Columbia University Computer Science Department Technical Report # CUCS 005-12 (2012)

Experiments of Image Retrieval Using Weak Attributes

Felix X. Yu[†], Rongrong Ji[†], Ming-Hen Tsai[§], Guangnan Ye[†], Shih-Fu Chang[§]
Columbia University, New York, NY 10027
[†]{yuxinnan, rrji, yegn}@ee.columbia.edu [§]{minghen, sfchang}@cs.columbia.edu

Abstract

Searching images based on descriptions of image attributes is an intuitive process that can be easily understood by humans and recently made feasible by a few promising works in both the computer vision and multimedia communities [4,7,9,11]. In this report, we describe some experiments of image retrieval methods that utilize weak attributes [11].

1 Introduction

Searching images based on descriptions of image attributes is an intuitive process that can be easily understood by humans and recently made feasible by a few promising works in both the computer vision and multimedia communities [4,7,9,11]. In this report, we describe some experiments of image retrieval methods that utilize weak attributes [11]. This technical report includes experimental results and discussions in addition to those in [11]. The experiments are done on the Labeled Faces in the Wild (LFW) [3], a-PASCAL/a-Yahoo [1], and a-TRECVID [11] datasets. For LFW, we have selected a small number of weak attributes, to visualize the learnt dependency model. For a-PASCAL/a-Yahoo and a-TRECVID, we evaluate the contribution of different types of weak attributes. In this report, our focus is evaluation of attribute based retrieval, instead of recognition or verification.

2 Labeled Faces in the Wild (LFW)

We implemented and tested the weak-attribute based image retrieval method described in [11]. The first evaluation is on the Labeled Faces in the Wild (LFW) dataset [3], which contains 9,992 images with manual annotations of 27 query attributes, including "Asian", "Beard", "Bald", "Gray Hair", *etc.* Following the setting of [9], we

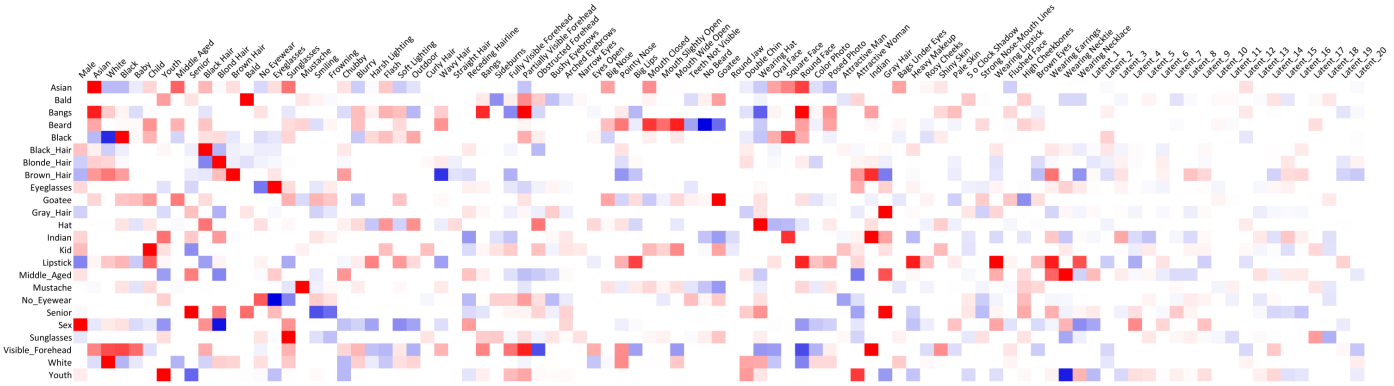

Figure 1: Learnt $\mathbf{w}$ for LFW dataset with sparsity $k = 40$ (best viewed in color). Vertical labels are query attributes, and horizontal labels are weak attributes. Red, blue and white represent positive, negative and no dependency respectively. High intensity means high dependency level. The sparse pattern of weak attributes results in faster training/testing and less overfitting. This learned matrix is semantically plausible for most dependency patterns. For instance, people "Wearing Lipstick" (query) is unlikely to be "Male" (weak), and "Kid" (query) is highly related to "Child" (weak). Note that not all mappings are semantically meaningful, due to the fact that the weak attributes, though with some specific names, can be "weak", *i.e.* inaccurate, in terms of semantic meanings.

randomly choose 50% of this dataset for training and the rest for testing. For visualization, the weak attributes for this dataset contain only attribute classifier scores from [5] (scores of 73 attribute classifiers designed for human faces) and latent variables produced by the graphical model[1].

In order to get the baseline results of individual classifiers (direct classifiers corresponding to the query attributes), we have omitted three attributes: "Long Hair", "Short Hair" and "No Beard" which are not covered by the classifier scores from [5]. Figure 1 shows the learnt dependency model $\mathbf{w}$ using our proposed method with sparsity $k = 40$. For visualization, we only show the weak attributes selected by single-attribute queries, *i.e.* each query attribute is uniquely mapped to 40 weak attributes only. Note the sparse pattern considered in our model is query specific, rather than a fixed one across different queries. For example, for a single query "Asian", the selected weak attributes might be "Asian" and "White", while for a double attribute query "Asian"+"Woman", the selected weak attributes might be "Asian" and "Male". In both learning and prediction processes involving the dependency model, only the selected weak attributes will be considered for each multi-attribute query.

Figure 2 shows the comparisons of our method to several existing approaches, including TagProp [2], Reverse Multi-Label Learning (RMLL) [8], Multi-Attribute based

[1]The value of latent variables are acquired by inferencing on the unsupervised graph, conditioned on the weak attributes. We have transformed the conditional marginal distribution of latent variables back to the real interval $[-1, 1]$ by linear mapping.

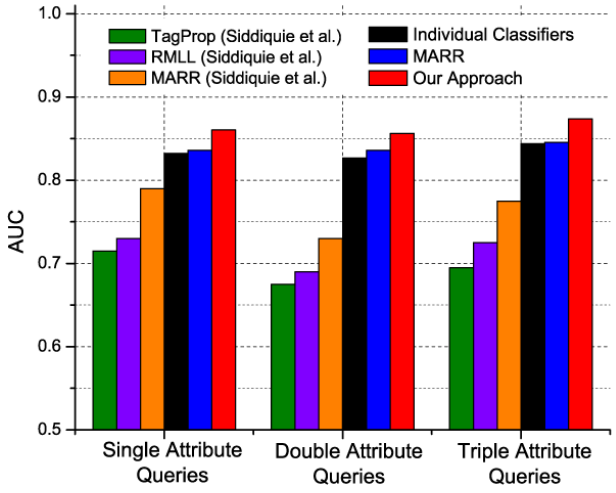

Figure 2: Retrieval performance on LFW dataset. The first three results are copied from [9], in which different individual classifiers are used. The last three results are based on our implementation.

Ranking and Retrieval (MARR) [9], individual classifier scores from [5], and our implementation of MARR based on individual classifier scores. Our method outperforms all the other competing methods consistently for all types of queries including single, double and triple attributes. It is interesting to note that in this experiment, the weak attributes are actually not "weak", in the sense that even individual classifiers outperform TagProp, RMLL and MARR reported in [9], in which different individual classifiers are used. The weak attributes are *weak* in the sense that they are trained from other sources, therefore are not directly related to the specific dataset at hand. Closely following [9], our implementation of MARR only slightly outperforms individual classifiers: the stronger baseline reduces the amount of improvement that can be obtained by utilizing dependency information within query attributes only.

3 a-PASCAL and a-Yahoo[2]

The dataset a-PASCAL [1] contains 12,695 images (6,340 for training and 6,355 for testing). Each image is assigned one of the 20 object class labels: people, bird, cat, *etc.* Each image also has 64 query attribute labels, such as "Round", "Head", "Torso", "Label", "Feather" *etc.* The a-Yahoo dataset is collected for 12 object categories from the Yahoo images search engine. Each image in a-Yahoo is described by the same set

[2]The first few paragraphs summarizing the datasets and weak attributes are copied from [11]. In the end of the section, we add analysis and discussions about the contributions by different types of weak attributes.

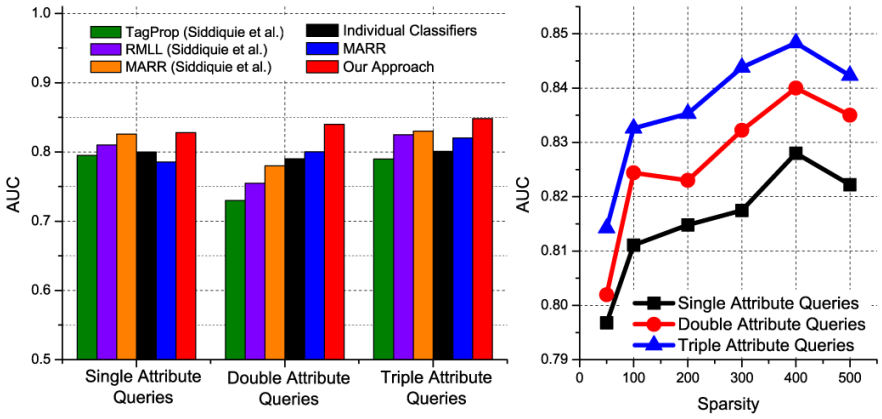

Figure 3: Retrieval performance on a-PASCAL dataset. This figure is copied from [11]. Left: AUC comparison based on optimal sparsity ($k = 400$). The first three results are copied from [9], under the same configurations compared to ours. Right: AUC of our approach with varying sparsity.

of 64 attributes, but with different category labels compared to a-PASCAL, including wolf, zebra, goat, donkey, monkey *etc.*

Following the setting of [1], we use the pre-defined training images of a-PASCAL as training set, and test on pre-defined test images of a-PASCAL and a-Yahoo respectively. We use the feature provided in [1]: 9,751-dimensional features of color, texture, visual words, and edges to train individual classifiers. Following [11], other weak attributes include:

- Scores from Classemes semantic classifiers [10]: 2,659 classifiers trained on images returned by search engines of corresponding query words/phrases;
- Discriminative attributes [1], which are trained using linear SVM by randomly selecting 1-3 categories as positive, and 1-3 categories as negative;
- Random image distances: the distance of each image to some randomly selected images based on the 9,751-dimensional feature vector;
- Latent variables, as detailed in Section 2.

This finally results in 5,000 weak attributes for each image.

Figure 3 shows performance evaluation results using the a-PASCAL benchmark, in comparison with the state-of-the-art approaches in [2,8,9]. Figure 4 shows the performance of weak attributes on the a-Yahoo benchmark compared to individual classifiers and MARR. From both of them, the weak attribute approach has demonstrated substantial performance gain compared to other methods. Detailed analysis of the two figures can be found in [11].

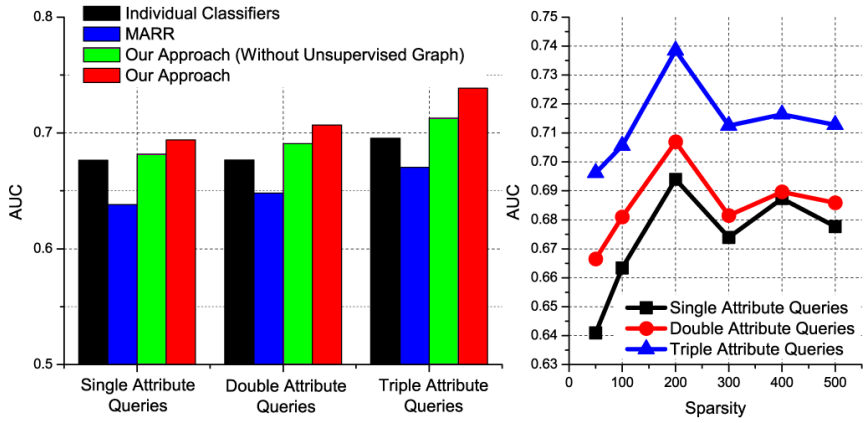

Figure 4: Retrieval performance on a-Yahoo dataset. This figure is copied from [11]. Left: AUC comparison based on optimal sparsity ($k = 200$). Right: AUC of our approach with varying sparsity.

To evaluate the contribution of heterogeneous types of weak attributes in the dependency model, we conduct experiments to compute the percentage of weights in the dependency model coming from different weak attribute types. Table 1 shows the results. For a-PASCAL, about 90% of the weights are from the weak attributes (excluding individual query attribute classifier scores). For a-Yahoo (cross-dataset retrieval, the training and test set are under different statistics), this value is even higher. It means that for both datasets, weak attributes play a key role. And for a-Yahoo, the individual query attribute classifier scores have a much worse generalization power (3.23 %) compared to a-PASCAL (10.49%), due to cross-dataset issues. We now discuss the contributions from different types of weak attributes.

Classemes. It is encouraging to notice that, for both a-PASCAL and a-Yahoo, Classemes [10] contribute considerably in the dependency model. Classemes represent a class of weak but semantically meaningful attributes trained on some *other* data sources. Given extensive emerging research on image attributes and classification, this kind of classifiers are becoming widely available, and can be readily utilized in the weak attribute framework to improve the performance. What is more encouraging is that Classemes are trained automatically from results of image search engines. Therefore, even in the original weak attribute training process, it requires no human labeling burdens.

Discriminative attributes. The contribution of discriminative attributes is much more substantial for a-PASCAL compared to a-Yahoo. The reason is that categories of a-Yahoo are very different from those of a-PASCAL, from which the discriminative attributes are trained. Thus this kind of attributes have far less description power on a-Yahoo. Generally speaking, the more relevant the weak attributes to the query

	Query Att Classifier	"Classemes"	Disriminative Att	Others
a-PASCAL	10.49%(64)	25.87%(2659)	48.73%(1350)	14.91%(927)
a-Yahoo	3.23%(64)	74.30%(2659)	17.23%(1350)	5.24%(927)
a-TRECVID	20.09%(126)	70.83%(2659)	3.84%(2000)	5.24%(1215)

Table 1: Contributions of different types of weak attributes. The table shows percentage of weights in the dependency model coming from different weak attribute types. The number in brackets represents the number of weak attributes in the corresponding type. In our case, "others" include random image distances and latent variables.

attributes, the more weights will be associated with them.

Others (latent variables and random distances in our case). For both a-PASCAL and a-Yahoo, other weak attributes do not contribute much in the dependency model. However, based on our experiments, discarding them does have a negative impact on the final performance. So likely they do contribute to some queries. The advantage of the proposed weak attribute based retrieval framework is that it can incorporate *all* kinds of weak attributes, even random distances, while automatically selecting a subset that are potentially useful.

We have also conducted experiments to visualize the dependency model similar to Figure 1. Ideally, the weak attributes in helping to answer a specific query should be semantically related to the query (the relation can be either positive or negative). For example, a query describing a person is likely to be mapped positively to weak attributes including classifiers scores related to person, template instances related to person images, or latent variables representing person related weak attribute groups *etc.* Also it should be negatively mapped to some weak attributes describing scene, object, texture *etc.* without person. Similar to Figure 1, the visualization has demonstrated some relations with semantically meaningful explanations. However, many elements in the learnt dependency model still cannot be explained. The reason is that a large number of weak attributes are weak recognizing semantic meanings. This is not surprising, because learning classifiers for semantically meaningful attributes remains difficult. Fortunately, the proposed weak attribute approach does not require weak attributes to be semantically plausible.

4 a-TRECVID

The last evaluation is done on a-TRECVID [11]. It contains 126 uniquely labeled query attributes, and 6,000 weak attributes, of 0.26 million images (video frames). The individual attribute classifiers are trained using bag-of-words SIFT features under the spatial pyramid configuration [6]. Following the setting of Section 3, weak attributes include individual classifier scores, Classemes, discriminative attributes, distance to

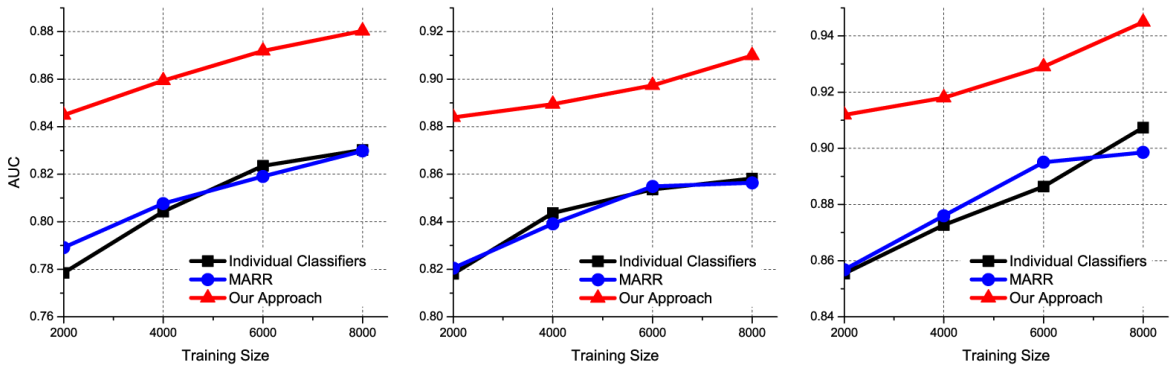

Figure 5: Retrieval performance on a-TRECVID dataset, with the varying training size. This figure is copied from [11]. From left to right: performance of single, double and triple attribute queries.

randomly selected images, and latent variables. Different from a-PASCAL dataset, we treat images from the same video as belonging to the same category. Figure 5 shows the performance of our approach comparing to individual classifiers and MARR.

From Table 1, we have found that about 80% of the weights are from the weak attributes (excluding individual query attribute classifier scores). This clearly demonstrates the contribution of weak attributes. The analysis of different types of weak attributes follows Section 3. It is interesting to notice that discriminative attributes do not contribute much for a-TRECVID. The possible reason is that weak attributes trained by the specific way are not directly related to the query attributes.

5 Conclusion

We evaluate performance of weak attributes based retrieval on LFW, a-PASCAL/a-Yahoo and a-TRECVID datasets. The learnt dependency model is visualized on LFW. We also analyze the contributions of heterogeneous types of weak attributes. The experiments verify the unique advantages of using weak attributes in the image retrieval framework [11]. It is encouraging to find that weak but semantically meaningful classifiers, *e.g.* Classemes [10], make strong contributions in our experiments.

Acknowledgement We would like to thank Dr. Michele Merler for sharing the LFW labels, Dr. Behjat Siddiquie and Dr. Neeraj Kumar for helpful discussions.

References

[1] H. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009. 1, 3, 4

[2] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Discriminative metric learning in nearest neighbor models for image auto-annotation. In *ICCV*, 2009. 2, 4

[3] G. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. *University of Massachusetts, Amherst, Technical Report 07-49*, 2007. 1

[4] N. Kumar, A. Berg, P. Belhumeur, and S. Nayar. A search engine for large collections of images with faces. In *ECCV*, 2008. 1

[5] N. Kumar, A. Berg, P. Belhumeur, and S. Nayar. Attribute and simile classifiers for face verification. In *CVPR*, 2009. 2, 3

[6] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006. 6

[7] A. Natsev, A. Haubold, J. Tešić, L. Xie, and R. Yan. Semantic concept-based query expansion and re-ranking for multimedia retrieval. In *ACM Multimedia*, 2007. 1

[8] J. Petterson and T. Caetano. Reverse multi-label learning. In *NIPS*, 2010. 2, 4

[9] B. Siddiquie, R. Feris, and L. Davis. Image ranking and retrieval based on multi-attribute queries. In *CVPR*, 2011. 1, 3, 4

[10] L. Torresani, M. Szummer, and A. Fitzgibbon. Efficient object category recognition using classemes. In *ECCV*, 2010. 4, 5, 7

[11] F. Yu, R. Ji, M.-H. Tsai, G. Ye, and S.-F. Chang. Weak attributes for large-scale image retrieval. In *CVPR*, 2012. 1, 3, 4, 5, 6, 7