walk in order to alleviate sparse tag distributions.

2.1. The feature and kernel used in [3]

We re-implemented the procedure in Guillaumin et al. [3]. We kept the tags that appear at least 16 times (i.e. among at least 0.2% of the images), resulting in a vocabulary of 392 tags. Then, we encoded the tags assigned to each image into a 392-dimensional binary vector $\mathbf{t}$ whose i-th component takes 1 if the i-th element of the vocabulary is present, and becomes 0 otherwise. Based on this feature representation, the similarity between two images is defined by a linear kernel $k_G(\mathbf{t}_i, \mathbf{t}_j) := \mathbf{t}_i^\top \mathbf{t}_j$ which counts the number of tags in the vocabulary shared by them. We remark that the kernel takes 0 including the diagonal, when no tags in the vocabulary are assigned to one of the images at least. Therefore, the small size vocabulary causes a lot of zero rows and columns in the kernel matrix.

2.2. Smoothing sparse features via Markov random walk

In order to reduce the number of images without assigned words, we considered a vocabulary of larger size by just removing tags that appear less than 3 times and that are shorter than 3 characters. Tags which stem from the same words (e.g. flower and flowers) are unified. In the end, our vocabulary comprised of 2415 tags. Although the number of images without assigned words decreases, this vocabulary leads to a sparse high-dimensional representation which seems to harm classification performances. Inspired by [9], we propose to smooth it via Markov random walk on a probabilistic relational graph on the tags in the vocabulary in order to alleviate the sparsity issue. The transition probability from i-th tag to j-th tag is defined as

$$P_{i \to j} = \frac{n_{ij}}{\sum_k n_{ik}}, \qquad (2)$$

where n_{ij} is the number of images which have both i-th and j-th tags. A given feature $\mathbf{t}^{(0)}$ can be smoothed via iterative application of the matrix P, i.e.

$$\mathbf{t}^{(k)} = P^k \mathbf{t}^{(0)} \qquad (3)$$

where the (i, j)-element of P is the transition probability $P_{j \to i}$. The limit of such smoothed vectors converges to the same stationary distribution vector for all input tag vectors which carries no discriminative information. Thus it makes no sense to iterate the walk ad infinitum. For our experiments we used only the first two steps and the original one $\mathbf{t}^{(k)}$, $k = 0, 1, 2$. The intuition behind this smoothing can be explained in the following example. Consider the case of two tags which co-occur frequently on the data set, but appear mutually exclusive in a pair of images which results in two features $\mathbf{t}_1^{(0)} = (1, 0)$ and $\mathbf{t}_2^{(0)} = (0, 1)$. Prior to smoothing a dimension-wise kernel like linear or Gaussian would result in low similarity between these two features. By applying the random walk P the weight from one tag flows to the other tag in both features so that the similarity between the two smoothed features $\mathbf{t}_1^{(1)} = (P_{1 \to 1}, P_{1 \to 2})$ and $\mathbf{t}_2^{(1)} = (P_{2 \to 1}, P_{2 \to 2})$ increases under a linear or RBF kernel.

3. Experimental setting

3.1. Data set

ImageCLEF2010 PhotoAnnotation data set contains 8000 labeled images from Flickr for training. Its annotation consists of 93 highly-variable concept classes: from well-defined objects like ‘Fish’ to rather ambiguously-defined abstract concepts like ‘Citylife’, ‘Cute’ or ‘Visual_Arts’ which makes the task highly challenging for any recognition system.

3.2. Statistics of tags

The user tags of these images were also provided for multi-modal prediction of the labels. In total, 60587 tags, comprised of 25367 unique ones, are assigned to the images. The average number of tags per image is 7.57 with a standard deviation of 6.95. As anticipated from the statistics, Figure 1 shows that most of the tags occur very infrequently. Due to the sparsity of the tag distribution, it makes

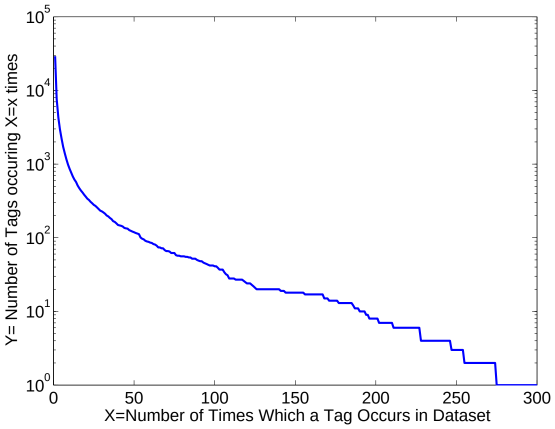


Figure 1. Number of unique tags as a function of their frequency of appearance over the data set (semi-log plot).

sense to reduce the number of words from all unique tags for constructing the bag-of-words features. Indeed, we observed performance gains by ignoring infrequent tags as is the case in document analysis [4]. This is because too rare tags do not provide statistically reliable information which can be generalized from one images to another.

On the other hand the most frequent tags in one class are not necessarily discriminative. The five most frequent tags for the concept 'Sunny' were 'explore', 'blue', 'canon', 'nikon', 'clouds'. The concept 'Cute' had similar uninformative tags like 'explore', 'nikon', 'flower', 'canon', 'water'. In contrast, we found more relevant ones in the most frequent tags for the concepts which showed significant gains by using tags. The concept 'Partylife' was mostly tagged with 'concert', 'live', 'fireworks', 'music', 'night'and for the concept 'Musical_Instrument' had 'concert', 'guitar', 'music', 'live', 'rock' as the most frequent tags. Therefore, we concluded that inverse document frequency terms used in document analysis are not appropriate for user tags. Although excluding uninformative tags could improve classification performance further, we did not pursue this option in this paper.

3.3. Visual features and kernels

We employed the bag-of-visual-words features [2] for information extraction from images. As the base features, we computed for each image grey SIFT [8], opponent SIFT, c-SIFT, rg-SIFT and rgb-SIFT features [16] on a dense grid of pitch size six. Note that all SIFT descriptors except for grey SIFT have dimensionality of 3×128. For each of the five color channels, 4000 visual words were generated from a randomly selected set of the base features in the training images by k-means clustering. Finally, we computed for each of the five color channels three spatially localized bag-of-words histograms over spatial tilings [5] 1×1, 2×2 and 3×1 [16], which results in 15 visual features in total. Our choice of features is a reduced set inspired by the winners' procedures of ImageCLEF2009 PhotoAnnotation and Pascal VOC2008 [14], which achieve comparable results to the state-of-the-art systems. Indeed, we checked via cross-validation that our baseline procedure achieved an AUC score of 83.60 over the 53 subclasses used in the previous ImageCLEF@ICPR2010 competition [10], which is on par with those of the third best group in the official challenge results.

We compute χ^2-kernels for all image features.

$$k(\boldsymbol{v}, \boldsymbol{w}) = \exp\left(-\frac{1}{\sigma}\sum_{i=1}^{d}\frac{(v_i - w_i)^2}{v_i + w_i}\right) \tag{4}$$

The kernel widths σ have been set to the mean of all pairwise χ^2-distances. Rows consisting solely of zeros have been excluded from that computation. All kernels were normalized to equal standard deviation in Hilbert space. This allows us to use the regularization constant to one in the SVM training as a good rule of thumb.

3.4. Textual features and kernels

As the previous work [3], we also used linear kernels

$$k^{(k)}(\mathbf{t}_i^{(k)}, \mathbf{t}_j^{(k)}) = \left\{\mathbf{t}_i^{(k)}\right\}^\top \mathbf{t}_j^{(k)}, \tag{5}$$

for $k = 0, 1, 2$, where $\mathbf{t}^{(k)}$ denotes the smoothed textual feature (Section 2.2). We will indicate these kernels as MRW(k) in our experimental results. When an image does not have any tags listed in the vocabulary, we have set the whole row/column of a kernel corresponding to this image to zero. Since the kernel function is an inner product, this implies that an image with missing textual feature is orthogonal to all other images. This is a generic way to treat missing features in kernel-based approaches. Such a kernel becomes non-negative definite, which does not harm SVM training, as the kernel mixture contains always a strictly positive definite visual kernel.

3.5. Classifier training

As a baseline ('Visual only'), SVM was trained with the average kernel k_V of the 15 kernels constructed from the various bag-of-words visual features (Section 3.3). The state-of-the-art classifier [3] used SVM with the average of the visual kernel k_V and the one k_G explained in Section 2.1 ('Visual+[3]'). We tried four combinations of our MRW kernels ($k^{(k)}$, $k = 0, 1, 2$ in Section 3.4) with the visual kernel k_v: 'Visual+MRW(0)', 'Visual+MRW(1)', 'Visual+MRW(2)' and 'Visual+MRW(all 3)'. Classification performances of these SVMs were compared with those of the baselines.

Table 1. Overall performance comparison of SVMs with the averaged kernels.

Methods	Mean AP score (s.d.)
Visual only	39.94 ± 1.18
Visual+[3]	45.87 ± 1.73
Visual+MRW(0)	47.34 ± 2.00
Visual+MRW(1)	47.69 ± 2.20
Visual+MRW(2)	46.60 ± 2.24
Visual+MRW(all 3)	**48.10** ± 2.24

We used the SHOGUN toolbox [12] for SVM training. The regularization constant has been fixed to one as it did not have much influence on the performance, if kernels are properly normalized. All results were evaluated using the Average Precision (AP) for each concept class using 20-fold cross-validation.

4. Results

Table 4 compares the mean AP scores over all concepts. The first conclusion from this table is that adding tag information is beneficial for concept classification with the recently published [3] as well as with our proposed method. To see differences in more detail we have performed for each concept separately a Wilcoxons signed rank test on the differences of the AP scores between our proposed method 'Visual+MRW(all 3)' and the 'Visual only' baseline. The test is based on the twenty results from the cross-validation splits. Figure 2 shows that for almost all classes we observe gains by adding the textual features. Furthermore for a majority of classes the gains are statistically significant.

Secondly, we can see that our proposed method outperforms the state-of-the-art approach [3]. The AP difference between 'Visual+MRW(all 3)' and 'Visual+[3]' was more than 2 points.

We may ask whether each of the MRW kernels is very competitive for a certain class and fails for others, because each class requires a certain level of smoothing between tags. Note that the MRW kernels differ by their level of smoothing (MRW(0): unsmoothed vs MRW(2): more smoothed). It could explain the good performance of the average kernel. We tested this by choosing for each class the class-wise best output over all three kernels. This class-wise selection over all three kernels achieves an AP score of 48.04 which is very close to the performance of the average kernels. Thus, our hypothesis is supported.

Unfortunately, we cannot evaluate classifiers based only on user tags on the whole CLEF data set, because we have a certain amount of images without tags or the textural features. Therefore, we checked the classification performances of the textural kernels on the restricted subsets of images which have the textural features after preprocessing. We observed the mean APs of the range 31-35 for the textural kernels, while we achieved the AP score of 40 with the visual kernel. When we averaged all kernels, the AP performance of SVM increased up to 49.30. Although the visual features are most informative, the high gain of the combined classifier implies that the information from tags seems to be very non-redundant relative to that from the visual descriptors. Furthermore, computation time for the tag features took seconds, while the visual features required minutes.

5. Discussions

As we reported in the previous section, the difference in gain varies strongly over concepts - however in a seemingly plausible manner. Despite limitations of the learning capabilities of our approach, they give an insight about what is tagged poorly or well. It is expectable that very *abstract concepts* like 'Boring', 'Abstract' or 'Technical' (located south in Figure 2) do not receive a bonus from tagging as they are not the usual focus of user tagging, when no guidelines about the concepts to be classified are provided. The same applies to the exposition and quality concepts like 'Overexposed' (located northwest in Figure 2). We never heard anybody exclaim "Wow that is nicely overexposed!" This is consistent when regarding exposition and quality as well as the concepts 'Shadow', 'Clouds' and 'Sky' as belonging to a class of *background observations* which often fail to catch the eye of a spontaneous observer. Comparably low gains can be observed in concepts which describe the *absence of something* like 'No_Visual_Season', 'No_Visual_Place'. Tagging usually denotes the presence of something apparent. Very *subjective concepts* like 'Fancy', 'Visual_Arts' and 'Aesthetic_Impression' do not give rise to big gains as well. It is not surprising that children and babies get better tagged in terms of learnability than old persons. High gains are observed by animals in general followed by 'Winter', 'Snow', 'Flowers' and natural landmarks (north), architectural landmarks like 'Bridge' and 'Church' as well as objects in the southeast.

In order to illustrate how tags help to achieve better predictions, we included example images in Figures 3 and 4. The thresholds for binary classifications are determined by a standard way. We compared here the classifier only with the visual kernel to our best procedure with Visual+MRW(all 3) and listed the concepts which the two methods returned inconsistent predictions. The correct ones are marked with green checks, while the wrong ones are indicated by red crosses. It can be seen that most of the cases, the proposed method with tags removed wrong predictions by the baseline and add correct concepts further. The first image has a tag 'dog' which is included in the label set. However, 'Musical_Instrument', 'Food', 'Still_Life' are not included

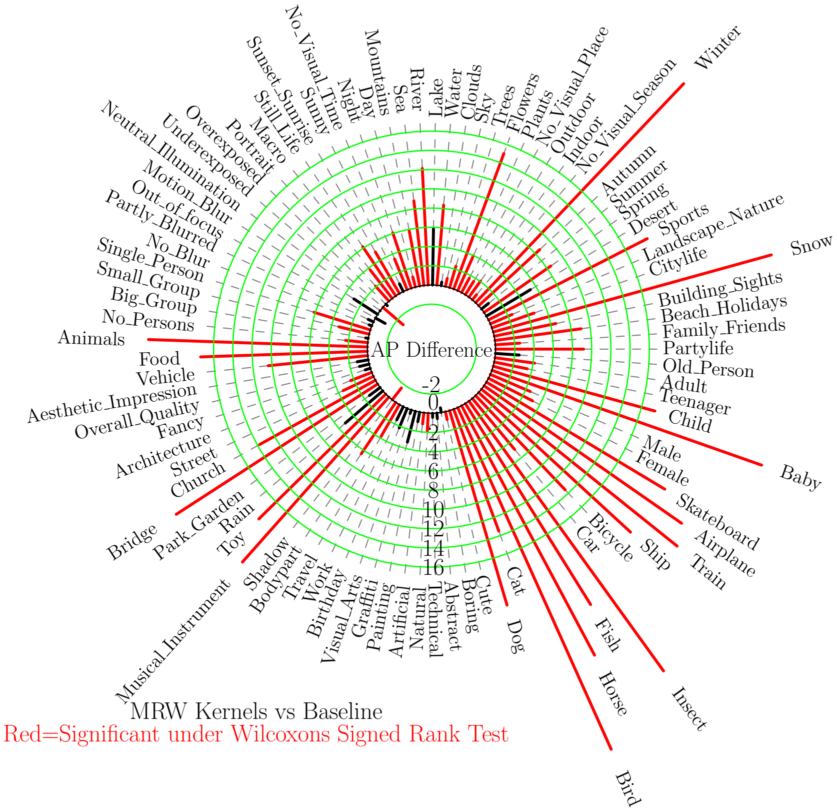
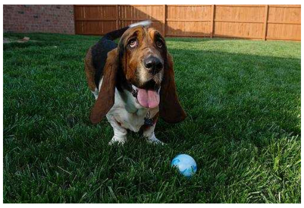
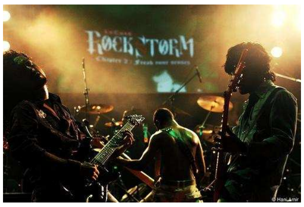


Figure 2. Class-wise AP score differences. Except for several classes with small losses, our procedure outperformed the baseline only with the visual kernels. Most of the gains are statistically significant.



Visual	Proposed	Tags	
	✓ dog	bassethound	basset
× Single_Person	✓ No_Persons	dog	hound
× male	✓ Animals	butters	murfreesboro
	✓ Sunny		



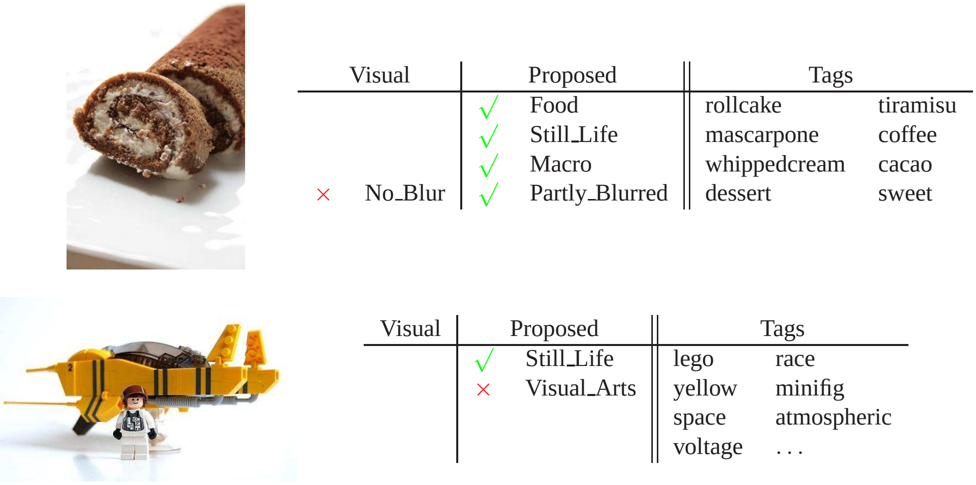
Visual	Proposed	Tags	
	✓ Musical_Instrument	guitar	bass
× No_Person	✓ Small_Group	live	music
	✓ male	rockstorm	metal
		maldives	. . .

Figure 3. Example images (1)

in the user tags. The relations between the labels and some of the relevant tags are learned by our multi-modal learning procedures.

6. Conclusions

In this paper, we proposed a smoothing technique for histogram-type representations with Markov random walk to alleviate their sparsity. By combining visual kernels and kernels from tag features with three different smoothness, we could outperform the state-of-the-art result [3] significantly. Further research should be done for applying more sophisticated model-based approaches to user tags.



Visual		Proposed		Tags	
		✓	Food	rollcake	tiramisu
		✓	Still_Life	mascarpone	coffee
		✓	Macro	whippedcream	cacao
×	No_Blur	✓	Partly_Blurred	dessert	sweet

Visual	Proposed		Tags	
	✓	Still_Life	lego	race
	×	Visual_Arts	yellow	minifig
			space	atmospheric
			voltage	. . .

Figure 4. Example images (2)

Acknowledgements

We thank Klaus-Robert Müller for valuable discussions. This work was supported by the Federal Ministry of Economics and Technology of Germany under the project THESEUS (01MQ07018), by the German Research Foundation (GRK 1589/1) and the FP7-ICT Programme of the European Community, under the PASCAL2 Network of Excellence, ICT-216886.

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[2] G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.

[3] M. Guillaumin, J. Verbeek, and C. Schmid. Multimodal semi-supervised learning for image classication. In *Proc. of IEEE Int. Conf. on Comp. Vis. & Pat. Rec. (CVPR '10)*, San Francisco, CA, USA, 2010.

[4] T. Hofmann. Probabilistic latent semantic indexing. In *Proc. of SIGIR Conf. on Research and Development in Information Retrieval*, pages 50–57, Berkeley, CA, USA, 1999.

[5] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. of IEEE Int. Conf. on Comp. Vis. & Pat. Rec. (CVPR '06)*, pages 2169–2178, New York, NY, USA, 2006.

[6] X. Li, C. G. Snoek, and M. Worring. Learning tag relevance by neighbor voting for social image retrieval. In *Proc. of ACM Int. Conf. on Multimedia Information Retrieval (MIR '08)*, pages 180–187, Vancouver, Canada, 2008.

[7] Y. Li, D. J. Crandall, and D. P. Huttenlocher. Landmark classification in large-scale image collections. In *Proc. of IEEE Int. Conf. on Computer Vision (ICCV '09)*, Kyoto, Japan, 2009.

[8] D. Lowe. Distinctive image features from scale invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[9] N.Craswell and M.Szummer. Random walks on the click graph. In *SIGIR Conf. Research and Development in Information Retrieval*, pages 239–246, 2007.

[10] S. Nowak. ImageCLEF photoAnnotation@ICPR2010. http://www.imageclef.org/2010/ICPR/, 2010.

[11] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[12] S. Sonnenburg, G. Rätsch, S. Henschel, C. Widmer, J. Behr, A. Zien, F. de Bona, A. Binder, C. Gehl, and V. Franc. The SHOGUN machine learning toolbox. *Journal of Machine Learning Research*, 11:1799–1802, 2010.

[13] A. Sun and S. S. Bhowmick. Image tag clarity: in search of visual-representative tags for social images. In *Proc. of ACM SIGMM Workshop on Social Media (WSM '09)*, pages 19–26, Beijing, China, 2009.

[14] M. Tahir, K. van de Sande, J. Uijlings, F. Yan, X. Li, K. Mikolajczyk, J. Kittler, T. Gevers, and A. Smeulders. SurreyUVA SRKDA method. http://pascallin.ecs.soton.ac.uk/challenges/voc/voc2008/workshop/tahir.pdf.

[15] J. Tang, S. Yan, R. Hong, G.-J. Qi, and T.-S. Chua. Inferring semantic concepts from community-contributed images and noisy tags. In *Proc. of ACM Int. Conf. on Multimedia (MM '09)*, pages 223–232, Beijing, China, 2009.

[16] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Trans. Pat. Anal. & Mach. Intell.*, 2010.

[17] H. Xu, J. Wang, X.-S. Hua, and S. Li. Tag refinement by regularized LDA. In *Proc. of ACM Int. Conf. on Multimedia (MM '09)*, pages 573–576, Beijing, China, 2009.