A Framework For Evaluating And Comparing Automatic Image Annotation Algorithms

Konstantinos Athanasakos, Vassilios Stathopoulos and Joemon M. Jose
{athanask, stathv, jj}@dcs.gla.ac.uk

Department of Computing Science, University of Glasgow,
Sir Alwyn Williams Building, Lilybank Gardens, Glasgow G12 8QQ, UK

Abstract Several Automatic Image Annotation (AIA) algorithms have been introduced recently, which have been found to outperform previous models. However, each one of them has been evaluated using either different descriptors, collections or parts of collections, or "easy" settings. This fact renders their results non-comparable, while we show that collection-specific properties are responsible for the high reported performance measures, and not the actual models. In this paper we introduce a framework for the evaluation of image annotation models, which we use to evaluate two state-of-the-art AIA algorithms. Our findings reveal that a simple Support Vector Machine (SVM) approach using Global MPEG-7 Features outperforms state-of-the-art AIA models across several collection settings. It seems that these models heavily depend on the set of features and the data used, while it is easy to exploit collection-specific properties, such as tag popularity especially in the commonly used Corel 5K dataset and still achieve good performance.

1 Introduction

During the last decade, we have witnessed a major transformation of the digital multimedia information field. A lot of effort has been invested in identifying modern and efficient ways of browsing, navigating and retrieving multimedia data, while the traditional challenge of bridging the semantic gap [1] remains unsolved. The ultimate goal of understanding multimedia content requires us to identify a way to effectively combine low-level features in order to reach a high-level understanding of objects and semantics portrayed in an image. A question arises however, as to whether a correlation between these two levels actually exists.

Automatic Image Annotation (AIA) attempts to learn the afore-mentioned correlation and build a dictionary between low-level features and high-level semantics [2]. The idea is to use a manually annotated set of multimedia data in order to train a system to be able to identify the joint or conditional probability of an annotation occurring together with a certain distribution of multimedia content feature vectors. Two ways have been suggested. The first one is using supervised machine learning techniques in order to classify an image into predefined categories. In this approach, the class and the non-class model have to be

defined for each class (category) of a collection. However, many risks arise, related to the number and nature of classes, the size and diversity of the training set, and images that may harm the descriptive model of a class. The second way is to use unsupervised classification. In this approach, a clustering algorithm such as k-means is used to identify a set of clusters from feature vectors extracted either globally or locally from images. The difficulties in this approach are related to deciding on the number of clusters and fine-tuning the model's parameters. These systems are most always computationally expensive and have excessive resource requirements.

However, such models [2,3,4,5] have been traditionally compared only on the "easy" dataset provided by Duygulu et al. [2]. Some of them [4,5] have been evaluated on more realistic collections as well, such as the TrecVid News dataset in order to support certain assumptions and statements regarding real-life multimedia collections. However, it is unclear whether the reported results are due to the descriptive power of the model, or are simply artifacts of the discriminating power of the employed descriptor in combination with the collection. We argue that a more comprehensive evaluation of AIA models is needed in order to show that the models' assumptions actually hold and that results are neither collection nor descriptor-specific. In light of this, a framework for evaluation and comparison of AIA models is presented in this work, which incorporates various collections and established content descriptors. We have used this framework to evaluate and compare two state-of-the-art image classification models, namely the Multiple Bernoulli Relevance Model (MBRM) [4] by Feng et al. and the Supervised Multiclass Labelling (SML) introduced by Carneiro et al. [5]. Our findings reveal that both models highly depend on the evaluation data. Moreover, a simple SVM approach using Global Features significantly outperforms the two models, suggesting that results presented thus far are due to the evaluation settings and not due to the algorithms themselves.

The rest of this paper is organised as follows. In Section 2, we refer to several AIA algorithms that have been introduced in the literature and we discuss their evaluation strategies. In Section 3, we describe our approach, namely the Evaluation Framework which we propose and which was used during the evaluation of the two AIA models. In Section 4, we present and analyse the results regarding the evaluation of these models, while in Section 5 we draw a conclusion discussing our findings, the limitations of this work and future work in this domain.

2 Related Work

In this section, we provide a survey of AIA models along with several remarks regarding their evaluation methodologies. Each one of these models attempts to incorporate underlying principles behind the generation, structure and organisation of a multimedia collection.

The first attempt to learn a way to automatically annotate images was in 2002 by Duygulu et al. [2]. They essentially created a lexicon which associated

terms with feature vectors. The model was evaluated on the Corel 5K collection. The dataset was made publicly available to allow reproducibility of the results and comparison with other systems. In 2003, Blei et al. introduced Latent Dirichlet Allocation [6] in an attempt to address the problem of a single document being associated with more than one latent topics. Later in the same year, they proposed a new model called Correspondence-LDA [7], in which the generation of an image's multimedia content was conditional on the underlying topics that generated it. The Corr-LDA was evaluated again on the Corel 5K, but using a different part of it and also different descriptors.

Lavrenko et al. in 2004 suggested a model called Continuous-space Relevance Model (CRM) [3] which assumed that each region of an image was conditional on the rest regions. This actually means that each region is generated based on its context, while later in the same year, Feng et al. through their Multiple Bernoulli Relevance Model (MBRM) [4] improved the previous model in order to be able to handle video collections and to be more suitable for multimedia collections with more realistic annotation distributions. Non-parametric models, such as the MBRM and the CRM [3] do not include a learning phase, rather attempt to estimate either Image-to-Image or Image-to-Class similarity. As stated by Boiman et al. in [8], probably the most important advantage of non-parametric models is that they do not require a training phase, which makes them ideal for dynamic datasets, in which learning-based models tend to require extensive periods of time while tuning class parameters. Also, the lack of a learning phase eliminates risks related to parameter overfitting. Their main disadvantage however lies in the huge gap in annotation time between these two classes of models. Both CRM and MBRM were evaluated on the dataset by Duygulu et al., while the MBRM was also evaluated on the TrecVid News collection.

Carneiro et al. in [5] use a variation of Mixture Models, which is introduced by Vasconcelos and Lippman in [9]. The scheme which is proposed is called Hierarchical Mixture Models and involves hierarchically clustering at first the actual data and then the clusters of one level in order to proceed to the next one. Regardless of the type of data and the application, the idea is rather promising, since in order to proceed to the next level, only the previous level's parameters are required. This significantly reduces the execution time of the Expectation-Maximisation process. With respect to its evaluation strategy, the SML was tested on the Corel 5K and Corel 30K collections and also on a less usual evaluation protocol suggested in [10].

These few examples of this kind of models reveal the various motivations and the various challenges AIA researchers are trying to tackle. There is however a problem with all of these models which is related to their evaluation. A review of their evaluation methodologies reveals some flaws and also a knowledge gap in this field. Some of these models [2,7,11,3,4,5] were compared on an unrealistic setting using a specific dataset from the Corel 5K collection provided by Duygulu et al. [2] in 2002. Kwasnicka and Paradowski [12] also compared several AIA methods on the Corel 5K collection, although their focus was more on the evaluation measures. However, as suggested by Westerveld and de Vries in

[13], the Corel dataset is far too "easy", while the TrecVid datasets essentially comprise an effort to build more realistic collections. Nevertheless, none of these models were directly compared to other models using these enhanced collections, since this would be an expensive and time-consuming procedure requiring the implementation of other models as well and carrying out more experiments. On the other hand, some models, such as the Corr-LDA [7] were not directly compared to any previous models. Moreover, although the SML has achieved the best performance so far on the Corel 5K dataset, we do not have enough evidence to support that this is due to the model and that SML would outperform previous models in other settings as well. Especially with the Corel 5K dataset, it would be easy to exploit collection-specific properties and still get good results. As such, we cannot be certain as to whether models are robust and independent of their setting and whether some perform better because of their descriptive ability or because of the discriminating ability of the features sets used.

3 Evaluation Framework

In this section, we describe the Evaluation Framework which we propose to be used for the evaluation of already introduced and future Automatic Image Annotation algorithms. It essentially defines a set of test collections, a sampling method which attempts to extract normalised and self-contained samples, a variable-size block segmentation technique with varying degrees of overlapping and a set of multimedia content descriptors. These are all discussed in details in the following sections.

3.1 Multimedia Collections

A very common challenge related to image classification algorithms and machine learning methods in general is the fact that these are usually dependent on the data on which they are applied. This actually means that their performance and discriminating ability varies significantly depending on the test collection which is used each time. In the case of image classification algorithms, the setting on which such an algorithm might be evaluated consists of a multimedia collection and the kind of features that will be used to represent its images.

Regarding multimedia collections, facts such as whether images depict single or multiple objects, and whether an annotation implies dominance of an object or simply its presence are some examples of these factors. Moreover a collection could be strongly or weakly labelled, depending on whether all instances of an object are annotated or not, while the existence of object hierarchies having tags such as "cat" and "tiger", "car" and "exotic car" or "water" and "ocean" might not only affect the performance of the algorithm, but also the results that one would expect. Collections also define the level of semantics that an algorithm should target for. Searching for objects is a totally different task than searching for scene categories or emotional states. It would perhaps require a different way of

treating images, namely segmenting and representing, thus again modifying the overall setting on which the algorithm would have to operate.

As such, an evaluation of a set of image classification algorithms would simply be incomplete, if it did not involve testing these algorithms on various settings in order to prove their robustness, namely whether they perform equally well under various settings. Therefore, a set of three multimedia collections was selected to be incorporated in our evaluation procedure. These are the Corel 5K [2], TrecVid 2007 [14] and Caltech 101 [15] collections.

Corel 5K is considered a rather easy setting, since Global Colour Features alone are considered to provide enough discriminative power for this collection. It was first used by Duygulu et al. [2] in the field of automatic image annotation algorithms, while since then, it has been used by each new model in the literature, in order for the results to be comparable to previously proposed models. The TrecVid 2007 dataset on the other hand comprises an extremely challenging setting. Since it is intended to be used for several high level tasks such as shot boundary detection and high level feature extraction, one can appreciate that using this dataset in the AIA domain will be equally difficult and unpredictable. Caltech 101 has a major advantage over other multimedia datasets, in that each image depicts a single object, thus removing any confusion associated with the multiple-labels paradigm. As such, it can be employed to learn precisely the class and non-class model of certain categories and objects. Although the categories are not described by the same number of images, the fact that images belong to only one category each allows for a sample which is fair towards all categories, namely it has the same number of images describing each category, while still being consistent and self-contained.

It is obvious that these collections present various settings ranging from controlled, "laboratory" ones to more realistic collections incorporating issues such as statistically unbalanced tag distributions, weak labelling and so on. Ideally, an AIA algorithm should be able to cope with all of the various challenges present in the afore-mentioned collections. However, no algorithm has been found and proved to meet this condition. In addition, as suggested by Westerveld and de Vries in [13], we might have to consider different performance measures in terms of granularity depending on the difficulty level of a collection.

3.2 Sampling Procedure

In this paper, the afore-mentioned collections were not used as a whole, rather we used a sampling procedure to extract a smoother and self-contained representative sample of each collection. By smoother, we mean that most of the tags would contain approximately the same number of images, and only a few, if any, would be described by significantly more example images. By self-contained, we mean that we would not discard any instances of the sampled classes which were included in the sampled images, as this would harm their class and non-class models. This sampling process was performed for two reasons. First, using the whole collections would require an immense amount of time to complete evaluating these algorithms, as in the case of memory-based models like MBRM which

require examining the whole training set each time a test image is being classified, while at the same time, it would not add significant value to the validity of our experiments. Second and more importantly, all of these collections have a highly unbalanced distribution of images over classes. There are a lot of classes which are inadequately described, a set of classes with a reasonable number of images belonging to them and a few which are very popular and frequent within each collection. Using the whole collections would probably create an easier setting for all of the algorithms for two reasons. When evaluating such an algorithm, popular tags would be more likely to be selected to be tested, while on the other hand, when classifying an image it would be more likely to annotate it with a more frequent tag. Moreover, we did not want to allow models to exploit attributes of collections which were unrelated to visual information, such as tag popularity. Hence, a sampling procedure was applied on all of the collections, which attempted to smooth these settings removing extreme conditions, namely classes which were either inadequately or very precisely described, while at the same time preserving the rest of the attributes of these collections.

The collections were first analysed, plotting the distribution of all N_{total} images over all of the C_{total} tags of each collection. In that way, it would be feasible to empirically determine on a reasonable number of images N_{min} with which each one of the classes should at least be described. This parameter N_{min} was set on a per collection basis. The second step was to remove any classes which were inadequately described, namely being described by a number of example images $N_C < N_{min}$. The result thus far would be having identified a part of the collection which contains only classes for which we have enough images (N_{min}) at our disposal. Next, we would randomly select a number of C_{sample} classes to form our sample. However, in order to also remove tags which were very frequent, we did not select the C_{sample} classes from the whole range of the remaining classes ($C_{remaining}$), but from the first $C_{sample-from}$ classes after sorting them based on the number of images belonging to them. As such, the result now would be having a sample of N_{sample} images from each collection which contained only medium-frequency classes. However, when selecting an image, we would consider all of the tags which belonged to the sampled C_{sample} classes regardless of the fact that this would make some tags appear as more popular than others. Discarding some of the instances of a class, might have a negative impact on the performance and the overall operation of an image classification algorithm, as it would be very difficult to define the class's class and non-class models.

In Figure 1, the reader is provided with the distributions of images over classes for the Corel 5K and the Caltech 101 collections. In the left column, the distribution of images over tags for the whole collection is plotted. In the middle, we have removed the inadequately described tags, which enables us to empirically determine, how many tags should be sampled (C_{sample}) and how many of the most popular tags, which appear on the right side of the graph should be discarded. Finally, at the right column, the distribution of images over classes for our sample of each collection is plotted. Moreover, in Table 1

	Corel 5K	TrecVid 2007	Caltech 101
N_{total}	4079	17675	8242
C_{total}	374	36	100
N_{min}	40	40	40
$C_{remaining}$	75	30	75
$C_{sample-from}$	60	30	70
C_{sample}	50	30	50
N_{sample}	1195	527	2009

Table 1. Statistics and parameters' values regarding the sampling procedure

the reader is provided with some statistics and the values of the afore-mentioned parameters for each collection.

However, in order to be fair with an algorithm and remove any chance of the results being based on luck, such an evaluation should be cross-validated. In our evaluation, we decided to evaluate the algorithms using $N = 10$ folds. However, it would be extremely difficult and even impossible to be able to split our samples of these collections into $N = 10$ totally separated, self-contained and statistically-balanced parts. Certain parts might not include any train images for some classes, or certain tags might not be tested in some folds. Therefore, we modified the afore-mentioned sampling procedure, executing it N times for each collection sampling each time $N_{fold} = N_{min}/N$ images for each one of the predetermined C_{sample} set of classes. In that way, the result of this process would be having N consistent, separated and self-contained samples from each collection.

3.3 Image Segmentation

Since fixed-size block-segmentation is an essential part of one of the models which were chosen to be evaluated, namely the MBRM model, block segmentation was also used while extracting local features. However, dividing images into equally-sized regions might be misleading, since even small-size objects may be split into two or more regions while large-sized ones are always seen in part and never in whole. Hence, the optimal size of blocks is dependent on the images and the collection itself. In order to overcome the two afore-mentioned obstacles of fixed-size block segmentation, we used variable-size block segmentation with varying degrees of overlapping. First, all images were resized to fit in a 512×512 window. Then, we empirically identified a set of block sizes $S = \{32, 64, 128, 256\}$ which would be meaningful when used in block segmentation given the average size of the images and the average size of the objects depicted in them. Depending on the size s_i of the square blocks, we would determine on the degree of overlapping. The step d between two neighbouring blocks was set to $d = s_i$ when $s_i \leq 32$, $d = s_i/3$ when $32 < s_i \leq 64$, and finally $d = s_i/4$ when $s_i > 64$. By

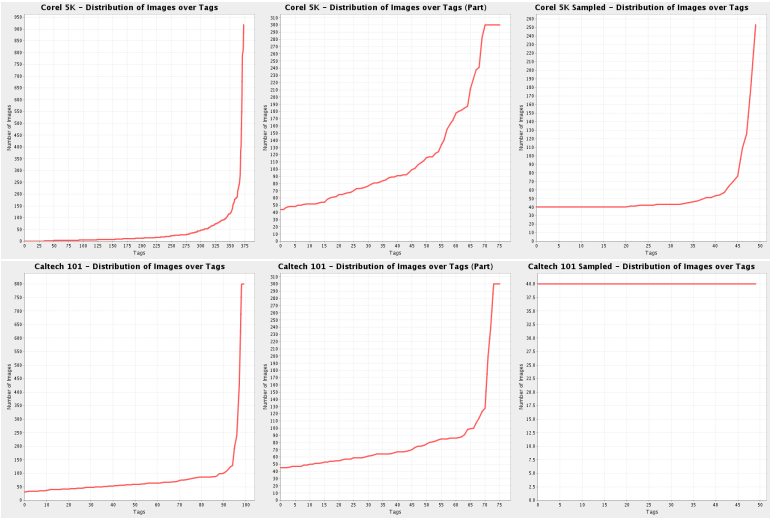

Figure 1. Distributions of images over classes in the Corel 5K (first row) and Caltech 101 (second row) collections.

considering overlapping multi-resolution block-segmentation, we ensure that objects and classes will be seen both in part and as a whole during the annotation process, which is a very desired property in object class recognition.

3.4 Content Descriptors

Image representation and feature extraction is an important and definitive step when attempting to use an automatic image annotation algorithm. It is important to identify the appropriate set of features, one which would provide not only the appropriate level of discrimination among images, but also enough compactness, so that the algorithm itself will not suffer from the challenging problems of computational complexity, immense resource requirements and the curse of dimensionality. In addition, it is not unusual for a multimedia collection to be known to yield better results when used in combination with a specific set of features, while on the other hand, certain image classification algorithms also perform better when used with certain sets of features. Hence an evaluation of image classification algorithms incorporating various features sets representing different attributes and characteristics of the same images from the same collections might shed some light into the operation of these algorithms through their variation in performance when applied on various such settings of collections and features sets.

When deciding on the features sets which would be incorporated in the evaluation process, the objective was to use standardised features sets, no matter how well they would actually perform. The goal of the present work was not to get better results, but to investigate patterns in the relative performance and the presence of any consistency between certain image classification algorithms.

As such, by using colour and texture features defined in the MPEG-7 Standard [16], such as Colour Histogram (CH), Edge Histogram (EH), and Homogeneous Texture (HT), as well as SIFT features introduced in 2004 by Lowe [17], it would be clear that we did not act in favour of a specific algorithm, while the results of this work would still be meaningful in the future, as it would be straightforward to implement a new algorithm, run experiments on the same collections using these standardised features sets and get comparable results.

4 Results

In this section, results showing mean per-word precision and recall for each setting individually are presented.

In Table 2, results of experiments with our implementation of MBRM and SML using MPEG-7 and SIFT Features respectively are presented for the three collections. Our results are significantly lower than the ones reported in the original papers [4,5]. The reason for this is that we used normalised parts of the collections, as well as other sets of features. On the other hand, in Table 3, the MBRM is contrasted to the simpler Support Vector Machines (SVM) approach using the SVM-light implementation [18].

First of all, with respect to the collections, we would say that Corel was the most "extreme" setting, followed by that of TrecVid 2007, and then the completely normalised sample of Caltech 101. By "extreme", we mean that only a few tags were more popular than others, while these had significantly more example images. Moreover, we would assume that, as TrecVid 2007 is supposed to be used for high level video tasks, it would be extremely difficult to detect similarity between frames using common image descriptors.

From Table 2, we can see that the variance of both Precision and Recall around the means was significantly high. We also see that only a small percentage of tags has $Recall > 0$ and most of these tags are popular tags in the collection. This is similar to previously reported results [11,4,5] on the Corel 5K collection. However, since we have removed most of the popular tags the numbers tend to be significantly smaller. This shows that previous optimistic results on Corel 5k are actually due to the tag distribution rather than the descriptive ability of the models. Interestingly, MBRM would always return the most popular words when evaluated on Corel 5K and TrecVid 2007. On the contrary, in Caltech 101, in which tag frequencies were completely normalised, more words were returned and the diversity among them was high. Also, regarding the TrecVid dataset, we see that MBRM had exactly the same response across all descriptors, meaning that similarity across images was not taken into account by the model. On the other hand, SML achieved the best performance on TrecVid 2007, followed by Corel 5K and Caltech 101. The bad performance on Caltech might be due to the fact that it is a single-label environment, and the actual number of classes depicted in an image was considered during the annotation process. The difference in performance between Corel 5K and TrecVid 2007 might be either due to the visual content of the images, or due to collection-specific properties. Nevertheless,

Collections	Corel 5K				TrecVid 2007				Caltech 101			
Models	MBRM			SML	MBRM			SML	MBRM			SML
Descriptors	CH	EH	HT	SIFT	CH	EH	HT	SIFT	CH	EH	HT	SIFT
# of words in total	70				30				50			
# of words with Recall>0	4	4	4	6	8	8	8	9	18	14	13	2
Precision and Recall on all words												
Mean Per-word Recall	0.034	0.045	0.045	0.046	0.194	0.194	0.194	0.130	0.125	0.265	0.270	0.015
Variance in Recall	0.151	0.193	0.193	0.175	0.356	0.356	0.356	0.269	0.207	0.275	0.286	0.077
Mean Per-word Precision	0.020	0.010	0.010	0.003	0.163	0.163	0.163	0.073	0.127	0.286	0.251	0.0009
Variance in Precision	0.121	0.044	0.044	0.011	0.296	0.296	0.296	0.160	0.214	0.316	0.268	0.005
Precision and Recall on words with Recall > 0												
Mean Per-word Recall	0.569	0.750	0.750	0.495	0.534	0.534	0.534	0.397	0.222	0.377	0.422	0.360
Variance in Recall	0.284	0.238	0.238	0.303	0.295	0.295	0.295	0.320	0.206	0.182	0.174	0.125
Mean Per-word Precision	0.334	0.172	0.174	0.036	0.449	0.449	0.449	0.188	0.227	0.360	0.284	0.021
Variance in Precision	0.374	0.054	0.054	0.012	0.232	0.232	0.232	0.235	0.218	0.244	0.217	0.015

Table 2. Mean Precision and Recall of MBRM (MPEG-7) and SML (SIFT).

overall in all collections, our results are not as optimistic as previously reported ones, and this seems to be related to the normalised tag distributions of our samples. However, although different categories of features were used with each model, the results between them are still comparable and can be interpreted in a generic way.

Moreover, we applied a Support Vector Machine using global MPEG-7 features on the Corel 5K collection and compared it with MBRM and SML. Results are presented in table 3, where we can see that a simple SVM with global features achieves better results than MBRM and SML, which are considered state-of-the-art methods. We have also implemented a SVM with local MPEG-7 features by using k-means to cluster local features and create visual terms. The local features are associated to their closest visual term (cluster centroid) and images are represented by the frequency of the visual terms they contain, similarly to a bag of word model used in Information Retrieval. Despite the quantisation errors introduced by the k-means algorithm, results are still better than MBRM and SML although not as good as using the SVM directly on the global MPEG-7 features.

Finally, with respect to SML, it was not feasible to combine it with local MPEG-7 Features. The image segmentation procedure which was used for extracting MPEG-7 local features led to a quite homogeneous representation of each image individually. The MBRM was not affected by this homogeneity since features were homogeneous only at the image level. SML however was not able to cluster the feature vectors representing each image with a mixture model of a reasonable number of components. As SML uses a mixture of Gaussians, it essentially makes strong assumptions about the nature and the properties of the features, thus making it feature-dependent. Hence, the SML would require a significantly larger dataset, and a descriptor which would provide an appropriate degree of heterogeneity at the image level.

Collections	Corel 5K						Caltech 101				TrecVid 2007			
Models	MBRM	SVM		MBRM	SVM		MBRM	SVM	MBRM	SVM	MBRM	SVM	MBRM	SVM
Descriptors	CH	GCH	CH	EH	GEH	EH	CH	GCH	EH	GEH	CH	GCH	EH	GEH
# of words in total	70						50				30			
# words (Recall>0)	4	32	26	4	37	29	18	20	14	23	8	15	8	16
Precision and Recall on all words														
Mean Recall	0.034	0.204	0.102	0.045	0.402	0.314	0.125	0.327	0.265	0.580	0.194	0.405	0.194	0.611
Recall Variance	0.151	0.236	0.159	0.193	0.430	0.250	0.207	0.221	0.275	0.325	0.356	0.265	0.356	0.374
Mean Precision	0.020	0.131	0.051	0.010	0.242	0.193	0.127	0.372	0.286	0.740	0.163	0.454	0.163	0.564
Precision Variance	0.121	0.226	0.087	0.044	0.301	0.183	0.214	0.158	0.316	0.363	0.296	0.347	0.296	0.336
Precision and Recall on words with Recall > 0														
Mean Recall	0.569	0.149	0.173	0.750	0.188	0.245	0.222	0.173	0.377	0.300	0.534	0.227	0.534	0.265
Recall Variance	0.284	0.095	0.144	0.238	0.195	0.169	0.206	0.080	0.182	0.124	0.295	0.140	0.295	0.144
Mean Precision	0.334	0.173	0.087	0.172	0.193	0.139	0.227	0.142	0.360	0.057	0.449	0.009	0.449	0.028
Precision Variance	0.374	0.269	0.092	0.054	0.294	0.182	0.218	0.234	0.244	0.288	0.232	0.258	0.232	0.288

Table 3. Comparison between MBRM and SVM using MPEG-7 Descriptors.

5 Conclusion

In this paper, we considered the lack of proper evaluation in the domain of Automatic Image Annotation. We found that the evaluation methodologies followed by AIA researchers are insufficient and do not support and prove the models' initial assumptions. Hence, we defined an Evaluation Framework, which is comprised by more than one multimedia collections and standardised descriptors, uses a sampling method to extract smoother, self-contained and representative samples and a multi-resolution block-segmentation method. We used this framework to evaluate and compare two state-of-the-art AIA models and we found that they heavily depend on the underlying test set. MBRM was found to return the most popular tags, while the SML was found to be extremely feature-dependent, and could not be integrated with standardised MPEG-7 Features. Thus, the high reported performance measures could be artifacts of the collections and not due to the descriptive power of the models. Finally, we have demonstrated that a simple SVM approach performs better than state-of-the-art models across several collections and descriptors.

We argue that as the number of experimental settings increases and as we keep their diversity high, we get more insight on a model's functionality, while strong and weak points emerge. As such, this study sets forward an evaluation paradigm for future annotation models, while the proposed framework should be integrated in the whole process of the development of a model, from the conceptualisation and the development phases until the validation and evaluation.

6 Acknowledgements

The research leading to this paper was supported by European Commission under contract FP6-027122 (Salero).

References

1. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Trans. Pattern Anal. Mach. Intell. **22**(12) (2000) 1349–1380
2. Duygulu, P., Barnard, K., de Freitas, J.F.G., Forsyth, D.A.: Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: Proceedings of the 7th European Conference on Computer Vision, London, UK, Springer-Verlag (2002) 97–112
3. Lavrenko, V., Manmatha, R., Jeon, J.: A model for learning the semantics of pictures. In Thrun, S., Saul, L., Schölkopf, B., eds.: Advances in Neural Information Processing Systems 16. MIT Press, Cambridge, MA (2004)
4. Feng, S., Manmatha, R., Lavrenko, V.: Multiple bernoulli relevance models for image and video annotation. Volume 2. (June-2 July 2004) II–1002–II–1009 Vol.2
5. Gustavo Carneiro, Antoni B. Chan, P.J.M., Vasconcelos, N.: Supervised learning of semantic classes for image annotation and retrieval. IEEE Trans. Pattern Anal. Mach. Intell. **29**(3) (2007) 394–410
6. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3** (2003) 993–1022
7. Blei, D.M., Jordan, M.I.: Modeling annotated data. In: Proceedings of the 26th annual international ACM SIGIR, New York, NY, USA, ACM (2003) 127–134
8. Boiman, O., Shechtman, E., Irani, M.: In defense of nearest-neighbor based image classification. In: IEEE CVPR 2008. (June 2008) 1–8
9. Vasconcelos, N., Lippman, A.: Learning mixture hierarchies. In: Advances in Neural Information Processing Systems II, Cambridge, MA, USA, MIT Press (1999) 606–612
10. Li, J., Wang, J.Z.: Automatic linguistic indexing of pictures by a statistical modeling approach. IEEE Trans. Pattern Anal. Mach. Intell. **25**(9) (2003) 1075–1088
11. Jeon, J., Lavrenko, V., Manmatha, R.: Automatic image annotation and retrieval using cross-media relevance models. In: Proceedings of the 26th annual international ACM SIGIR Conference, New York, NY, USA, ACM (2003) 119–126
12. Kwasnicka, H., Paradowski, M.: On evaluation of image auto-annotation methods. In: Proceedings of the 6th Int. Conf. on Intelligent Systems Design and Applications, Washington, DC, USA, IEEE Computer Society (2006) 353–358
13. Westerveld, T., de Vries, A.P.: Experimental evaluation of a generative probabilistic image retrieval model on 'easy' data. In: In Proceedings of the SIGIR Multimedia Information Retrieval Workshop 2003. (Aug 2003)
14. Ayache, S., Quénot, G.: Trecvid 2007 collaborative annotation using active learning. TRECVID'2007 Workshop (November 2007)
15. Fei-Fei, L., Fergus, R., Perona, P.: One-shot learning of object categories. IEEE Trans. Pattern Anal. Mach. Intell. **28**(4) (April 2006) 594–611
16. Manjunath, B.S., Salembier, P., Sikora, T.: Introduction To Mpeg-7: Multimedia Content Description Interface. John Wiley & Sons (2002)
17. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision **60** (2004) 91–110
18. Joachims, T.: Making large-scale support vector machine learning practical. (1999) 169–184