Weak Attributes for Large-Scale Image Retrieval*

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

Abstract

Attribute-based query offers an intuitive way of image retrieval, in which users can describe the intended search targets with understandable attributes. In this paper, we develop a general and powerful framework to solve this problem by leveraging a large pool of weak attributes comprised of automatic classifier scores or other mid-level representations that can be easily acquired with little or no human labor. We extend the existing retrieval model of modeling dependency within query attributes to modeling dependency of query attributes on a large pool of weak attributes, which is more expressive and scalable. To efficiently learn such a large dependency model without overfitting, we further propose a semi-supervised graphical model to map each multi-attribute query to a subset of weak attributes. Through extensive experiments over several attribute benchmarks, we demonstrate consistent and significant performance improvements over the state-of-the-art techniques. In addition, we compile the largest multi-attribute image retrieval dateset to date, including 126 fully labeled query attributes and 6,000 weak attributes of 0.26 million images.

1. Introduction

The idea of "attribute" has been advocated in the computer vision community recently, where its effectiveness was demonstrated in various applications such as object recognition [5,13,14,27], and image/video search [12,23, 26]. In this paper, we focus on attribute based image retrieval. Specifically, we tackle the problem of large-scale image retrieval with multi-attribute queries. In such a scenario, the user provides multiple query attributes, to describe the facets of the target images. For instance, to retrieve images of a bird, one could describe the physical traits of feather, beak, and body *etc*. The task is to retrieve images containing *all* of the query attributes. We assume only a small portion of the database have the query attributes labeled before hand, and yet our goal is to search the entire large-scale image corpus.

A straightforward solution for the above problem is to build classifiers for the query attributes of interest, and sum the independent classifier scores to answer such multi-attribute queries [12]. A promising alternative, as shown in [23], is to analyze the dependencies among query attributes and leverage such multi-attribute interdependence to mitigate the noises expected from the imperfect automatic classifiers and thereby achieve robust query performance. An illustrative example of the above dependency model is shown in Figure 1.

However, [23] relied only on the pre-labeled query attributes to design the dependency model, limiting its performance and scalability. On one hand, user labeling is a burdensome process. On the other hand, the number of such pre-labeled attributes is limited: only a small set of words were chosen, for instance "Car", "Tree", "Road" *etc.* for street scenes, and "Bed", "Chair", "Table" *etc.* for indoor scenes. In particular, there are only 64 attributes in a-PASCAL benchmark [5] and similarly small number of attributes considered in other attribute datasets. Such a small amount of attributes are far from sufficient in forming an expressive feature space, especially for searching a large image corpus of diverse content.

In this paper, a **Weak Attribute** based paradigm is proposed to address the above challenges. It provides a principled solution for large-scale image retrieval using multi-attribute queries.

Weak Attributes *are a collection of mid-level representations, which could be comprised of automatic classifier scores, distances to certain template instances, or even quantization to certain patterns derived through unsupervised learning, all of which can be easily acquired with very little or no human labor.*

Different from query attributes, which are acquired by human labeling process, all kinds of weak attributes are

*Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20070. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

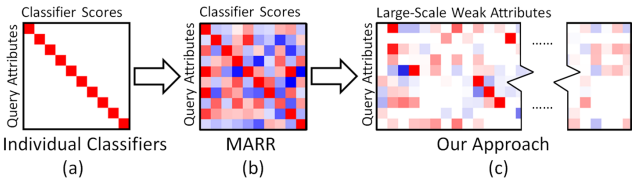

Figure 1. Different approaches to multi-attribute queries: (a) Direct independent matching of query attributes with corresponding classifiers. (b) Modeling dependency/correlation among a small set of query attributes (MARR) [23]. (c) Modeling dependency of query attributes on a much larger pool of weak attributes. To avoid overfitting and reduce complexity, we also impose sparsity in the dependency model.

generated automatically by machine[1]. We specifically refer to such attributes as "weak" because they may or may not be directly related to the query attributes. For example, hundreds or thousands of visual classifiers such as Classemes [24], Columbia374 [28], and automatic attributes [2], have been developed and made available, though they typically do not have direct correspondence with the target query attributes. Different from query attributes, weak attributes may not have clear semantic meanings. Examples are discriminative attributes [5] (A is more like a dog than a cat); relative attributes [20] (A is more natural than B); comparative values of visual features [13] (A is similar to B, or car in A has a shape similar to Volvo S60); and even values generated from latent topic models.

We are interested in the fact that the dimensionality of weak attributes (say thousands or more) is much higher than that of the query attributes. The high dimensionality ensures the weak attribute space to be sufficiently expressive, based on which a robust retrieval model can be developed. As shown in Figure 1, to bridge the gap between the limited query attribute (for user) and the expressive weak attribute space (for machine), we extend the attribute dependency model from the narrow one *among query attributes only* to a much more general one *mapping query attributes to weak attributes*, as detailed in Section 3.1.

Learning a large-scale dependency model based on a small amount of training labels is not trivial, due to the complexity of the learning process and the risk of overfitting. To address the above issues, we propose to impose sparsity in the dependency model, so that for both training and prediction, only a limited number of weak attributes are selected when answering each multi-attribute query (Section 3.2). To achieve this goal, we further develop a novel semi-supervised graphical model to incorporate statistics from both the labeled training data and the large amount of unlabeled data (Section 3.3). We will demonstrate through extensive experiments that our approach improves significantly over existing methods (Section 4), and can largely boost the flexibility of the retrieval model when dealing with cross-dataset variants (Section 4.1) and large-scale scenarios (Section 4.2). Our work has four unique contributions:

- We propose *weak attributes* that unify various kinds of mid-level image representations which can be easily acquired with no or little human labor.
- We apply weak attributes to image retrieval, by modeling dependency of *query attributes on weak attributes* under the framework of structural learning.
- To achieve efficiency and avoid overfitting, we propose a novel *semi-supervised graphical model* to select a sparse subset of weak attributes for each query. This makes the proposed method applicable to large and general datasets.
- We compile the largest multi-attribute image retrieval dataset to date, named a-TRECVID, including 126 fully labeled query attributes and 6,000 weak attributes of 0.26 million images extracted from videos used in the 2011 TRECVID Semantic Indexing (SIN) track[2].

2. Related Work

Attributes. Attributes or so-called "concepts" based image representation is well motivated in [7,17], and exploited in recent literature. For instance, [5] proposed an attribute-centric approach to help in reporting unusual aspects of known objects, and naming unknown objects; [27] proposed to jointly learn visual attributes, object classes and visual saliency in a unified framework, with consideration of attribute interdependency. Besides object recognition, attribute based approaches are also shown to be effective for video/image search [12,23,26]. In addition, there are works aiming to discover attributes either automatically through the web [2] or semi-automatically with human in the loop [19], and to explore new types of attributes [20].

Structural Learning. [25] introduced structural SVM to develop classifier with structural output. This technique has been well advocated in document retrieval [16] and multi-label learning [22] *etc*, due to its capability of incorporating different kinds of losses, such as precision, recall, F_β score and NDCG, into the optimization objective function. [23] proposed an interesting solution for multi-attribute image retrieval by using a structural learning paradigm to model interdependency across query attributes.

Multi-Keyword Queries. Besides the utilization of attributes, there are works on image search based on multi-keywords queries [6,11]. The main limitation of these approaches is the requirement that every image in the database

[1]Scores of the attribute classifiers, which are trained based on data sources existent before hand, can also be treated as weak attributes, for the reason that they are easily acquired without additional human labors.

[2]Attribute is called "concept" in TRECVID SIN track.

should be tagged manually or automatically to determine the presence of the tag words that may be used for query. This can be mitigated by propagating tags to new images with methods such as PAMIR [8] or Tag-Prop [9]. However, the above works did not take into account the dependency between query terms. Our work is also related to "query expansion" for multi-keyword queries in multimedia retrieval literature [18]. Different from query expansion, our approach "expands" a query to a sparse subset of weak attributes from a large pool, and the final dependency model is jointly learned across all possible queries under the framework of structural learning.

3. Weak Attributes for Image Retrieval

In the weak attribute based retrieval paradigm, our system first determines a sparse subset of weak attributes for each multi-attribute query, through a novel semi-supervised graphical model (Section 3.3). The selection process is optimized under a formulation of maximizing mutual information between query attributes and weak attributes (Section 3.2). Then, for each multi-attribute query, only the selected weak attributes are considered in the subsequent attribute-based retrieval process (Section 3.1), ensuring efficiency and avoiding overfitting. In the following, we first start with the retrieval method using structural SVM that models the dependency of query attributes on weak attributes.

3.1. Retrieval Model

Our retrieval model is based on structural SVM. Similar to [23], it can be easily modified for the image ranking scenario, by some minor changes with an appropriate query relevance function.

Retrieval. Let $Q \subset \mathcal{Q}$ be a multi-attribute query, where $\mathcal{Q}$ is the complete set of all possible query attributes. Let $\mathcal{X}$ be the set of weak attributes, and $\mathcal{Y}$ denotes the set of images. The multi-attribute retrieval is to select a set of images $Y^* \subset \mathcal{Y}$ as the structured response to a given Q:

$$Y^* = \arg\max_{Y \subset \mathcal{Y}} \mathbf{w}^T \psi(Q, Y), \tag{1}$$

where[3]

$$\mathbf{w}^T \psi(Q, Y) = \sum_{q_i \in Q} \sum_{x_j \in \mathcal{X}} w_{ij} \sum_{y_k \in Y} \phi(x_j, y_k). \tag{2}$$

Here, $\phi(x_j, y_k)$ is the value of weak attribute x_j of image y_k. Compared to [23], our key insight is to model the dependency (characterized by $\mathbf{w}$) of query attributes on a large set of weak attributes, not just within the small query attributes set itself, as illustrated earlier in Figure 1. Equation 1 can be solved efficiently in $O(|\mathcal{Y}|)$.

[3]For easier presentation, $\mathbf{w}$ is written as matrix form here: $\mathbf{w} \in \mathbb{R}^{|\mathcal{Q}| \times |\mathcal{X}|}$, where w_{ij} is the dependency/correlation of the i-th query attributes to the j-th weak attribute.

Training. Given a set of labeled images $\mathcal{Y}_l$, whose ground truth query attributes are known and weak attribute scores are computed, the training step is to learn $\mathbf{w}$ to predict the subset of images ${Y_t}^* \subset \mathcal{Y}_l$ for a given multi-attribute query $Q_t \subset \mathcal{Q}$. We follow the standard max-margin training formulation as follows:

$$\begin{aligned} \arg\min_{\mathbf{w}, \xi} \quad & \mathbf{w}^T \mathbf{w} + C \sum_t \xi_t \\ \forall t \quad & \mathbf{w}^T (\psi(Q_t, {Y_t}^*) - \psi(Q_t, Y_t)) \geq \triangle({Y_t}^*, Y_t) - \xi_t, \end{aligned} \tag{3}$$

where Y_t is any subset of images different from ${Y_t}^*$. $\triangle({Y_t}^*, Y_t)$ is the loss function, which can be set as Hamming loss, precision, recall and F_β score *etc.* [22,23].

Equation 3 can be solved by the cutting plane method [25]. It solves Equation 3 initially without any constraints, and then during each iteration adds the most violated constraint of the current solution. The most violated constraint during each iteration is generated by:

$$\arg\max_{Y_t \subset \mathcal{Y}_l} \triangle({Y_t}^*, Y_t) + \mathbf{w}^T \psi(Q_t, Y_t), \tag{4}$$

which can be solved in $O(|\mathcal{Y}_l|)$ with the Hamming loss, and $O(|\mathcal{Y}_l|^2)$ with the loss of the F_β score [22].

3.2. Query Adaptive Selection of Weak Attributes

For a large weak attribute pool, the model ($\mathbf{w}$ in Equation 1) contains a prohibitively large number of $|Q| \times |\mathcal{X}|$ parameters to be learnt (500,000 if there are 100 query attributes and 5,000 weak attributes). This is computationally expensive and may cause overfitting, given the fact that images containing query attribute labels are usually difficult to get, and thus in much smaller amount compared to the test images.

We solve the above issues by imposing query adaptive selection of weak attributes on the dependency model. Given a query $Q \subset \mathcal{Q}$, the objective is to get a small set of weak attributes $X_Q \subset \mathcal{X}$ relevant to Q, so that for both training (Equation 3) and testing (Equation 1), only the corresponding elements of $\mathbf{w}$ are considered:

$$\mathbf{w}^T \psi(Q, Y) = \sum_{q_i \in Q} \sum_{x_j \in X_Q} w_{ij} \sum_{y_k \in Y} \phi(x_j, y_k). \tag{5}$$

This idea is important, since from both intuition and the experiment results which will be presented later, only a small subset of weak attributes in the large weak attribute pool are related to a specific multi-attribute query.

We formulate the above weak attribute selection problem as maximizing mutual information, which is a general measurement of relevance:

$$\max_{X_Q \subset \mathcal{X}} I(Q; X_Q) \quad s.t. \quad |X_Q| = k, \tag{6}$$

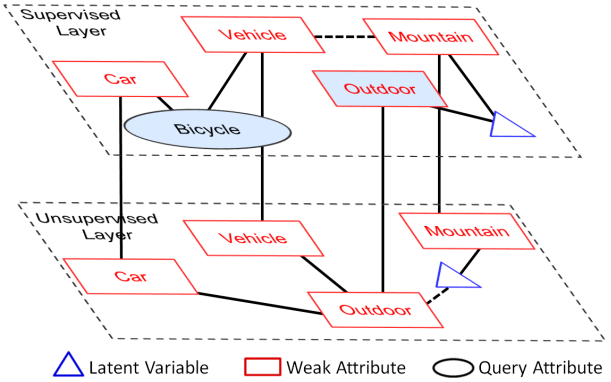

Figure 2. The proposed semi-supervised graphical model for discovering weak attributes that have the highest correlations with the query attributes in any specific query (the dotted line means connection through other nodes). The power of the two-layer model can be best shown by the relation between query attribute 'Bicycle" (shaded) and the weak attribute "Outdoor" (shaded) through links across layers. Such relation is unclear by considering the supervised layer alone.

where k, the sparsity, is the desired number of selected weak attributes. General feature selection methods based on entropy criterions can be found in [21].

Equation 6 is hard to solve, because there are $\binom{|\mathcal{X}|}{k}$ combinations in selecting $X_Q \subset \mathcal{X}$. We instead consider $x_i \in \mathcal{X}$ one at a time, and set X_Q as the top-k x_i with highest mutual information values:

$$\max_{x_i \in \mathcal{X}} I(Q; x_i), \tag{7}$$

where $I(Q; x_i) = H(Q) - H(Q|x_i)$. $H(Q)$ is a constant for a given query Q. $H(Q|x_i)$ can be expanded as:

$$H(Q|x_i) = -\sum_{x_i} P(x_i) \sum_{Q} P(Q|x_i) \log P(Q|x_i), \tag{8}$$

where $P(x_i)$ is the marginal distribution of weak attribute x_i in the test set. $P(Q|x_i)$ plays a key role in bridging each weak attribute $x_i \in \mathcal{X}$ to a specific multi-attribute query Q. It is easy to model $P(Q|x_i)$ based on training data that has query attribute ground truth. However, it may bias the model because the training set can be very small, and possibly under different statistics compared to the test set.

3.3. Semi-Supervised Graphical Model

To find weak attributes that have the highest relevance with the query attributes, we have designed a novel semi-supervised graphical model, which estimates $P(Q|x_i)$ efficiently with statistics from both training and test data.

The Model. The semi-supervised graphical model (Figure 2) is a two-layer probabilistic graphical model[4]. The first layer is the *supervised layer*, which is constructed based on the training data with query attribute labels. This layer consists of nodes representing both query and weak attributes (including latent variables discovered in the graph construction process). It is intended to model their joint distributions on the training data. The second layer is the *unsupervised layer*, which is constructed based on the target test data, with no query attribute labels available. This layer only consists of nodes representing weak attributes (including latent variables discovered in the graph construction process). It is to characterize the joint distribution of weak attributes on the target test data.

We choose latent tree [4] as our graphical model in each layer. Tree models fall within a class of tractable graphical models. They are efficient in inference and widely used in prior works of context modeling or object recognition *e.g.* [27]. The learnt latent variables can be treated as additional weak attributes, which in some sense summarize information of certain attribute node groups.

The two layers are connected through weak attributes that appear in both layers. Therefore the model can leverage information from both the labeled training data and the unlabeled test data, and thus we call it "semi-supervised". The graphical model can capture the high-order dependency structure of attributes. It greatly improves the generalization power of the proposed method in "cross-dataset" retrieval (Section 4.1), and retrieval with very small amount of training data (Section 4.2).

Alternating Inference. We now describe how to estimate $P(Q|x_i)$ based on the proposed semi-supervised graphical model. Direct inference is difficult, because when considering the two layers together, the graphical model may not be acyclic and thus untractable. To address this issue, we have developed a method called Alternating Inference, as summarized in Algorithm 1. The idea is to do inference iteratively on the unsupervised and supervised layers

Algorithm 1 Alternating Inference

Given a query $Q \subset \mathcal{Q}$, compute $P(Q|x_i = 1)$ (without loss of generality), $\forall x_i \in \mathcal{X}$, as needed in Equation 8.
while Not convergent (the change of $P_{sup}(x)$ is large) **do**
  Inference on the unsupervised graph to get its marginal distribution $P_{unsup}(x|x_i = 1), \forall x \in \mathcal{X}$.
  Update the margin of $x \in \mathcal{X}$ of the supervised graph $P_{sup}(x) \leftarrow P_{unsup}(x|x_i = 1)$.
  Inference on the supervised graph, to get updated $P_{sup}(x), \forall x \in \mathcal{X}$.
  Update the margin of $x \in \mathcal{X}$ of the unsupervised graph: $P_{unsup}(x) \leftarrow P_{sup}(x)$.
end while
Compute joint distribution $P_{sup}(Q)$ in the supervised graph as output $P(Q|x_i = 1)$.

[4]Before learning this model, weak attributes are converted to binary variables in order to use the same discrete format as the query attributes.

in an alternate fashion. In each iteration, marginal probabilities of the weak attributes are estimated in one layer and passed to the other layer for inference. Then the process is reversed for refining the margins in the previous layer. The inference over each layer can be done efficiently by belief propagation [3]:

Let $\psi_{ij}(x_i, x_j)$ be the potential function obtained by the latent tree model, and $N(i)$ be the set of nodes connected to node i. And let $\mu_{j\to i}(x_i)$ be the message passing from node j to node i. Then the margin $P(x_i) = \phi_i(x_i)\prod_{j\in N(i)}\mu_{j\to i}(x_i)$ can be computed efficiently using belief propagation by just traversing the tree twice, where $\mu_{j\to i}(x_i)$ is computed recursively as

$$\sum_{x_j}\phi_j(x_j)\psi_{ji}(x_j, x_i)\prod_{k\in N(j)\setminus i}\mu_{k\to j}(x_j). \qquad (9)$$

To improve the stability of the algorithm, we initialize marginal distribution of every query attribute x_i in the supervised layer to be $P(x_i = 1) = 0.5$ and $P(x_i = 0) = 0.5$.

To compute the joint probability $P_{sup}(Q)$ in the last step of Algorithm 1, we rewrite it as product of conditional probabilities. For example, if query $Q = \{q_1, q_2, q_3\}$, $P_{sup}(Q) = P_{sup}(q_1|q_2, q_3)P_{sup}(q_2|q_3)P_{sup}(q_3)$. To compute conditional probabilities, *e.g.* $P(x|q_2 = 1, q_3 = 0)$, we set the margins $P(q_2 = 1) = 1$ and $P(q_3 = 0) = 1$, and then run the procedures above. It is easy to show that, we can get $P_{sup}(Q)$ by performing Alternating Inference $2^{|Q|} - 1$ times, which is small given the fact that $|Q|$ is small. This assumption is valid for the reason that most images are only associated with a limited number of query attributes. For instance, the average number of attributes present in one image of a-PASCAL datasets is 7.1, while this number for a-TRECVID dataset is only 2.6, for images that have at least one label. In our configuration, we set $|Q| \leq 3$.

We found empirically that the convergence of Algorithm 1 is fast, usually within less than 10 iterations. Therefore, weak attribute selection based on Equation 7 can be computed efficiently, and Equation 3 can be solved efficiently with Equation 5.

4. Experiments

Our implementation of structural SVM is based on [10] with its Matlab wrapper[5], under the 1-slack formulation. We use regCLRG [4] to learn the latent tree graphical model for each layer of the semi-supervised graphical model. This method is found to be effective in terms of both efficiency and performance. The UGM package [6] is used for tree graphical model inference. Following [23], Hamming loss for binary classification is used as the loss function throughout the experiments:

$$\triangle(Y_t^*, Y_t) = 1 - \frac{|Y_t \cap Y_t^*| + |\bar{Y}_t \cap \bar{Y}_t^*|}{|\mathcal{Y}_l|}. \qquad (10)$$

Accordingly, our evaluation is based on mean AUC (Area Under Curve), which is a standard measurement commonly used to evaluate performance of binary classification tasks, in our case, image retrieval. Note that the AUC measure of a random guess system is 0.5. The framework of weak attributes and structural learning is general. Other loss functions such as precision, recall and NDCG can also be utilized, in which case the evaluation measurement should be changed accordingly.

In the following sections, we report and discuss the experiment results on a-PASCAL, a-Yahoo and a-TRECVID datasets. Separate testing of weak attributes on more datasets can be found in [29].

4.1. a-PASCAL and a-Yahoo

Our first evaluation is on the a-PASCAL dataset [5], which contains 12,695 images (6,340 for training and 6,355 for testing) collected from the PASCAL VOC 2008 challenge[7]. 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.* Another evaluation is on a-Yahoo dataset [5], including 2,644 test images, collected for 12 object categories from the Yahoo images search engine. Each image in a-Yahoo is described by the same set of 64 attributes, but with different category labels compared to a-PASCAL, including wolf, zebra, goat, donkey, monkey *etc.*

Following the setting of [5], we use the pre-defined training images of a-PASCAL as training set, and test on predefined test images of a-PASCAL and a-Yahoo respectively. We use the feature provided in [5]: 9,751-dimensional features of color, texture, visual words, and edges to train individual classifiers. Other weak attributes include:

- Scores from Classemes semantic classifiers [24]: 2,659 classifiers trained on images returned by search engines of corresponding query words/phrases;
- Discriminative attributes [5], 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;

[5] http://www.vlfeat.org/~vedaldi/code/svm-struct-matlab.html

[6] http://www.di.ens.fr/~mschmidt/Software/UGM.html

[7] http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2008/

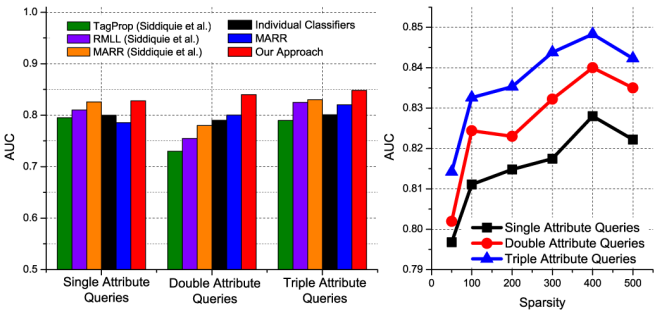

Figure 3. Retrieval performance on a-PASCAL dataset. Left: AUC comparison based on optimal sparsity ($k = 400$). The first three results are copied from [23], under the same configurations compared to ours. The last three results are based on our implementation. Right: AUC of our approach with varying sparsity.

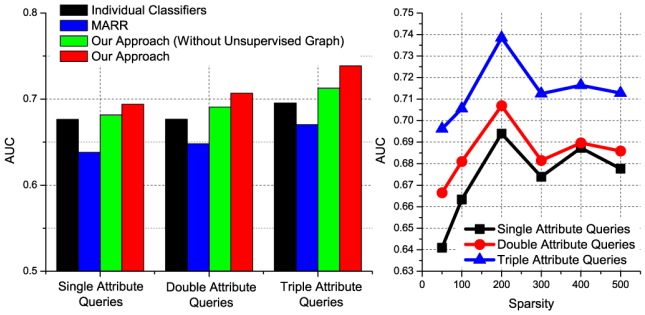

Figure 4. Retrieval performance on a-Yahoo dataset. Results reported here are based on our implementation. Left: AUC comparison based on optimal sparsity ($k = 200$). Right: AUC of our approach with varying sparsity.

- Latent variables, as described in Section 3.3.

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

Figure 3 shows the comparisons of our method to several existing approaches, including TagProp [9], Reverse Multi-Label Learning (RMLL) [22], Multi-Attribute based Ranking and Retrieval (MARR) [23], individual classifier scores from [13], and our implementation of MARR. Our approach outperforms all other methods for all types of queries, especially with large margins for double and triple query scenarios (Figure 3 Left). An example of retrieval result is shown in Figure 6. We also evaluate the effect of the sparsity level k, as shown in Figure 3 (Right). Our approach reaches the best performance with sparsity $k = 400$ (only 8% of all the weak attributes). Beyond this point, the performance begins to drop, possibly due to overfitting. This validates the assumption we made earlier that for each query, only a partial set of weak attributes are related. In terms of speed, our implementation requires 10 hours for training, with sparsity $k = 400$, on a 8-core 2.8GHz Intel workstation. The prediction can be done in real time.

Figure 4 shows the performance of our methods on the a-Yahoo benchmark compared to individual classifiers and MARR. Image categories of a-Yahoo and a-PASCAL are different, resulting in different data statistics. Therefore, training on a-PASCAL and testing on a-Yahoo can be understood as a "cross-dataset" task, which is ideal for evaluating the power of the proposed semi-supervised graphical model. From Figure 4 (Left), the performance of MARR is worse than that of individual classifiers, most likely due to cross-dataset issues. In turn, our method outperforms individual classifiers for all types of queries.

To validate the merit of integrating the supervised graph and unsupervised graph into a semi-supervised model (Section 3.3), we have further evaluated the proposed model without the unsupervised layer. As expected, the performance drops compared to the semi-supervised model (Figure 4 Left). This is an evidence validating the contribution of the semi-supervised graphical model in mapping query attributes to a large weak attribute pool.

From Figure 4 (Right), the optimal sparsity of a-Yahoo ($k = 200$) is lower than that of a-PASCAL ($k = 400$), meaning that for cross-dataset scenario, less dependency patterns between query attribute and weak attributes are generalizable. Nevertheless, our semi-supervised approach can successfully uncover such sparse patterns.

For both a-PASCAL and a-Yahoo, we have found that more than 90% weights of the dependency model are from the weak attributes (excluding individual query attribute classifier scores). This clearly demonstrates the contribution of weak attributes in the proposed framework.

4.2. a-TRECVID

To further test the effectiveness of our model, we have compiled the largest multi-attribute image retrieval dataset to date, named a-TRECVID[8]. It contains 126 uniquely labeled query attributes, and 6,000 weak attributes, of 0.26 million images. The 126 query attributes are listed in Table 1. This dataset is compiled from the TRECVID 2011 Semantic Indexing (SIN) track common annotation set[9] by discarding attributes with too few positive images, and images with too few local feature detection regions. The original dataset includes about 0.3 million video frames extracted from videos with durations ranging from 10s to just longer than 3.5 minutes. The total length of the videos is about 200 hours. Originally, there are 346 fully labeled, unique query attributes for the video frames [1]. The attributes are mostly from the concepts defined in the LSCOM multimedia ontology [17].

The individual attribute classifiers are trained using bag-of-words SIFT features under the framework of [15], with 1000-dimensional dictionary formed by k-means clustering, and 2 level spacial pyramid. Following the setting

[8]The images, labeled attributes, and computed weak attributes are described in http://www.ee.columbia.edu/dvmm/a-TRECVID

[9]http://www-nlpir.nist.gov/projects/tv2011/#sin

Boy	Airplane	Airplane_Flying	Trees	Classroom	Government-Leader	Highway	Politicians	Dark-skinned_People
Car	Bicycles	Daytime_Outdoor	Child	Reporters	Animation_Cartoon	Kitchen	Black_Frame	Domesticated_Animal
Gun	Building	Ground_Vehicles	Lakes	Teenagers	Apartment_Complex	Meeting	Blank_Frame	Eukaryotic_Organism
Face	Cheering	People_Marching	Animal	Carnivore	Female_Human_Face	Outdoor	Wild_Animal	Female_News_Subject
Girl	Mountain	Walking_Running	Beards	Herbivore	Head_And_Shoulder	Running	Anchor_Person	Construction_Vehicles
Hand	Suburban	Civilian_Person	Driver	Quadruped	Human_Young_Adult	Singing	Asian_People	Instrumental_Musician
Road	Swimming	Female_Reporter	Indoor	Old_People	Male_News_Subject	Streets	Sitting_Down	Waterscape_Waterfront
City	US_Flags	Hispanic_Person	Person	Scene_Text	Military_Aircraft	Walking	Urban_Scenes	Residential_Buildings
News	Clearing	Male_Human_Face	Forest	Body_Parts	Adult_Female_Human	Glasses	Female_Person	Celebrity_Entertainment
Room	Speaking	Office_Building	Hockey	Caucasians	Man_Wearing_A_Suit	Skating	Overlaid_Text	Male-Face-Closeup
Actor	Standing	House_Of_Worship	Mammal	Junk_Frame	Military_Personnel	Talking	Single_Person	Demonstration
Adult	Bicycling	Press_Conference	Athlete	Urban_Park	Religious_Building	Traffic	Amateur_Video	Studio_Anchorperson
Beach	Boat_Ship	Roadway_Junction	Dancing	Vertebrate	Single_Person_Male	Valleys	Male_Reporter	Female-Face-Closeup
Birds	Cityscape	Adult_Male_Human	Flowers	Male_Person	Speaking_To_Camera	Windows	Man_Made	Text_On_Artificial_Bk

Table 1. 126 query attributes of a-TRECVID, selected from a pool of 346 concepts defined in TRECVID 2011 SIN task, by discarding attributes with too few positive images.

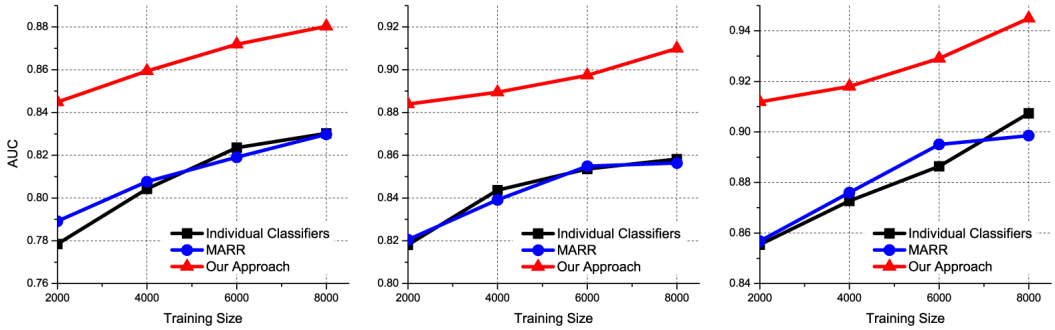

Figure 5. Retrieval performance on a-TRECVID dataset, with the varying training size. From left to right: performance of single, double and triple attribute queries.

of Section 4.1, weak attributes include individual classifier scores, Classemes, discriminative attributes, distance to randomly selected images, and latent variables. Different from a-PASCAL dataset, there is no category labels in a-TRECVID. We therefore treat images from the same video as belonging to the one category. Thus, the number of categories of a-TRECVID is much larger than that of a-PASCAL, and we have selected 1,000 more discriminative attributes for this dataset. This leads to 6,000 weak attributes per image.

Figure 5 shows the performance of our approach comparing to individual classifiers and MARR. MARR is only marginally better than individual classifiers, for the reason that the limited feature space is not scalable for the large-scale setting. Our method significantly outperforms both individual classifiers and MARR by 5% – 8.5%. This experiment validates our assumption that the proposed approach can handle large-scale image retrieval with extremely small training size. In particular, when using 2,000 images (0.8% of the whole dataset) for training, our method already outperforms both individual classifiers and MARR approaches with 8,000 images for training. An example of retrieval results with 6000 training images is shown in Figure 6.

In addition, 80% weights of the dependency model are from the weak attributes (excluding individual query attribute classifier scores). This has verified the contribution of weak attributes in our framework.

5. Conclusion

We introduce *weak attributes* that unify different kinds of mid-level image representations which can be easily acquired with no or little human labor. Based on the large and expressive weak attribute space, robust retrieval model can be developed. Under the framework of structural learning, we extend attribute dependency model originally defined over a small close set of query attributes to a more general and powerful one that maps query to the entire pool of weak attributes.

To efficiently learn the dependency model without overfitting, we select a sparse set of weak attributes for each query by using by a novel semi-supervised graphical model. It further enables our approach to be effective for cross-dataset and large-scale scenarios.

We have carried out extensive evaluations on several benchmarks, demonstrating the superiority of the proposed method. In addition, we compile the largest multi-attribute image retrieval dataset to date, named a-TRECVID, including 126 fully labeled query attributes and 6,000 weak at-

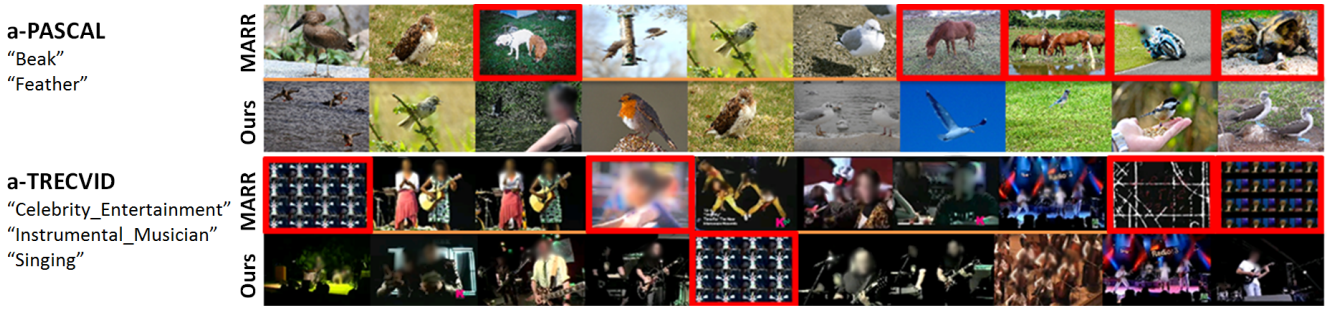

Figure 6. Top-10 results of MARR and our approach based on two query examples of a-PASCAL and a-TRECVID. Images with red frames are false positives. Note that in the third image of "Ours" for a-PASCAL, there is a bird in the background.

tributes of 0.26 million images.

Acknowledgement We would like to thank Dr. Michele Merler, Dr. Behjat Siddiquie and Dr. Neeraj Kumar for their help. We also thank Dr. Dong Liu, Prof. Ferran Marques and anonymous reviewers for their insightful suggestions.

References

[1] S. Ayache and G. Quénot. Video corpus annotation using active learning. In *ECIR*, 2008.

[2] T. Berg, A. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. In *ECCV*, 2010.

[3] C. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[4] M. Choi, V. Tan, A. Anandkumar, and A. Willsky. Learning latent tree graphical models. *JMLR*, 12:1771–1812, 2011.

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

[6] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories using google's image search. In *ICCV*, 2005.

[7] V. Ferrari and A. Zisserman. Learning visual attributes. In *NIPS*, 2007.

[8] D. Grangier and S. Bengio. A discriminative kernel-based model to rank images from text queries. *PAMI*, 30(8):1371–1384, 2008.

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

[10] T. Joachims, T. Finley, and C. Yu. Cutting-plane training of structural svms. *Machine Learning*, 77(1):27–59, 2009.

[11] J. Krapac, M. Allan, J. Verbeek, and F. Jurie. Improving web image search results using query-relative classifiers. In *CVPR*, 2010.

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

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

[14] C. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009.

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

[16] Q. Le and A. Smola. Direct optimization of ranking measures. http://arxiv.org/abs/0704.3359, 2007.

[17] M. Naphade, J. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *Multimedia, IEEE*, 13(3):86–91, 2006.

[18] 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.

[19] D. Parikh and K. Grauman. Interactively building a discriminative vocabulary of nameable attributes. In *CVPR*, 2011.

[20] D. Parikh and K. Grauman. Relative attributes. In *ICCV*, 2011.

[21] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *PAMI*, 27(8):1226–1238, 2005.

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

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

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

[25] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *JMLR*, 6(2):1453–1484, 2006.

[26] D. Vaquero, R. Feris, D. Tran, L. Brown, A. Hampapur, and M. Turk. Attribute-based people search in surveillance environments. In *WACV*, 2009.

[27] Y. Wang and G. Mori. A discriminative latent model of object classes and attributes. In *ECCV*, 2010.

[28] A. Yanagawa, S.-F. Chang, L. Kennedy, and W. Hsu. Columbia university's baseline detectors for 374 lscom semantic visual concepts. *Columbia University ADVENT Technical Report # 222-2006-8*, 2007.

[29] F. Yu, R. Ji, M.-H. Tsai, G. Ye, and S.-F. Chang. Experiments of image retrieval using weak attributes. *Columbia University Computer Science Department Technical Report # CUCS 005-12*, 2012.