

Dear reviewers,

Thank you kindly for your extensive comments on our manuscript. We value your input highly and have done our best to revise our manuscript in light of your comments. In the following, we address your comments in the order in which they were given in the letter from the editor to the authors.

**Reviewer #3:**

*The work described in this manuscript is an evaluation / application-type report. The authors apply a general multi-label classification method (PCT) they published elsewhere (Vens et al, Machine Learning 2008, earlier incarnations since late 1990s) to the problem of medical image classification (X-ray images in particular). The authors evaluate the utility of PCT across four different feature types, two ensemble methods, two different fusion architectures, and contrast it to SVM classification that does not utilize the multi-label structure of annotations. In the end, they conclude that a certain subset of features coupled with one ensemble approach and one level of fusion with PCT outperforms unstructured SVMs on two competition datasets.*

=>>>>

The major concern of Reviewer #3 is that the conclusions we make are overly general. In essence, we would not be justified in stating that ensembles of PCTs perform better than SVMs if we only show their superior performance under a very specific set of conditions. To paraphrase the comment by the reviewer, we should not conclude that PCTs outperform unstructured SVMs in general, if we only show that for a certain subset of features coupled with one ensemble approach, one type of fusion and two competition datasets of the same type, namely X-Ray images.

To address the last of these issues, i.e., the comparison on only one type of dataset, we include in the revised version of the paper a comparison on a new dataset consisting of general images. The conclusions drawn from experiments on this dataset are exactly the same as those for the X-Ray image datasets, thus adding evidence of the generality of our approach and the conclusions we make from the experimental evidence. The other concerns, namely on the subset of features, the type of ensemble approach and type of fusion, were already addressed in the original submission. We explore two ensemble approaches, namely bagging of PCTs and random forests of PCTs. Both are shown to have superior performance than the SVM based approach. Concerning the different sets of features, the conclusion on the superiority of PCTs over SVMs holds across all of the different types of features and their combinations and not only on a specific subset of features. The conclusions also hold regardless of the type of fusion used: that is both for low-level and high-level fusion, the relative performance of the PCTs and SVMs remains the same, with low-level fusion giving better performance overall.

We have now reformulated our statements in the conclusion section to make this clear.

*The work presented is certainly interesting in that it demonstrates a possible utility of structured*

*approaches over a somewhat comprehensive set of features. Yet, there are several major issues that concern me here:*

*1) To me, the paper demonstrates the utility of ensemble methods rather than structured approaches. Every single row (feature/fusion combination) in Tables 1 & 2 shows that PCT alone underperforms (significantly) compared to the unstructured SVM. Only with the addition of ensemble methods does the structured approach gain ground. Unfortunately, the authors do not pursue the next obvious step: contrast ensembles of PCTs to ensembles of SVMs. Only this setting would more definitely demonstrate the utility of structured approach. Of course, one could have also considered structured SVMs here, which one should comment on.*

*=>>>>*

When we designed the experimental setup, our goal was to compare the performance of ensembles of PCTs with the approach that is most widely used by the image annotation community: SVMs (Mensink et al. 2010). It is true that the SVMs outperform single PCTs. However, we offer four explanations as to why we compare ensembles of PCTs and SVMs.

First, we would like to note that the SVMs (trained in a one-vs-all strategy) are already an ensemble that consists of  $|C|$  classifiers, where  $|C|$  is the sum of nodes in the annotation hierarchy. The SVMs in this setting can be viewed as an ensemble that consists of  $|C|$  classifiers.

Second, Vens et al. (2008) show that training PCTs per class is inferior to a PCT for the whole hierarchy. Thus, we use PCTs for the whole hierarchy as a base classifier.

Third, the ensembles are able to lift the predictive performance of a single classifier in the case of classification and regression. While it is well known that ensembles lift the predictive performance of a single classifier in the case of classification and regression trees, it is not obvious that the lift carries over to PCTs for predicting structured outputs (HMC in our case). In the case when the base classifiers are decision trees, Bauer and Kohavi (1999) conclude that the increase in performance is related to the trees being unpruned, i.e., overfitting. On the other hand, Blockeel et al. (2006) state that PCTs for HMC overfit less than the single classification approach. Having in mind these two conflicting influences, it is not obvious whether an ensemble of PCTs will significantly increase the predictive performance of a single PCT. Moreover, the use of PCTs for HMC (and ensembles thereof) has not been investigated in the context of image annotation.

Fourth, the machine learning community hasn't reached a consensus whether and how ensembles of SVMs should be constructed. To begin with, the literature suggests that bagging gives best predictive performance when unstable learners are used as base classifiers (such as decision trees and neural networks). An unstable classifier is the one that will change greatly, when a small change in the learning set occurs (Breiman, 1996). Next, a theoretical and empirical evaluation of ensembles from SVMs is performed in (Evgeniou 2000; Evgeniou et al., 2000). There they consider two types of ensembles: bagging of SVMs (each SVM is constructed on bootstrap replicate) and voting SVMs (each SVM is constructed using different kernel and on different feature sub-space). The findings of this study, in this context were: "... with appropriate

*tuning of the parameters of the machines, combining SVMs does not lead to performance improvement compared to a single SVM.*” and that “*With accurate parameter tuning (model selection) single SVMs and ensembles of SVMs perform similarly.*” On the other hand, there exists some approaches that justify the usage of SVMs in the context of an ensemble (Hyun-Chul et al., 2003; Valentini and Dietterich, 2002; Wang et al., 2007). Valentini and Dietterich (2002) consider bagging of low-bias-SVMs and heterogeneous ensembles and combination of SVMs with different kernel parameters (in their case RBF kernel with different  $\sigma$ ). Hyun-Chul et al. (2003) consider bagging and boosting of SVMs. Wang et al. (2007) first perform clustering of the instances. Then small quantities of representative instances from the clusters are chosen as training subsets to construct the SVMs. However, these works are done typically in the context of binary or multi-class classification and mainly on a small number of domains (typically three per study) UCI domains. Moreover, there are also practical implications in terms of efficiency of such an ensemble, especially when a prediction for unseen example needs to be generated. Since in our work we do perform parameter tuning for the SVMs, we believe that making an ensemble of SVMs will not bring further (significant) improvements of the predictive performance. To the best of our knowledge, these (and similar to them) approaches are not used by the image annotation community.

There exist few implementations of structured SVMs. However, the most well-known, such as *SVM-struct* (Joachims 2010; Tsochantaridis et al., 2004), do not offer facilities for HMC. Those that do, are very recent (Gärtner and Vembu, 2009), have high computational complexity and are not used by the image annotation community. We compare our performance to image annotation approaches that are currently state-of-the-art in this area (Guillaumin et al., 2009; Makadia et al., 2008; Mensink et al., 2010). We show that our approach exhibits superior performance over these approaches (see also response to Reviewer #2).

E. Bauer, and R. Kohavi (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine Learning*, 36(1), 105-139.

H. Blockeel, L. Schietgat, J. Struyf, S. Dzeroski, and A. Clare (2006). Decision trees for hierarchical multilabel classification: A case study in functional genomics, *Knowledge Discovery in Databases: PKDD 2006*, LNCS vol. 4213, pp. 18-29.

L. Breiman (1996). Bagging Predictors, *Machine Learning* 24(2), p. 123–140

T. Evgeniou (2000). Learning with kernel machine architectures, PhD thesis, Massachusetts Institute of Technology - MIT, 2000

T. Evgeniou, L. Perez-Breva, M. Pontil, and T. Poggio (2000). Bounds on the generalization performance of kernel machines ensembles, In *Proceedings of 17th International Conference on Machine Learning*, Stanford, California

T. Gärtner, and S. Vembu (2009). On structured output training: hard cases and an efficient alternative. *Machine Learning* 76(2-3), p. 227-242

M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid (2009). Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation, *International Conference on Computer Vision*, 309–316, 2009

K. Hyun-Chul, P. Shaoning, J. Hong-Mo, K. Daijin, and B. Sung Yang (2003). Constructing support vector machine ensemble, *Pattern Recognition* 36(12), p. 2757-2767

- T. Joachims (2010). SVMstruct - Support Vector Machine for Complex Outputs, web page: [http://www.cs.cornell.edu/People/tj/svm\\_light/svm\\_struct.html](http://www.cs.cornell.edu/People/tj/svm_light/svm_struct.html), accessed on 17.11.2010
- A. Makadia, V. Pavlovic, and S. Kumar (2008). A New Baseline for Image Annotation. Computer Vision – ECCV 2008, LNCS vol. 5304, pp. 316-329
- T. Mensink, G. Csurka, F. Perronnin, J. Sanchez, and J. Verbeek (2010). LEAR and XRCE's Participation to Visual Concept Detection Task - ImageCLEF 2010, CLEF (Notebook Papers/LABs/Workshops), 2010
- I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun (2004). Support Vector Learning for Interdependent and Structured Output Spaces, Proceedings of 21st International Conference on Machine Learning, ICML
- G. Valentini, and T. G. Dietterich (2002). Bias—Variance Analysis and Ensembles of SVM, LNCS vol. 2364, 2002, pp. 27-38
- C. Vens, J. Struyf, L. Schietgat, S. Džeroski, and H. Blockeel (2008). Decision trees for hierarchical multi-label classification, Machine Learning 73(2), p. 185–214
- C. Wang, H. Yuan, J. Liu, T. Zhou, and H. Lu (2007). A Novel Support Vector Machine Ensemble Based on Subtractive Clustering Analysis, LNCS vol. 4426, pp. 849-856

2) *What is the overall **\*general\*** message of this work? One may conclude (putting aside my comments above) that on Xray images PCT with ensembles has a certain benefit over SVMs. But what are the drawbacks of PCT, in the context of images? Would this setting with equal conclusions generalize to other image classification problems? Why does this particular feature combination outperform others? (A consequence of the spatial pyramid choices, number of words in the SIFT formulation,...) Why the low level fusion? Basically, the discussion section restates results from the two tables but offers no other insightful discussions to the reader that would shed light on the inner working of the approach nor help him/her generalize conclusions to other contexts. This seriously undermines the value of this manuscript.*

=>>>>

We consider the question on what is the overall general message of this work to be the crucial question by Reviewer #3. To address this question, we have reformulated the conclusions section to give a direct answer and we have expanded the results and discussion section. In short, the overall general message of our work is that ensembles of PCTs for HMC are a superior alternative, both in terms of performance and in terms of efficiency, to the most commonly used approach in image annotation, that is collections of SVMs.

Our conclusions are general since we explore the two approaches under a wide range of conditions. To begin with, we consider three different datasets: two medical X-ray images and one general photos. Next, we consider several state-of-the-art feature extraction approaches and combinations thereof. Furthermore, we consider two types of feature fusion, i.e., low- and high-level fusion. All in all, our approach shows better performance under all of the mentioned conditions, both in terms of predictive performance (Tables 1, 2 and 4) and efficiency (Table 3).

The discussion section primarily focuses on answering the questions stated in the “Experimental questions” section (Section 5.3 from the manuscript). We have expanded the discussion section to include explanations and clarifications on the issues raised by the reviewer,

such as drawbacks of PCTs, why low-level fusion, computational complexity. The discussion section now provides further comments on the sources of difference in the performance. Moreover, to show the generality of our approach, we have performed experiments on a database with general photos. The conclusions are the same as for the medical images.

3) *I find the manuscript somewhat sloppily put together.*

*a) Most of the text is the discussion of features and the experimental evaluation (again, setting aside my comments in 2 above.) PCT is described in one paragraph that tells one nothing about how PCT is actually constructed, neither for training nor (especially) in query evaluation. The authors refer to their work in ML journal. Even their preceding conference work (SIKDD'08) had more details about the PCT framework. For completeness I certainly would like to see a more comprehensive description of the PCT algorithm in this manuscript.*

=>>>>

Done. We have added a new section “Predictive clustering trees” where we explain in more details the Predictive clustering trees framework.

*b) There are a number of places where the notation is not clear, terms are not defined, or essential information is missing. For instance:*

*- All references are missing titles*

=>>>>

Done. Corrected, the problem was the wrong class in the tex file.

*- Is spatial pyramid used for EHD? This is not stated in the EHD section but later on (p. 15) the authors claim that spatial pyramid is used for EHD features.*

=>>>>

Done. We have corrected this, for the EHD descriptor we didn't use spatial pyramid.

*- p. 22: "highest probabilities reported from the SVM classifier" - where do prediction probabilities of SVM come from? Platt-type normalization?*

=>>>>

For the general photo annotation we used Platt's probabilities, while for defining the threshold in the medical image annotation experiments we are using the distances of the test sample to the hyperplanes.

*- What is the asterisk notation in 6b? "don't care"?*

=>>>>

Done. We modified this.

*- Please restate what LL and HL means in table captions.*

=>>>>

Done.

*4) The authors state that one benefit of PCT vs SVM is scalability (at training time). There is very little empirical evidence presented, eg. running times. Also, it is the ensemble of PCTs that outperforms the SVM and, at evaluation time, the ensemble has additional overhead over the single SVM evaluation. It would be worth commenting more explicitly on this.*

=>>>>

We have included a new Table (namely Table 3 from the manuscript) with total training time and test time per image for all descriptors and their combinations for all of the considered learning methods. These results now clearly show that random forests of PCTs are much more efficient than SVMs both in terms of training time and testing time.



**Reviewer #2:**

*This paper presents a multi-label classification system for medical image annotation. The proposed system is mainly based on the ensemble of the predictive clustering trees and four different visual feature extraction methods are also applied to the system. The experiments are conducted on IRMA database for performance evaluation and the experimental results show that the system outperforms the ordinary SVM based approach. The topic of medical image annotation is very interesting and the paper has indeed conducted some interesting experiments, for instance, fusion different visual features for multi-label classification, however, the novelty of the paper seems insufficient for publication in the journal. The technical contribution of the paper is modest: the presented ensemble method seems simple; the method of predictive clustering tree had been published in previous literature; and the four visual feature extraction methods are also well known for the community.*

=>>>>

This paper presents contributions to the fields of ensemble learning, predicting structured outputs and image annotation. First, the performance lift from a single PCT to an ensemble of PCTs does not follow automatically, as explained bellow. In this work, we show that ensembles can lift the performance of their base classifiers even in the case when the output is a structure. Next, we show that the methods that exploit the structure of the output can perform better than the methods that perform flat classification. Here, we emphasize the last contribution: image annotation. We focus on the selection of the appropriate feature extraction technique for medical images and their combinations. We present novel results that show that some other classifiers (than the typically used SVMs) can perform better, not only in terms of efficiency but also in terms of predictive power. The results from the experiments offer new insights in the area of medical image annotation. Furthermore, we demonstrate the generality of the proposed method by comparing its performance with state-of-the art approaches on a recent database with general photos.

While it is well known that ensembles lift the predictive performance of a single classifier in the case of classification and regression trees, it is not obvious that the lift carries over to PCTs for predicting structured outputs (HMC in our case). In the case when the base classifiers are decision trees, Bauer and Kohavi (1999) conclude that the increase in performance is related to the trees being unpruned, i.e., overfitting. On the other hand, Blockeel et al. (2006) state that PCTs for HMC overfit less than the single classification approach. Having in mind these two conflicting influences, it is not obvious whether an ensemble of PCTs will significantly increase the predictive performance of a single PCT. Moreover, the use of PCTs for HMC (and ensembles thereof) has not been investigated in the context of image annotation.

E. Bauer, and R. Kohavi (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine Learning*, 36(1), 105-139.

H. Blockeel, L. Schietgat, J. Struyf, S. Dzeroski, and A. Clare (2006). Decision trees for hierarchical multilabel classification: A case study in functional genomics, *Knowledge Discovery in Databases: PKDD 2006, LNCS vol. 4213*, pp. 18-29.

*Automatic image annotation is an important topic and has been studied for nearly decade. So, there is a great deal of literature on the topic and some of the state-of-the-art approaches are on multi-label classification for image annotation. Although, this paper deals with the problem on medical image annotation, it is still strongly related with the general problem. But, there is the lack of sufficient related literature reviews in the paper. For the performance evaluation, the proposed system is not compared with the state-of-the-art image annotation approaches, such as "A.Makadia et al. A New Baseline for Image Annotation", "M. Guillaumin et al. Tagprop. " and other multi-label classification based approaches.*

=>>>>

To address this valid comment, we have added a new section “Related work” and performed additional evaluation on a database with general photos. In this section we give a short overview of the current state-of-the-art work in the field of image annotation. Considering the mentioned papers, we would like to point out a recent study performed by Mensink et al. (2010) which showed that per-label-trained-linear SVM classifiers outperform TagProp (Guillaumin et al. 2009). Moreover, Guillaumin et al. (2009) show that TagProp outperforms the system presented in Makadia et al. (2008). Furthermore, the best results on the current challenges/competitions detection and annotation tasks, such as the PASCAL Visual Object Classes challenge, the ImageCLEF medical image annotation task and the ImageCLEF visual concept detection and annotation tasks are obtained using binary classifiers for each visual concept. As binary classifier, they usually use SVM with  $\chi^2$  kernel, which is the baseline in our case.

We also performed additional experiments on the ImageCLEF@ICPR2010 database and compare our results with the results obtained using SVMs with  $\chi^2$  kernel. On this database, random forests of PCTs outperform SVMs both in terms of predictive power and efficiency. Since, SMVs with  $\chi^2$  kernel outperform TagProp (Mensink et al., 2010), we can conclude that our method also outperforms both TagProp (Guillaumin et al, 2009) and the baseline from Makadia et al. (2008).

T. Mensink, G. Csurka, F. Perronnin, J. Sanchez, and J. Verbeek (2010). LEAR and XRCE's Participation to Visual Concept Detection Task - ImageCLEF 2010, CLEF, 2010

M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid (2009). Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation, International Conference on Computer Vision, 309–316, 2009

A. Makadia, V. Pavlovic, and S. Kumar (2008). A New Baseline for Image Annotation. Computer Vision – ECCV 2008, LNCS vol. 5304, pp. 316-329

*The writing of the paper is generally understandable, but the style of the reference is uncommon, all the titles of the reference papers are missing.*

=>>>>

Done. The problem was in the wrong class in the tex file, this problem is corrected.



**Reviewer #1:**

*This is a well written paper with good experimental design and reporting. My only issue with the paper is the significance of the problem being addressed. All medical imaging modalities following DICOM standards allow the entering the information recovered from the PR system, and is a mandatory component of all clinical protocols that I am aware of. More evidence on the need for automating this type of annotation would seem to be needed. What is the probability of operator error relative to the method presented in the paper? Would this be used for retrospective annotation (pre DICOM) or for verification purposes? More context on the problem and how the proposed solution fits into the clinical workflow would improve the paper.*

=>>>>

We have added text in the introduction that clarifies and explains the issues raised by this reviewer. The text is along the following lines:

“Automatic image annotation or image classification is an important step in image retrieval. In the medical domain, using information directly extracted from images to annotate/categorize them will improve the quality of image annotation in particular, and more generally the quality of patient care. Properly classified medical image data can help medical professionals in fast and effective access to data in their teaching, research, training, and diagnostic problems. The results of the classification step can also be used for multilingual image annotation as well as for DICOM header correction.

Automatic image annotation can be used for retrospective annotation (pre DICOM). It can also be used as help for human annotators (i.e., radiologists), where the annotations that are suggested by the system are corrected/verified/confirmed by the human annotator. The limits of performance of an automated annotation system that learns from example images annotated by humans, is the rate /probability of operator error/agreement of annotators.

Automatic image annotation uses a computer system which automatically assigns metadata in the form of captions or keywords to a digital image. Typically, image analysis first extracts feature vectors. Together with the training annotations, they are then used by a machine learning algorithm to learn to automatically assign annotations. The performance of the computer system largely depends on the availability of strongly representative visual features, able to characterize different visual properties of the images, and the use of effective algorithms for training classifiers for automatic image annotation.”