

Hierarchical Medical Image Annotation Using SVM-based Approaches

Igor F. Amaral, Filipe Coelho, Joaquim F. Pinto da Costa and Jaime S. Cardoso

Abstract—Automatic image annotation or image classification can be an important step when searching for images from a database. Common approaches to medical image annotation with the Image Retrieval for Medical Applications (IRMA) code make poor or no use of its hierarchical nature, where different dense sampled pixel based information methods outperform global image descriptors.

In this work we address the problem of hierarchical medical image annotation by building a Content Based Image Retrieval (CBIR) system aiming to explore the combination of three different methods using Support Vector Machines (SVMs): first we concatenate global image descriptors with an interest points Bag-of-Words (BoW) to build a feature vector; second, we perform an initial annotation of the data using two known methods, disregarding the hierarchy of the IRMA code, and a third that takes the hierarchy into consideration by classifying consecutively its instances; finally, we make use of pairwise majority voting between methods by simply summing strings in order to produce a final annotation.

Our results show that although almost all fusion methods result in an improvement over standalone classifications, none clearly outperforms each other. Nevertheless, these are quite competitive when compared with related works using an identical database.

I. INTRODUCTION

The image is one of the most important tools in medicine since it provides a method for diagnosis, monitoring drug treatment responses and disease management of patients with the advantage of being a very fast non-invasive procedure, having very few side effects and with an excellent cost-effect relationship. As new image acquisition devices are constantly being developed, to increase efficiency and produce more accurate information, and data storage capacity increases, a steady growth of the number of medical images produced can be easily inferred. A good example of this trend lies in the Radiology Department of the University Hospital of Geneva where, alone, produced from 12.000 medical images a day in 2002 [1] to 50.000 medical images a day in 2007 [2]. With such an exponential increase of medical data in digital libraries, it is becoming more and more difficult to execute certain analysis on search and information retrieval related tasks.

Because textual information retrieval is already a matured discipline, the use of captions for image indexation, where key description about its content and context is stored,

Manuscript received July 15, 2010.

I. F. Amaral is with INESC Porto, Faculdade de Ciências, Universidade do Porto, iamaral@inescporto.pt

F. Coelho and J. S. Cardoso are with INESC Porto, Faculdade de Engenharia, Universidade do Porto, {filipe.e.coelho, jaime.cardoso}@inescporto.pt

J. P. da Costa is with Departamento de Matemática, Faculdade de Ciências, Universidade do Porto, jpcosta@fc.up.pt

plays an important role in the categorization of a medical database. However, facing the amount of images therein, manual annotation is a time consuming and cumbersome task, prone to error and perception subjectivity. A study by Güld [3] reports around 15% of annotation errors in Digital Imaging and Communications in Medicine (DICOM) headers. Also, the whole DICOM header is very often lost as a consequence of image compression.

In the last few decades advances in Content Based Image Retrieval (CBIR) prompted researchers towards new approaches in information retrieval for image databases [4]. Although related but not yet developed as its textual obverse, CBIR is nowadays an independent and very active area of research where some success in constrained medical applications was already found. Still in problems where the image domain is broader only unsatisfactory results were achieved [5]. While classification of medical images based in purely visual features is still an ongoing challenge, developments in this field are expected to improve the performance of current retrieval engines and hospital information systems and, therefore, liable to be integrated in Picture Archival and Communicating Systems (PACS).

A. The IRMA code

Annotation of medical images require a nomenclature of specific terms to describe its content. The Image Retrieval for Medical Applications (IRMA) code [6] is a mono-hierarchical multi-axial standard for the purpose of medical image classification, designed to avoid incompleteness, ambiguity and lack of causality existent in its counterparts. It consists in four independent axes describing different content within the image: the technical (T) axis code describes the image modality; the direction (D) axis code describes body orientation; the anatomical (A) axis code for the body region examined; and the biological (B) axis code for body system examined (Fig. 1). The full IRMA code for one particular image consists in 13 characters (IRMA: TTTT-DDD-AAA-BBB).



IRMA: 1121-220-230-700

T - x-ray, projection radiography,

analog, overview image

D - sagittal, left-right lateral

A - cranium, neuro cranium

B - musculoskeletal system

Fig. 1. Example of a radiograph annotated with the Image Retrieval for Medical Applications (IRMA) code.

II. RELATED WORKS

The evolution of the state-of-the-art for automatic annotation of medical images methods based on purely visual features and using the IRMA code can be tracked in the cross-language image retrieval campaign (ImageCLEF) medical image annotation tasks¹ that took place from 2005 until 2009. These tasks aimed to explore, develop and promote automatic annotation techniques and strategies for semantic information extraction in medical images databases with little or no annotation, as well as benchmarking the purposed systems.

In 2005 [7] an Image Distortions Model (IDM) using images pixel based information and similarity distances achieved the best result for a database of 10.000 radiographs belonging to 52 classes without IRMA code annotation. Other methods that performed equally well consisted also in combinations of IDMs using pixel information and texture features as well as image patches and decision trees. For 2006 [8] Support Vector Machines (SVMs) were extensively used for annotation together with a diverse amount of image descriptors in a 11.000 radiographs from 116 classes once again without IRMA code annotation. However, it was a log-linear maximum entropy model using image patches that surpassed all submitted methods. The IRMA code was first introduced for a 12.076 radiographs database from 116 classes in 2007 [9]. This year, dense sampled local descriptors in a Bag-of-Words (BoW) with SVMs for annotation attained the best results. Very close to these were other approaches also relying on SVMs for image patches and local relational models. Nearest Neighbor (NN) classifiers for feature vectors based in a large quantity of descriptors clearly underperformed when compared with the best works presented. In 2008 [10] the number of images from the database increased to 13.000 IRMA code annotated radiographs belonging to 196 classes. From the few submitted runs to accomplish the task the most successful was again based in a BoW from local image features together with an SVM for annotation. Every tasks from 2005 until 2008 were again purposed for the final 2009 task [11], where an extensively trained SVM for image patches in a BoW won the competition, followed closely by the winner method for the 2008 task.

Previously, from 2005 to 2006, annotation of the medical databases provided consisted in a number representative of a class. Even with the introduction of the IRMA code in 2007 for this challenge, all approaches systematically ignored its hierarchy and focused only in two methodologies: considering the code as a whole class of objects, a *flat* annotation, or by considering each axis separately, a *axis-wise* annotation. This goes against the purpose of the challenges and the nature of the code itself. Also, methods relying in local image descriptors with SVMs outperform other methods based in global image descriptors. Yet, since the databases under analysis belong to a restricted domain, global image descriptors may still be suitable for medical image annotation

since feature extraction can be performed very quickly when compared with the BoW construction. As a result, since time efficiency vs. classification performance can be an issue for very large databases, there should be an effort in pursuit and test this type of image descriptors.

III. METHODOLOGY

Annotation of medical images in this work undergoes several stages: first we extract information from the images and form a feature vector; hence we train several SVMs to create a model from the data for annotation accordingly to the mentioned approaches, flat and axis-wise, and another approach herein tested: a *position-wise* method; finally we use majority voting, by summing strings, for a pairwise fusion between all three methods. A general flowchart of our procedure can be found in Fig. 2.

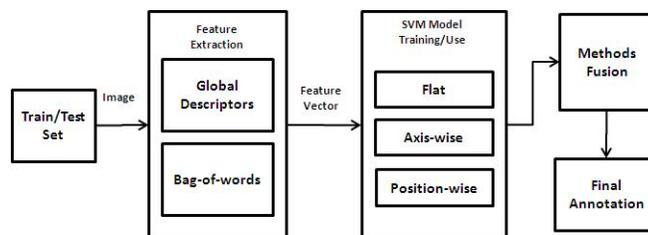


Fig. 2. Framework flowchart for our medical image annotation system.

A. Feature Extraction

To extract information from the images we used both global and a local image descriptor in a BoW approach. Feature selection was made accordingly to the desired image properties that we aimed to discriminate: *color*, *texture* and *shape*. All global descriptors were extracted using the Local and Web Image Retrieval Engine, *Img(Rummager)*², or code provided by the authors.

1) *Global image descriptors*: We used three MPEG-7 global descriptors - the *Scalable Color*, based on a histogram computed in the HSV color space and encoded by a Haar transform. It focuses on color representation, and its performance increases with the number of coefficients used. In this work, 64 coefficients were used for the Haar transform, resulting in a 64-value feature vector; the *Color Layout*, that attempts to overcome the limitations of global color features by dividing the image into blocks and extracting information from each one. It operates in the *YCrCb* color space and the implementation used returns a 12-value feature vector; the *Edge Histogram* (EH), that represents the spatial distribution of four directional edges and one non-directional edge. It uses sketch-like information from the image, which is semantically relevant for concept detection. The implementation returns an 80-value feature vector [12].

Moreover we used two composite descriptors that use fuzzy systems to combine more than one image property: the *Compact Color and Edge Directivity Descriptor*

¹<http://www.clef-campaign.org>

²<http://savvash.blogspot.com/2008/06/imgummager-in-now-available-for.html>

(CCEDD) [13], that uses a fuzzy version of the five digital filters proposed by the EH to extract texture information and two fuzzy systems that map the colors of the image in a 10-color custom palette. This results in a 60-value feature vector; the *Compact Fuzzy Color and Texture Histogram* (CFCTH) [14] captures color information using the same process as CCEDD and extracts texture information using the high frequency bands of the Haar wavelet transform in a fuzzy system. This descriptor returns a 80-value feature vector.

Other two image global descriptor were also used: the *Spatial Envelope* (GIST) descriptor, introduced in the context of scene recognition [15]. It describes the spatial layout by capturing features such as naturalness, openness, expansion, depth, roughness, complexity, ruggedness and symmetry. Images were resized to 256x256 pixels and the GIST descriptor was extracted for 32x32 non-overlapping sub-windows in 8 different directions, yielding a feature vector of 256 values; the *Tamura Textures* [16], a series of measures accounting for the coarseness, contrast and directionality texture information of an image which produces a 18 feature vector.

2) *Bag-of-words*: The *Speeded Up Robust Features* (SURF) local descriptor [17], a 128 bins histogram based in sums of two-dimensional Haar wavelets was extracted from a fast Hessian blob detector interest points to construct a 512 dictionary of visual words in a BoW [18]. The sensitivity of the detector was dynamically changed to detect between 256 and 512 points. A sample of 330.000 local descriptors from the training set were gathered and clustered without supervision into 512 centers using a k-means algorithm [19]. Then, the histogram representing the distribution of visual words for a particular image was built by assigning every SURF histogram for an interest point to its correspondent center bin using the NN with a Euclidian metric.

B. Annotation

For the annotation process we relied on SVMs [20] with a Radial Basis Function (RBF) kernel due to their performance in the 2007-2009 ImageCLEF medical image annotation tasks. We set up a framework in MATLAB[®] ³ using the popular LIBSVM [21] multi-class implementation. We performed an extensive grid-search on the common approaches to this problem, flat and axis-wise strategies, to optimize the kernel parameters (γ, C) using 10-fold cross validation. Because LIBSVM normalizes all input data, this search was performed around $1/k$ where k is the number of features. A total of 108 models were trained for each axis in the axis-wise method as well as the flat strategy. For γ we took into consideration the interval between one magnitude order immediately greater and smaller than $1/k$, for a total of 9 values, and for the cost we considered $C = 2^n$ where $n = \{-4, -3, \dots, 7\}$.

As a third strategy we introduced a position-wise method. Each image is classified one axis at the time but, unlike the axis-wise method, conceptualization of the image content

does not take the full meaning of the axis into consideration, i.e., we do not separate classes accordingly to the full axis code. Instead, we first consider the highest hierarchical position of the axis, its root, and use the whole training set to perform an initial classification. Afterwards, we reduce the training set to those images who match the initial classification, a *semantic reduction* of the training set, and classify the hierarchically subsequent inferior position. We undergo this top-down process thorough the axis tree until it is completely classified. We undertake the same methodology for all axes and assemble the final IRMA code. Using this method raised a problem in choosing the right (γ, C) parameters for the RBF kernel.

After the annotation from the three methods separately we make pairwise fusions of these by summing strings. During this phase it is possible to assign a *wildcard* ('*') when two annotations are non-consensual for a specific position, denoting a 'not know' meaning.

IV. EXPERIMENTAL RESULTS

For this work we use the IRMA 2007 medical image database (Fig. 3). This collection from the Department of Radiology at the RWTH Aachen University Hospital consists in 12.000 grayscale radiographs in Portable Network Graphics (PNG) format, divided in a 11.000 images training set and a 1.000 images test set, contain 116 unique classes of objects with intra-category similarity and inter-class variability. The distribution of classes is highly uneven to reflect the medical routine of a radiology department. However, in this database, each class has a minimum of 10 images. Furthermore, images were scaled proportionally to their original size, keeping aspect ratio, in order to fit a 512x512 maximum pixel window. The first three positions of the IRMA code T axis for all images within this database are the same, which reflects their modality similarity. This image collection is a subset of the larger 17.000 radiographs IRMA database.

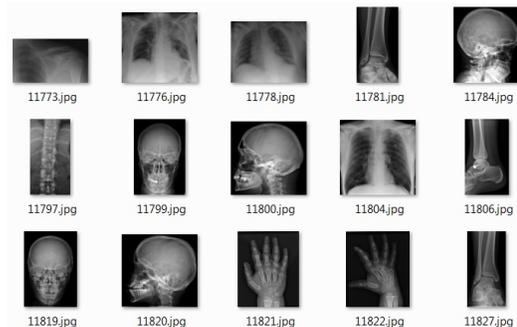


Fig. 3. Examples of images from the 2007 Image Retrieval for Medical Applications (IRMA) database.

A. Feature extraction

Concatenation of all descriptors used resulted in a 954-value feature vector. Even so, because some image descriptors act in different color spaces but operate in grayscale images, a total of 121 feature vector positions always possess

³<http://www.mathworks.com>

zero values. For the Tamura textures we also noticed that only the 16 bins directional histogram actually contained some information.

During the interest points detection, for the SURF descriptor extraction, some images could not reach a number between 256 and 512 points even at a low sensitivity threshold. For these images the maximum number of detections was considered, where the worst case comprehended roughly 100 interest points.

B. Annotation

The evaluation of the performance of the methods is based in two error measures: the *error rate*, the percentage of codes that have at least one error in one position within one axis, and the *error count*, a normalized error schema [22] that introduces a greater penalty for misclassification in higher hierarchical instances than for less precise classification in lower hierarchical positions. A wildcard assignment for a position implies a half maximum penalty in error count.

From the best RBF kernel parameters (Table I), standalone classification shows that for the test set the flat method outperforms the axis-wise and position-wise approaches (Table II). Such behavior is consistent with the results attained in related works for the same database [9]. Some preliminary experiences for a RBF kernel parameter optimization for the highest hierarchical position in axes D, A and B of the position-wise method showed that the best configuration is not far away from the parameters for the axis-wise method. Even so, with a similar parameterization to the axis-wise method, the position-wise method can achieve a good performance.

TABLE I
BEST PARAMETRIZATION RESULTS FROM RADIAL BASIS FUNCTION (RBF) KERNEL

	Flat	Axis-wise			
		T	D	A	B
γ	0.0001	0.00035	0.00035	0.00035	0.00035
Cost	16	4	4	4	4
Accuracy (%)	88,8	99,7	90,4	93,00	99,0

We noticed a large confusion between classes ‘1123-127-500-000’ and ‘1123-120-500-000’, responsible for roughly 5% of misclassifications in all methods. In the axis-wise and position-wise methods we also verified the existence of *semantic meaningless* codes. Since the annotation in these methods is performed on each axis independently, a misclassification in one axis can lead to a final IRMA code that does not exist in the spectrum of 116 possible classes. Curiously this fact remained unnoticed by previous works with the same database [9].

Using the detection of meaningless codes we proceed to pairwise fusion (Table III) from two ways: normal fusion and by replacing this codes for the correspondent flat classification in the flat/axis-wise and flat/position-wise fusions. Here, we expected that the replacement of such codes by an annotation method that uses a conceptualization of the image as a whole would lower the error count and rate. This,

TABLE II
ANNOTATION RESULTS FOR STANDALONE SUPPORT VECTOR MACHINE (SVM) MODELS

Model	Training set		Test set	
	Err. Cnt	Err. rate (%)	Err. Cnt	Err. rate (%)
Flat	30,8	11,9	31,4	13,3
Axis-wise	32,7	14,5	37,2	16,6
Position-wise	36,3	16,2	39,9	17,4

in fact, happened, but with less good results than expected. Since the contribution for the error count of each axis is 0.25 and because a meaningless code has at least one misclassified axis, we expected to decrease the error rate by $N_m/1000$, and in the error count by $0.25 \cdot N_m$, where N_m is the number of meaningless codes found. Obviously these expected results are only for replacements by correctly classified codes.

Results from fusion where the flat method is involved led to a decrease of error count but at an increase of error rate. This means that even if more wrong or ‘not known’ codes are introduced, their importance accordingly to the error count scheme is smaller. Fusion between axis based methods underperform due to the high error rate produced during standalone classifications. A comparison from the replacement or not of semantic meaningless codes in both training and test sets verifies that the error rate does decrease as expected but this is not reflected in a correspondent decrease of the error count. Therefore we can assume that there is a correlation between the existence of a semantic meaningless code for an image originated by the axis based methods and the existence of an error their flat counterpart annotation.

From all methods presented the lowest error rate is achieved in the flat annotation and the lowest error count results from the flat/position-wise fusion. Comparing this results when compared to similar works regarding the same database [9] it can be seen that our best result with an error count of 28.6 is close to the state of the art error count of 26.8. This would place flat/position-wise run as the 3rd best for all IRMA 2007 medical image database classification runs. However, while the state of the art presents an error rate of 10.3%, our best run has a much higher value of 17.4%. For other results close to this error rate, the error count ranged from 44.6 to 48.4. This difference of performance is even larger if we consider the normal Flat/Position-wise fusion, with an error count of 29.4 at an error rate of 20.6% with the error count of results with a similar error rate, where these are read between 51.3 and 72.4.

V. CONCLUSIONS AND FUTURE WORK

In this paper we presented a methodology for hierarchical medical image annotation based in three different approaches using global and local features together with SVMs. While we aimed to explore the hierarchy in a semantic reduction position-wise method it is clear that this strategy is very susceptible of error since a wrong classification in an early stage of the process implies subsequent wrong classifications in the hierarchically inferior positions. Nevertheless there

TABLE III
ANNOTATION RESULTS FOR SUPPORT VECTOR MACHINE (SVM) MODEL FUSION

Dataset	SVM Models	Normal fusion		Replace meaningless codes		
		Err. Cnt	Err. rate (%)	Err. Cnt	Err. rate (%)	N_m
Train	Flat/Axis-wise	26,9	16,1	28,3	14,6	36
	Flat/Pos-wise	27,7	18,6	27,8	15,7	47
	Axis-wise/Pos-wise	30,6	18,4	-	-	-
Test	Flat/Axis-wise	29,1	18,3	30,1	16	42
	Flat/Pos-wise	29,4	20,6	28,6	17,4	43
	Axis-wise/Pos-wise	34,9	20	-	-	-

is still room to improve this method in terms of feature selection depending on the axis or classification stage and SVM kernel configuration. For this we expect this method to perform better in databases where the number of unique classes is higher, like the IRMA 2008 database.

Fusion of annotations from different methods can also be target of further research towards relevance feedback mechanisms that decide whenever a fusion is worthwhile given the number of wildcards returned. To foster this analysis, the use of the error count as a measure of distance between two images is also a possibility.

For the semantic meaningless codes while it is certain that an annotation error exists, the exact location of the error is unknown. An important conclusion to be drawn from here is that aside the IRMA code is stated as possessing independent axis there seems to be, after all, a dependence among these in order to assemble a valid final code. Further work aiming to comprehend such dependences and the semantic meaningless behavior will most surely improve our results.

VI. ACKNOWLEDGMENTS

This work was funded by the Portuguese Innovation Agency (ADI) through project QREN reference 3472 “Semantic PACS”. The IRMA database is a courtesy of TM Deserno, Dept. of Medical Informatics, RWTH Aachen, Germany. We would like to thank Thomas M. Deserno and Petra Welter for the IRMA database support, Savvas Chatzichristofis for the support on Img(Rummager) engine and Pedro Quelhas and Hélder Oliveira for the insightful discussions on this work.

REFERENCES

- [1] H. Müller, N. Michoux, D. Bandon, and A. Geissbuhler, “A review of content-based image retrieval systems in medicine - clinical benefits and future directions,” *International Journal of Medical Informatics*, vol. 73, no. 1, pp. 1–23, 2004.
- [2] X. Zhou, A. Depeursing, and H. Müller, “Hierarchical classification using a frequency-based weighting and simple visual features,” *Pattern Recogn. Lett.*, vol. 29, no. 15, pp. 2011–2017, 2008.
- [3] M. O. Güld, M. Kohlen, D. Keysers, H. Schubert, B. Wein, J. Bredno, and T. M. Lehmann, “Quality of dicom header information for image categorization,” in *Intl. Symposium on Medical Imaging*, ser. Proc. SPIE, vol. 4685, San Diego, CA, Feb. 2002, pp. 280–287.
- [4] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, “Content-based image retrieval at the end of the early years,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, 2000.
- [5] L. R. Long, S. Antani, T. M. Deserno, and G. R. Thoma, “Content-based image retrieval in medicine: retrospective assessment, state of the art, and future directions,” *Int J Healthc Inf Syst Inform*, vol. 4, no. 1, pp. 1–16, January 2009.
- [6] T. M. Lehmann, H. Schubert, D. Keysers, M. Kohlen, and B. B. Wein, “The irma code for unique classification of medical images,” in *In: Medical Imaging. Volume 5033 of SPIE Proceedings*, 2003, pp. 109–117.
- [7] P. Clough, H. Müller, T. Deselaers, M. Grubinger, T. M. Lehmann, J. Jensen, and W. Hersh, “The CLEF 2005 cross-language image retrieval track,” in *Working Notes of the 2005 CLEF Workshop*, Vienna, Austria, 2005.
- [8] H. Müller, T. Deselaers, T. M. Lehmann, P. Clough, E. Kim, and W. Hersh, “Overview of the ImageCLEFmed 2006 medical retrieval and medical annotation tasks,” in *CLEF 2006 Proceedings*, ser. Lecture Notes in Computer Science (LNCS), vol. 4730. Alicante, Spain: Springer, 2007, pp. 595–608.
- [9] T. Deselaers, T. M. Deserno, and H. Müller, “Automatic medical image annotation in ImageCLEF 2007: Overview, results, and discussion,” *Pattern Recognition Letters*, vol. 29, no. 15, pp. 1988–1995, 2008.
- [10] T. Deselaers and T. Deserno, “Medical image annotation in imageclef 2008,” in *CLEF Workshop 2008 / Evaluating Systems for Multilingual and Multimodal Information Access*, Aarhus, Denmark, September 2009.
- [11] T. Tommasi, B. Caputo, P. Welter, M. O. Güld, and T. M. Deserno, “Overview of the clef 2009 medical image annotation track,” in *Proceedings of the 9th CLEF workshop 2009*, ser. Lecture Notes in Computer Science (LNCS), Corfu, Greece, September 2009.
- [12] T. Sikora, “The mpeg-7 visual standard for content description-an overview,” vol. 11, no. 6, pp. 696–702, June 2001.
- [13] S. A. Chatzichristofis and Y. S. Boutalis, “CEDD: Color and edge directivity descriptor: A compact descriptor for image indexing and retrieval,” in *ICVS*, ser. Lecture Notes in Computer Science, A. Gasteratos, M. Vincze, and J. K. Tsotsos, Eds., vol. 5008. Springer, 2008, pp. 312–322.
- [14] S. Chatzichristofis and Y. Boutalis, “FCTH: Fuzzy color and texture histogram-a low level feature for accurate image retrieval,” in *Proceedings of the 9th International Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS*, 2008, pp. 191–196.
- [15] A. Oliva and A. Torralba, “Modeling the shape of the scene: A holistic representation of the spatial envelope,” *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.
- [16] H. Tamura, S. Mori, and T. Yamawaki, “Textural features corresponding to visual perception,” *IEEE Transaction on Systems, Man, and Cybernetics*, vol. 8, no. 6, pp. 460–472, 1978.
- [17] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, “Surf: Speeded up robust features,” *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.
- [18] L. Fei-fei, “A bayesian hierarchical model for learning natural scene categories,” in *In CVPR*, 2005, pp. 524–531.
- [19] D. Pelleg and A. Moore, “Accelerating exact k -means algorithms with geometric reasoning,” in *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, S. Chaudhuri and D. Madigan, Eds. New York, NY: AAAI Press, August 1999, pp. 277–281.
- [20] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer, 1995.
- [21] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [22] T. Deselaers, J. Kalpathy-Cramer, H. Müller, and T. M. Deserno, “Hierarchical classification for imageclef 2007 medical image annotation,” 2007, available at <http://www-i6.informatik.rwth-aachen.de/~deselaers/imageclef07/hierarchical.pdf>.