Learning Multiple Relative Attributes with Humans in The Loop

Buyue Qian, Xiang Wang, Nan Cao, Yu-Gang Jiang, Ian Davidson

Abstract—Semantic attributes have been recognized as a more spontaneous manner to describe and annotate image content. It is widely accepted that image annotation using semantic attributes is a significant improvement to the traditional binary or multi-class annotation due to its naturally continuous and relative properties. Though useful, existing approaches rely on an abundant supervision and high quality training data, which limit their applicability. Two standard methods to overcome small amounts of guidance and low quality training data are transfer and active learning. In the context of relative attributes this would entail learning multiple relative attributes simultaneously and actively querying a human for additional information. This work addresses the two main limitations in existing work: i) it actively adds humans to the learning loop so that minimal additional guidance can be given, and ii) it learns multiple relative attributes simultaneously and thereby leverages dependence amongst them. In this paper, we formulate a joint active learning to rank framework with pairwise supervision to achieve these two aims which also has other benefits such as the ability to be kernelized. The proposed framework optimizes over a set of ranking functions (measuring the strength of the presence of attributes) simultaneously and dependently on each other. The proposed pairwise queries take the form of "which one of these two pictures is more natural?", and can be easily answered by humans. Extensive empirical study on real image datasets shows that our proposed method, compared to several state-of-the-art methods, achieves superior retrieval performance while requires significantly less human inputs.

Index Terms—Active Learning; Learning to Rank; Humans-in-the-loop; Image Recognition; Relative Attributes.

1 Introduction

One core component, when constructing advanced image management systems, is to design an effective content based image retrieval (CBIR) scheme [1] [2] [3] [4]. Earlier studies have looked at using relevance feedback to interactively produce a desirable retrieval function [5] [6] [7] [8]. Though useful, traditional relevance feedback based image retrieval has two main limitations: (i) it assumes semantic attributes as binary predicates indicating whether an image contains certain properties, and (ii) it can be difficult for users to label images since images can be complicated and ambiguous. Conventional image annotation mainly focuses on providing the estimation of categorical labels, such as predicting whether an image contains "people" or "mountain". However, humans tend to describe an image using more natural language by giving rich semantic descriptions in a non-boolean manner. Relative attributes was recently proposed to provide natural textual descriptions and annotations for images using learning-to-rank techniques [9]. Though relative attributes approaches have shown practical success in various applications like outdoor scene recognition and product search [10], there are two limitations of relative attributes that are still under studied, which if well addressed can further improve its applicability. (i) Relevant attributes can be learnt *jointly* [11], such that the relations amongst multiple attributes can be exploited. (ii) While most of the existing attribute learning methods are passive models relying on the abundance of training data (which may not be the case in practice), *active* attributes learning [12] [13] proactively asks humans informative questions so as to improve the learning accuracy with minimal efforts.

Motivation. Inspired by previous studies [9], we aim to improve relative attributes learning in the following two aspects: (1) *Relatedness* and (2) *Activeness*. Relatedness indicates that rather than learning each attribute individually, one would achieve better learning performance through exploiting the dependence amongst multiple attributes. Activeness denotes an efficient way to integrate the strengths of computational modeling and human perception. Another consideration in our mind when we design the algorithm is to make the (active) questions easier to answer. Labeling an image requires global view of the data, especially for ranking problems, since images can be complex involving multiple labels. Therefore, we choose to make use of pairwise constraints, as ordering a pair of images only requires local view of the data. Though pairwise active learning to rank prompts more questions, the unit cost per question is lower. This is particularly true in the applications requiring domain knowledge, such as medical image analysis. In practice, pairwise constraints can be easily obtained through various methods, such as Mechanical Turk, or generated from implicit user feedback, side information, or crowdsourcing. As shown

- *Buyue Qian, Xiang Wang, and Nan Cao are with IBM T. J. Watson Research, Yorktown Heights, NY, 10598.*
  E-mail: {bqian, wangxi, nancao}@us.ibm.com
- *Yu-Gang Jiang is with School of Computer Science, Fudan University, Shanghai, China.*
  E-mail: ygj@fudan.edu.cn
- *Ian Davidson is with Dept. of Computer Science, University of California, Davis, CA 95616.*
  E-mail: davidson@cs.ucdavis.edu

in Fig. 1, the human-machine interaction in our humans-in-the-loop scheme is accomplished by actively asking users to compare pairwise image samples.

Challenges. Compared with typical attribute learning, active and multi-attribute learning provides the following three distinctive challenges. (1) *How to model the relationships between attributes?* In this paper such relations are captured by allowing multiple attributes to share a common ranking basis. Then the learning is formulated as a large "margin" problem, which is later kernelized to form a nonlinear model. As we adopt pairwise supervision, which is different from typical active learning studied in the context of querying labels, the next challenge would naturally be (2) *What is an informative image pair to query?*, and given a pair of images (3) *Which attribute should we query?*

Proposal. We propose a query selection strategy with the following properties: (i) the ranking model is most uncertain about the ordering of the image pair to query (local significance); (ii) the overall rank of this pair to the remaining images is also unclear to the ranking model (global significance); (iii) querying is conducted w.r.t. a single *salient* attribute for each pair. The goal in the training stage is to simultaneously find a set of ranking functions that satisfies a maximum number of pairwise orderings, which are provided by both the initial training data and later human inputs. In the test stage, for a given collection of images, each ranking function assigns a real-valued score to every image indicating the relative presence of an individual attribute.

2 Related Work

Our work involves two areas in pattern recognition and image retrieval - *image attributes learning* and *humans-in-the-loop*. We shall now review in more detailed specific papers for both areas.

Image attributes learning. Recognition of categorical attributes has been extensively studied in computer vision. For instance, Ferrari *et al* [14] proposes a probabilistic generative model to predict color and texture patterns, based on which many object recognition systems are built [15], [16]. Beyond simply recognizing an object, some existing studies provide more intelligent ways to predict image structures or semantically describe an image [17], [18], [19], [20], for example, there is an interesting application [21] to generate a sentence to describe an image based on its visual content. A comprehensive survey and empirical comparison of objects ranking methods can be found in [22]. Instead of learning predefined attributes, researchers also have looked at automatic discovery of image semantic attributes using existing knowledge [23], [24], and crowd-sourcing [25]. Recent research tends to learn relative attributes as it is a more natural way to describe images. Kumar *et al* [26] explore comparative facial attributes for face verification, and generate semantic facial descriptions such as "Lips like Barack Obama". Wang *et al* [27] make use of explicit similarity-based supervision, such as "A serval is like a leopard". Parikh *et al* [9] propose a SVM based approach to explore relative attributes of images, for example "the rounder pillow". Shrivastava *et al* [28] present an attribute learning approach in semi-supervised setting. Though useful, previous studies on relative attributes solely focus on learning a single attribute each time, and the relations amongst multiple attributes are ignored. This paper extends previous research in terms of learning multiple relevant attributes jointly, such that dependence amongst attributes are exploited.

Humans-in-the-loop. The idea of humans-in-the-loop, which allows human to input additional information to guide the learning, has been explored in many previous studies [29], [30]. The objective of human-aided machine vision is to effectively improve the visual recognition or retrieval accuracy with a minimal amount of human effort [31], [32], [33]. Existing works have looked at different ways to utilize human knowledge. Relevance feedback interleaves image retrieval with human response, [34] allows users to mark the relevance of image search results, which are then used to refine the search queries. Whittle Search [10] allows users to provide relative attribute feedback on a set of reference images. Parikh and Grauman [35] present an interesting human-machine interaction that aims to interactively discover a discriminative vocabulary of nameable attributes for images. More recently, Biswas and D. Parikh [36] designed an interactive learning paradigm to actively acquire informative negative samples. Our recent work in active learning looked at active learning to rank [37], easier active learning from relative queries [38], and fast query selection for large scale learning to rank [39]. In this paper, based on previous studies, we present a joint learning framework that actively queries humans for additional information and also learns a set of relative attributes simultaneously.

3 Method

In this section, we start from giving the overview and preliminaries of the problem to address. We shall then present the details of the proposed method.

3.1 An Overview

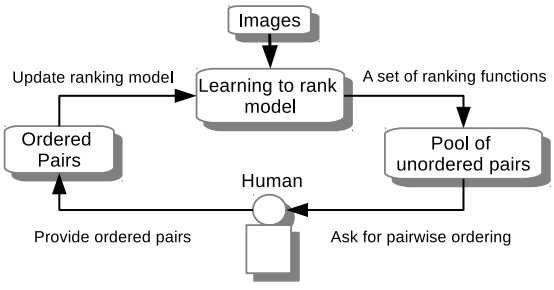


Fig. 1 The cycle of active multi-attribute learning.

Similar to regular active learning system, there are two key components in an active multi-attribute system: (i) a learning to rank model and (ii) a query selection strategy.

The learning to rank model takes training data and outputs retrieval functions. We in our approach require the learning to rank model to have the following two properties: (1) the learning to rank model learns a set of retrieval functions jointly and simultaneously, so that the dependency amongst multiple relative attributes can be captured and exploited; (2) the learning to rank model takes pairwise constraints so that the relative answers (for example, image A is more relevant to the keyword "mountain" than image B) provided by humans can be encoded and utilized by the learning model.

The query selection strategy aims to find the most informative questions to ask humans, such that the retrieval performance can be effectively improved with minimal human efforts. We in our approach require the query selection strategy to have the following two properties: (1) the query strategy finds a pair of images each time (different to typical active learning where we only find a single image for labeling at each time); (2) for an identified pair of images, the query strategy finds the most informative attribute, with respect to which the querying should be conducted.

TABLE 1 Notation Table

Variables	Definitions
$\mathbf{x}_i$	Feature vector of image i
$\mathbf{X}$	The collection of image feature vectors
M	Number of attributes (or categorical labels)
R_t	The set of strongly ordered image pairs for attribute t
S_t	The set of weakly ordered image pairs for attribute t
τ_t	Ranking function of attribute t
w_0	The common ranking basis
v_t	Ranking variation of attribute t w.r.t. the base
w_t	Ranking hyperplane (vector) of attribute t
ξ_{ij}^t	The slack variable for a strongly ordered image pair (i,j) of attribute t
γ_{ij}^t	The slack variable for a weakly ordered image pair (i,j) of attribute t
λ	The parameter to emphasize the commonality of multiple attributes
C	The parameter to penalize the slacks
α_{ij}^t	Lagrange multiplier for a strongly ordered image pair (i,j) of attribute t
β_{ij}^t	Lagrange multiplier for a weakly ordered image pair (i,j) of attribute t
$\mathcal{LS}(\mathbf{x}_i, \mathbf{x}_j, t)$	Local significance of an image pair (i,j) w.r.t attribute t
$\mathcal{GS}(\mathbf{x}_i, \mathbf{x}_j, t)$	Global significance of an image pair (i,j) w.r.t attribute t
K_{ij}	Entry (i,j) of a kernel matrix K
(i^*, j^*, t^*)	The most informative query, image pair (i,j) w.r.t. attribute t
p	Parameter to emphasize the global significance
$\mathcal{N}(\bullet)$	A normalizing operation, to make a vector $\bullet$ non-negative and sum-to-one

Fig. 1 illustrates the conceptual diagram of our active (human-aided) multi-attribute learning framework. A learning cycle starts with the simultaneous training of multiple ranking functions, which is accomplished using the ranking method described in Section 3.3 (linear) or its kernelized version described in 3.4 (nonlinear). Then using the query selection strategy presented in Section 3.5, we locate *an informative pair of images with its salient attribute*, and ask humans for its pairwise ordering with respect to the identified salient attribute. The answer/ordering obtained from humans will later be encoded as an additional constraint to update the ranking/retrieval functions. This process repeats until a desirable accuracy is achieved or a certain stopping criterion is satisfied. We shall in the rest of this section present our proposed learning model and query selection strategy for active multi-attribute learning, present solutions for the optimization, explain the intuitions behind the math, and discuss the practical usage of our formulations.

3.2 Preliminaries

Given a set of n images represented in $\mathbb{R}^d$ space by feature vectors $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$, we define a ranking problem which aims to learn a set of ranking functions w.r.t. M attributes simultaneously. As the prior knowledge, for each attribute we are given two sets of ordered pairs: a set of *strongly* ranked pairs of images $R_t = \{(i,j)\}$ such that $(i,j) \in R_t$ means image i has *confidently* stronger presence of the attribute t than image j; and a set of *weakly* ranked pairs of images $S_t = \{(i,j)\}$ such that $(i,j) \in S_t$ means image i has *slightly* stronger presence of the attribute t than image j. Since we aim at developing an inductive ranking algorithm, the training set for each attribute cannot be empty at the beginning of training, i.e. $R_t \cup S_t \neq \varnothing$. The training set grows as the algorithm iteratively queries knowledge from human experts resulting in more ordered pairs add to R_t and S_t. The notations of our formulation are summarized in TABLE 1.

The goal is to learn a linear ranking function τ_t for each attribute. Let w_t denote the ranking hyperplane for attribute t, the ranking score of an image $\mathbf{x}$ can be obtained using

$$\tau_t(\mathbf{x}) = w_t^T \mathbf{x}, \tag{1}$$

To capture the relation amongst the multiple attributes, we assume that the multiple ranking functions share a common base w_0. The intuition behind the assumption is that all ranking functions w_t come from a particular probability distribution such as a Gaussian, which implies that all w_t are "close" to some mean function w_0. In particular, w_t can be written as the sum of the base and a variation v_t, for every attribute t, as

$$w_t = w_0 + v_t \tag{2}$$

where the ranking base vector w_0 is "large" and the variation vectors v_t are "small" when the multiple attributes are similar to each other. The ranking basis w_0 can be viewed as capturing the commonality of

multiple attributes, and the variation vector v_t can be interpreted as capturing the unique (compared to other attributes) aspects of attribute t. Importantly, having a common underlying w_0 allows transferring between the attributes. In particular it allows deficient label to be overcome by other attribute labeling.

3.3 Learning Multiple Attributes

As we adopt pairwise supervision, each ordered image pair in R_t or S_t can be encoded as a pairwise constraint. Then the ranking problem is to optimize w_t on ordered images such that a maximum number of the given constraints are satisfied. As pointed out in [9], deriving the optimal w_t by simply satisfying a maximum number of pairwise orderings is intractable. Instead, we estimate both w_0 and v_t using a large "margin" approach [40]. Additionally, we further relax the problem using two non-negative slack variables ξ^t and γ^t for each attribute t, since completely obeying the constraints is undesirable if the data contains errors or noise. A smooth ranking hyperplane that ignores a few pairwise orderings is better than one that loops around the outliers. This allows the ranking functions to be slightly on the wrong side of the given constraints. Slack variables are penalized to prevent the trivial solutions which involve large slacks that allow any hyperplane, the multiple attributes learning problem is then formulated as:

Primal Problem:

$$\begin{aligned}
\min_{w_0,v_t,\xi,\gamma} \quad & \frac{1}{2}w_0^T w_0 + \frac{\lambda}{2M}\sum_{t=1}^{M} v_t^T v_t + C\sum_{t=1}^{M}\sum_{i,j}\left(\xi_{ij}^t + \gamma_{ij}^t\right) \\
s.t. \quad & (w_0+v_t)^T(\mathbf{x}_i - \mathbf{x}_j) \geq 1 - \xi_{ij}^t, \quad if\ (i,j) \in R_t; \\
& (w_0+v_t)^T(\mathbf{x}_i - \mathbf{x}_j) \geq -\gamma_{ij}^t, \quad if\ (i,j) \in S_t; \\
& \xi_{ij}^t \geq 0; \quad \gamma_{ij}^t \geq 0. \qquad \forall i,j,t
\end{aligned} \tag{3}$$

where λ is a positive regularization parameter controlling how much the solutions w_t differ from each other, and C is a constant for the slacks penalizing the errors that each of the final models w_t makes on the training data. Intuitively, setting a large λ tends to make all models similar, while a small λ tends to make the attributes less related in terms of a nearly zero w_0. The objective function could be interpreted as finding a trade off between having each model w_t maximally stretch the ordered pairs and having each model close to the average model w_0. The first set of constraints shown in Eq.(3) is used to enforce the ranking distance $w_t^T(\mathbf{x}_i - \mathbf{x}_j)$ on a *strongly* ordered pair to be greater than one, while the second set of constraints only requires a slight ranking distance on *weakly* ordered pairs.

The primal optimization described in Eq.(3) is convex [41], and equivalent to a SVM problem but on M attributes with pairwise differences $(\mathbf{x}_i - \mathbf{x}_j)$. Therefore, the optimization can be solved using a variation of decomposition SVM solvers, such as SVM^{light} [42]. As the slack penalty term on variables ξ^t and γ^t in Eq.(3) can be viewed as the hinge loss, which is not differentiable. There are two differentiable loss functions that can be used to substitute the hinge loss, such that the optimal solution can be recovered by applying gradient based optimization [9], [43]: (i) hinge loss can be approximated using the Huber loss [43] and (ii) the penalty term can be formulated as quadratic loss.

3.4 Kernelization and The Dual

In practical image retrieval systems it is unlikely that a simple linear function performs desirably, thereby nonlinear approaches are preferred if the computational complexity is not significantly raised. Using the kernel trick, a nonlinear ranking function could be easily obtained from a linear one followed by a mapping function ϕ, which projects the original image feature vectors to a higher (even infinite) dimensional space where there is a better chance that more pairwise constraints are satisfied. Applying a mapping function ϕ, the linear retrieval shown in Eq.(1) becomes nonlinear.

$$\tau_t(\mathbf{x}) = w_t^T \phi(\mathbf{x}) \tag{4}$$

To make use of the kernel, we need to derive the dual optimization of Eq.(3). There are three major reasons that we solve the multiple attributes problem in its dual format [43]: (1) The duality theory provides a convenient way to handle the constraints; (2) There are less variables need to be recovered in the dual formulation (four in primal form but only two in the corresponding dual form); (3) The dual optimization could be expressed in terms of the dot products of image feature vectors, thus making it possible to apply kernel functions. According to the Lagrangian theory and KKT (Karush-Kuhn-Tucker) conditions, strong duality holds in the primal problem. Therefore, the dual problem can be derived by the following procedure. The Lagrange primal function (denoted by L) can be obtained by subtracting the products of the Lagrange multipliers (α^t, β^t) and the two sets of pairwise constraints. The Lagrange primal function L can be expressed as:

$$\begin{aligned}
& \min_{w_0,v_t,\xi,\gamma}\ \max_{\alpha,\beta} L = \\
& \frac{1}{2}w_0^T w_0 + \frac{\lambda}{2M}\sum_{t=1}^{M} v_t^T v_t + C\sum_{t=1}^{M}\sum_{i,j}\left(\xi_{ij}^t + \gamma_{ij}^t\right) \\
& -\sum_{t=1}^{M}\sum_{(i,j)\in R_t} \alpha_{ij}^t\left((w_0+v_t)^T(\mathbf{x}_i - \mathbf{x}_j) - 1 + \xi_{ij}^t\right) \\
& -\sum_{t=1}^{M}\sum_{(i,j)\in S_t} \beta_{ij}^t\left((w_0+v_t)^T(\mathbf{x}_i - \mathbf{x}_j) + \gamma_{ij}^t\right) \\
s.t. \quad & \xi_{ij}^t \geq 0; \qquad \gamma_{ij}^t \geq 0; \\
& \alpha_{ij}^t \geq 0; \qquad \beta_{ij}^t \geq 0
\end{aligned} \tag{5}$$

Note that the non-negative slack constraints could be kept as they can be easily handle directly. We then take

the first-order derivatives of this Lagrange function L with respect to both w_0 and v_t, and set them to zero, i.e., $\frac{\partial L}{\partial w_0} = 0$ and $\frac{\partial L}{\partial v_t} = 0$.

$$\begin{aligned}\frac{\partial L}{\partial w_0} &= w_0 - \sum_{k=1}^{M} \sum_{(i,j)\in R_k} \alpha_{ij}^k (\mathbf{x}_i - \mathbf{x}_j) \\ &\quad - \sum_{k=1}^{M} \sum_{(i,j)\in S_k} \beta_{ij}^k (\mathbf{x}_i - \mathbf{x}_j) = 0 \qquad (6)\end{aligned}$$

$$\begin{aligned}\frac{\partial L}{\partial v_t} &= \frac{\lambda}{M} v_t - \sum_{(i,j)\in R_t} \alpha_{ij}^t (\mathbf{x}_i - \mathbf{x}_j) \\ &\quad - \sum_{(i,j)\in S_t} \beta_{ij}^t (\mathbf{x}_i - \mathbf{x}_j) = 0 \qquad (7)\end{aligned}$$

From the above equations, we can derive the kernelized versions of w_0 and v_t, which are written as

$$w_0 = \sum_{k=1}^{M} \left(\sum_{(i,j)\in R_k} \alpha_{ij}^k (\mathbf{x}_i - \mathbf{x}_j) + \sum_{(i,j)\in S_k} \beta_{ij}^k (\mathbf{x}_i - \mathbf{x}_j) \right) \qquad (8)$$

$$v_t = \frac{M}{\lambda} \left(\sum_{(i,j)\in R_t} \alpha_{ij}^t (\mathbf{x}_i - \mathbf{x}_j) + \sum_{(i,j)\in S_t} \beta_{ij}^t (\mathbf{x}_i - \mathbf{x}_j) \right) \qquad (9)$$

where α^t and β^t are two sets of Lagrange multipliers that are used to encode the pairwise constraints into the objective. For the convenience of expression, let A denote the dot product of two pairs of images $(\mathbf{x}_i - \mathbf{x}_j)^T(\mathbf{x}_k - \mathbf{x}_l)$. By substituting the new w_0 and v_t (Eq.(8) and (9)) into the Lagrange primal function L, we can derive the dual as follows (the full deviation of the dual problem can be found in the supplemental material).

Dual Problem:

$$\begin{aligned}\max_{\alpha,\beta} \quad & \sum_{t=1}^{M} \sum_{(i,j)\in R_t} \alpha_{ij}^t - \frac{1}{2} \sum_{t=1}^{M} \sum_{v=1}^{M} \sum_{(i,j)\in R_t} \sum_{(k,l)\in R_v} \alpha_{ij}^t \alpha_{kl}^v A \\ & - \frac{1}{2} \sum_{t=1}^{M} \sum_{v=1}^{M} \sum_{(i,j)\in S_t} \sum_{(k,l)\in S_v} \beta_{ij}^t \beta_{kl}^v A \\ & - \sum_{t=1}^{M} \sum_{v=1}^{M} \sum_{(i,j)\in R_t} \sum_{(k,l)\in S_v} \alpha_{ij}^t \beta_{kl}^v A \\ & - \frac{M}{2\lambda} \sum_{t=1}^{M} \sum_{(i,j)\in R_t} \sum_{(k,l)\in R_t} \alpha_{ij}^t \alpha_{kl}^t A \\ & - \frac{M}{2\lambda} \sum_{t=1}^{M} \sum_{(i,j)\in S_t} \sum_{(k,l)\in S_t} \beta_{ij}^t \beta_{kl}^t A \\ & - \frac{M}{\lambda} \sum_{t=1}^{M} \sum_{(i,j)\in R_t} \sum_{(k,l)\in S_t} \alpha_{ij}^t \beta_{kl}^t A \\ s.t. \quad & 0 \le \alpha_{ij}^t \le C; \quad 0 \le \beta_{ij}^t \le C. \qquad (10)\end{aligned}$$

where each term has interpretable roles, and we can clearly see how the multiple attributes are incorporated in the dual formulation. The first term comes from the ranking margin "1" in the first constraint of Eq.(3), and is used to enforce a large ranking distance on strongly ordered pairs. The interaction between all attributes is assessed in terms 2 to 4, which measure the relatedness of the multiple attributes and are used to form the common ranking base w_0. Terms 5-7 (the last 3 terms in Eq.(10)) only focus on each attribute individually, and are used to estimate the variation vectors v_t.

The proposed approach requires to select two parameters: (i) the parameter C which penalizes the training error and (ii) the parameter λ which is used to emphasize the commonality of multiple attributes rather than the individuality of each attribute. The two parameters could be selected using the cross-validation method or a validation set. According to the convexity of the primal optimization, the dual formulation is also convex [44]. Consequently, the dual optimization could be recovered as a kernel SVM problem but on M attributes and inner products of pairwise differences, and can be solved using a variant of kernel SVM solvers, such as $libsvm$ [45]. Alternating the gradient based optimizations over α^t and β^t could also produce the global optimum. In our empirical study, we simply solve the dual optimization using CVX, which is a matlab package for specifying and solving convex programs [46], [47]. We will make the code used in our experiment publicly available soon.

Finally, to derive the inductive form of the ranking score for a test image, let K denote the kernel matrix of $\mathbf{X}$ and $\mathbf{x}_*$ denote an unseen/test image. Then the kernelized ranking function τ_t (to produce ranking scores using kernel matrix instead of feature vectors) with respect to attribute t can be written as:

$$\begin{aligned}&\tau_t(\mathbf{x}_*) = \qquad (11)\\ &\sum_{k=1}^{M} \left(\sum_{(i,j)\in R_k} \alpha_{ij}^k (K_{i*} - K_{j*}) + \sum_{(i,j)\in S_k} \beta_{ij}^k (K_{i*} - K_{j*}) \right) \\ &+ \frac{M}{\lambda} \left(\sum_{(i,j)\in R_t} \alpha_{ij}^t (K_{i*} - K_{j*}) + \sum_{(i,j)\in S_t} \beta_{ij}^t (K_{i*} - K_{j*}) \right)\end{aligned}$$

3.5 Pairwise Query Selection Strategy

The primary goal of human-in-the-loop is to improve the retrieval performance with minimal human efforts, therefore, the key question to address is how to find the most informative pairwise ordering (currently unknown) w.r.t. a particular attribute. We define an influential query as requiring the two properties $\mathcal{LS}$ and $\mathcal{GS}$ which we now describe.

$\mathcal{LS}$ (*Local Significance*): The ordering of an image pair is (1) unclear to the ranking base (w_0) but (2) clear to the variance vector v_t of attribute t. Requirement (1) encourages to query a pair of images whose ordering is inconsistent between the multiple attributes, and querying of such pair can impact all ranking functions.

Requirement (2) enforces the querying to be performed on a salient attribute, since querying an attribute that does not exist in the images is meaningless. Therefore, we wish to query a pair that is uncertain to the base model w_0 w.r.t. a salient attribute of the images. We adopt the term "local" because $\mathcal{LS}$ only measures the uncertainty of our model (w.r.t. w_0) on a particular pair of images, but the overall ranking of image is ignored.

$\mathcal{GS}$ (*Global Significance*): The *overall* rank of an image is unclear to our ranking functions, i.e., considering all other images the model is uncertain about the position of an image in the total ranked list. This encourages to query images that are located near a dense area in the feature space, whose rank can impact the ranking scores of many images. By using an entropy measure we inherently choose those images that are in dense regions in the feature space, and hence if queried will have the greatest impact. We adopt the term "global" since this significance measure considers the importance of an image in the context of all images.

We shall in the rest of this subsection define the two measures and discuss their properties. To produce a query selection strategy that performs stably across different scenarios, we combine the $\mathcal{LS}$ and $\mathcal{GS}$ measures and let them help each other. Let (i^*, j^*, t^*) be the currently most informative pair of images (i^*, j^*) whose salient attribute is t^*, i.e., the next question to ask humans is to *"order image i^* and j^* w.r.t. the strength of the presence of attribute t^*"*. Our query selection is formulated as the product of the two significance measures:

$$(i^*, j^*, t^*) = \arg\max \quad \mathcal{LS}(\mathbf{x}_i, \mathbf{x}_j, t) \times \mathcal{GS}(\mathbf{x}_i, \mathbf{x}_j, t)^p \quad (12)$$

where p is a tuning parameter that balances the influence of the local and global significance measures.

Definition of $\mathcal{LS}$ (Local Significance). $\mathcal{LS}$ aims to locate the uncertain pairs whose orderings are unclear to the base model w_0, and also their corresponding salient attributes. Hence, the proposed $\mathcal{LS}$ consists of two parts: (i) a small ranking distance on w_0 which implies that the base model is confused about the ordering of a pair of images, but (ii) a large ranking distance on v_t which implies that we prefer the active querying to be performed w.r.t. a strongly presented attribute. In this sense, the $\mathcal{LS}$ is estimated using:

$$\mathcal{LS}(\mathbf{x}_i, \mathbf{x}_j, t) = |w_0^T(\mathbf{x}_i - \mathbf{x}_j)|^{-1} + |v_t^T(\mathbf{x}_i - \mathbf{x}_j)| \quad (13)$$

In this above equation, the first term encourages small ranking distance on the base model w_0. This induces the following three possible cases. (1) The ordering of image i and j is uncertain to all M attributes, thus querying its pairwise ordering would possibly benefit the overall ranking performance. (2) The multiple attributes have most disagreement on this pair, in this case it would be also beneficial to query this pair as it is the most confusing pair in a multi-attributes ranking problem. (3) The pair consists of two similar images. The second term in Eq.(13) leans towards large ranking distance on the variation v_t, and makes the following three possible cases: (4) Identify a saliently presented attribute thus making the active queries more "meaningful"; (5) Potentially has great impact to the ranking model if a human says that the ranking order specified by v_t should be reversed; or (6) Possibly makes no change to the ranking function if v_t ranks the two images far apart (> 1) and the human's answer just confirms this.

While cases (1) (2) (4) (5) are advantageous to efficiently improve the retrieval performance, cases (3) and (6) would not introduce much new knowledge to the ranking model and thus should be avoided. Consequently, the query selection cannot solely depend on $\mathcal{LS}$. We shall later in this subsection propose the definition of $\mathcal{GS}$ to alleviate the occurrence of the undesirable cases listed above. The kernelized version of $\mathcal{LS}$ could be obtained by substituting the new w_0 and v_t (Eq.(8) and (9)) into Eq.(13), as shown below:

$$\begin{aligned}
&\mathcal{LS}(\mathbf{x}_i, \mathbf{x}_j, t) = \\
&|\sum_{u=1}^{M} \sum_{(k,l)\in R_u} \alpha_{kl}^u (K_{ki} - K_{kj} - K_{li} + K_{lj}) \\
&+ \sum_{u=1}^{M} \sum_{(k,l)\in S_u} \beta_{kl}^u (K_{ki} - K_{kj} - K_{li} + K_{lj})|^{-1} \\
&+ \frac{M}{\lambda} |\sum_{(k,l)\in R_t} \alpha_{kl}^t (K_{ki} - K_{kj} - K_{li} + K_{lj}) \\
&+ \sum_{(k,l)\in S_t} \beta_{kl}^t (K_{ki} - K_{kj} - K_{li} + K_{lj})|
\end{aligned} \quad (14)$$

Definition of $\mathcal{GS}$ (Global Significance). Following the intuition that a globally significant image (overall rank is unclear) is the one whose pairwise orderings to all other images is unclear, the $\mathcal{GS}$ also can be assessed using the ranking distance $w_t^T(\mathbf{x}_i - \mathbf{x}_j)$. Consider an extreme case that there is an image $\mathbf{x}_i$ whose ranking distances to all other images are exactly the same, in our ranking model this is expressed as an uniform value of $w_t^T(\mathbf{x}_i - \mathbf{x}_j)$ for $\forall$ $\mathbf{x}_j \in \mathbf{X}$. If this is the case, we can conclude that our ranking model w_t has no idea about where to rank image $\mathbf{x}_i$, thus querying $\mathbf{x}_i$ would greatly enrich the training set. In particular, we assess $\mathcal{GS}$ using entropy, and the $\mathcal{GS}$ on an image pair is simply the sum of the entropy of each individual image. To estimate the entropy, we first normalize all possible ranking distances of a particular image to be non-negative and sum-to-one. Let $\mathcal{N}(\bullet)$ denote the operation of this type of normalization, the definition of $\mathcal{GS}$ is expressed as

$$\begin{aligned}
&\mathcal{GS}(\mathbf{x}_i, \mathbf{x}_j, t) = \\
&- \sum_{\mathbf{x}_k \in \mathbf{X}} \mathcal{N}\left(w_t^T(\mathbf{x}_i - \mathbf{x}_k)\right) \log \mathcal{N}\left(w_t^T(\mathbf{x}_i - \mathbf{x}_k)\right) \\
&- \sum_{\mathbf{x}_k \in \mathbf{X}} \mathcal{N}\left(w_t^T(\mathbf{x}_j - \mathbf{x}_k)\right) \log \mathcal{N}\left(w_t^T(\mathbf{x}_j - \mathbf{x}_k)\right)
\end{aligned} \quad (15)$$

There are two possible cases for a large $\mathcal{GS}$ value: (1) locate the images pairs whose overall rank to others are

unclear, and (2) possibly pick up noise or outliers. Case (1) is helpful in active query selection since it would improve the overall ranking performance, while case (2) is not constructive. However, the occurrence of case (2) would be significantly reduced by combining with $\mathcal{LS}$, since noise or outliers by its definition only appears in the sparse area of the feature space and it is unlikely that two outliers have a small $\mathcal{LS}$. The kernelized version of $\mathcal{GS}$ could be obtained by substituting the new w_0 and v_t (as shown in Eq.(8) and (9)) into the linear global significance function Eq.(15). For the convenience of mathematical expression, let $\mathcal{G}(\mathbf{x}_i, \mathbf{x}_k, t)$ denote:

$$\begin{aligned}&\mathcal{G}(\mathbf{x}_i, \mathbf{x}_k, t) = \\ &\sum_{z=1}^{M} \sum_{(u,v)\in R_z} \alpha_{uv}^z \left(K_{ui} - K_{vi} - K_{uk} + K_{vk}\right) \\ &+ \sum_{z=1}^{M} \sum_{(u,v)\in S_z} \beta_{uv}^z \left(K_{ui} - K_{vi} - K_{uk} + K_{vk}\right) \\ &+ \frac{M}{\lambda} \sum_{(u,v)\in R_t} \alpha_{uv}^t \left(K_{ui} - K_{vi} - K_{uk} + K_{vk}\right) \\ &+ \frac{M}{\lambda} \sum_{(u,v)\in S_t} \beta_{uv}^t \left(K_{ui} - K_{vi} - K_{uk} + K_{vk}\right)\end{aligned} \tag{16}$$

Then the kernelized $\mathcal{GS}$ is formulated as:

$$\begin{aligned}&\mathcal{GS}(\mathbf{x}_i, \mathbf{x}_j, t) = \\ &- \sum_{\mathbf{x}_k \in \mathbf{X}} \mathcal{N}\left(\mathcal{G}\left(\mathbf{x}_i, \mathbf{x}_k, t\right)\right) \log \mathcal{N}\left(\mathcal{G}\left(\mathbf{x}_i, \mathbf{x}_k, t\right)\right) \\ &- \sum_{\mathbf{x}_k \in \mathbf{X}} \mathcal{N}\left(\mathcal{G}\left(\mathbf{x}_j, \mathbf{x}_k, t\right)\right) \log \mathcal{N}\left(\mathcal{G}\left(\mathbf{x}_j, \mathbf{x}_k, t\right)\right)\end{aligned} \tag{17}$$

There are a few potential limitations in the proposed method. (i) Overfitting is more likely to occur in the dual formulation compared to its primal form which solves a linear problem. (ii) The success of learning multiple attributes requires an appropriate regularization parameter λ, whose value implies the extent of the dependence amongst multiple attributes. We shall in the next section experimentally show that the two concerns can be alleviated if the parameters are carefully chosen.

4 Empirical Study

All code used to produce the results in this paper will be made publicly available to ensure reproducibility of results. In the experiment, we attempt to understand the strengths and relative performance of our active multiple attributes learning approach which we refer to in this section as `Proposed+Active`. In particular, we in this section study four core questions. **The first question is:** *(1) How well our multiple attributes learning method compares to the following state-of-the-art single attribute learning baselines*:

1) `RankSVM` [48], a state-of-the-art ranking algorithm, solves the ranking SVM problem in its primal form using gradient-based optimization.
2) `Relative Attributes` [9], also a large "margin" ranking approach but considering "weakly" ordered pairs. The objective function is solved in its primal form using Newton's method.

`Relative Attributes` is very similar to the single-attribute and linear version of the proposed method, and can be viewed as the linear comparison to the kernel solution. Note that all reported results of our method in this section are from the kernel solution.

We shall see that our multiple attributes approach significantly outperforms the two baseline methods where each attribute is learnt individually (as shown in Fig. 5 – 3). Given this, **The second question is:** *(2) Is the good performance due to the multi-attribute learning setting or the nonlinear property (Kernelization) of our approach?*. To investigate this we also test our approach in its single-attribute mode. We can conclude that learning multiple attributes simultaneously is beneficial if our multi-attribute learning approach outperforms its single-attribute version.

3) `Proposed+Active-Single`: our attribute learning method presented in Section 3, but learns the ranking function for each attribute separately.

The third question we address is: *(3) How well our query strategy performs?*. Since the proposed approach to the best of our knowledge is the first attempt at active multi-attribute learning (i.e. that needs to select both a pair of images and an attribute to query), we can **only** compare to random querying. In particular, we explore the following scenarios:

4) Random query selection for all the three ranking models (including our proposed multi-attribute ranking `Proposed` and the two competing techniques, `RankSVM` and `Relative Attributes`), i.e. `Proposed+Random`, `RankSVM+Random` and `Relative Attributes + Random`.
5) Our query selection strategy with the two baselines ranking models, i.e. `RankSVM+Active` and `Relative Attributes + Active`.

The parameters in the two baseline methods, `RankSVM` and `Relative Attributes`, are set as follows. In our experiments the initial ranking hyperplane is random, and the two parameters for the Newton's method are: the maximum number of linear conjugate gradients is set to 20, and the stopping criterion for conjugate gradients is set to 10^{-3}. For our `Proposed+Active` approach, we adopt RBF kernel (the length scale of RBF is selected using cross validation), and simply set the coefficient p to 1 giving equal weights to the local and global significance, i.e. $\mathcal{LS}$ and $\mathcal{GS}$. The penalty constant for the slack variables is set to 1. In our evaluation, the retrieval performance is assessed using a standard ranking accuracy measure – normalized discounted cumulative gain (NDCG), which is defined as

$$\mathrm{NDCG}(\tau_t) = N \sum_i \frac{(2^{r_t(i)} - 1)}{\log_2(1+i)} \tag{18}$$

where i in indicates the rank of an image, $r_t(i)$ is the rating (strength of the presence of attribute t) of image ranked at i, N is a normalization constant such that a perfect rank list receives a NDCG score of value one. We choose NDCG as the accuracy measure since it is more sensitive to the top ranked images.

4.1 Experiment 1 - Structural MRI Scans

4.1.1 Dataset and Experimental Setting

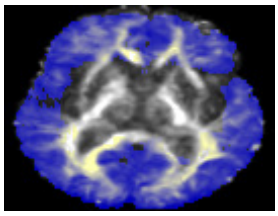


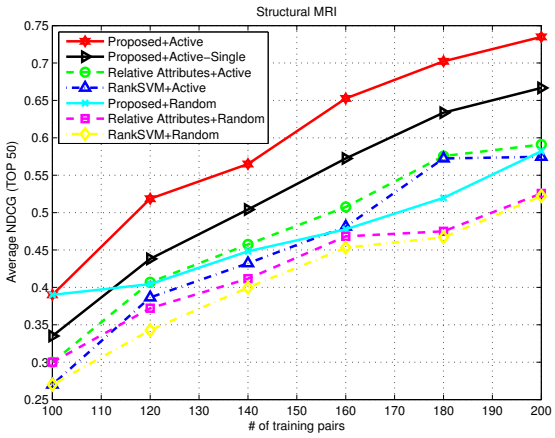
Fig. 2 Example structural MRI scan

Attributes	Relative
Researcher	Ex≺Se≺Ep≺Sp
Manager	Sp≺Se≺Ep≺Ex
Writer	Ex≺Sp≺Ep≺Se

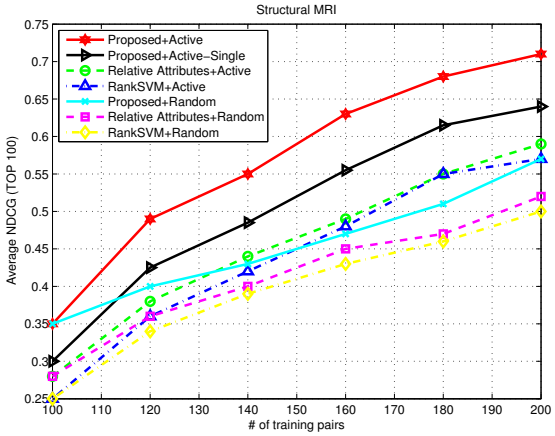
TABLE 2 Relative cognitive measure assignments used on structural MRI scans. The semantic descriptions (relative attributes) of persons are shown in the left column, and the corresponding assignments of cognitive measures are listed in the right column. The four cognitive functioning measures are as follows: (Se) semantic, (Ep) episodic, (Ex) executive and (Sp) Spatial. $\prec$ denotes less important than, which in practice means ranking score less than.

The structural MRI scans used in our experiment were acquired from real clinic cases of 632 patients, whose cognitive functioning scores (semantic, episodic, executive and spatial, ranged from -2.8258 to 2.5123) were also acquired at the same time using cognitive function tests. This is a *new dataset* and will be made publicly available. The raw MRI scans are in *GRAY* format (T1-weighted), which only reveals structural integrity of the gray matter of a brain. In the raw scans, each voxel has a value between 0 and 1, where 1 indicates that the structural integrity of the neuron cell bodies at that location is perfect, while 0 implies either there are no neuron cell bodies or they are not working. The raw scans are preprocessed (including normalization, denoising and alignment) and then restructure to 3D matrices with a size of $134\times102\times134$. To further reduce the dimension of data, we divide the brain into a set of biological regions (46 regions in total, such as Hippocampal, Cerebellum, Brodmann, Fornix, ...) and take the average value of each region to represent the entire region. An example structural MRI scan and the gray matter regions (marked by blue) are shown in Fig. 2. Compared with measuring a person's cognitive functioning, people tends to describe a person using more natural language, such as "he is a research type person". For this purpose, we first learn four ranking functions to predict the four cognitive measures of a person, and then based on the predicted cognitive scores (by the four learnt ranking functions) we assign each person with an understandable attribute based on the rules listed in TABLE 2. The ground truth pairwise orderings (to generate the pairwise constraints) and binary ratings (to calculate the NDCG scores) are generated from the ground truth cognitive scores. The three relative attributes of persons and the corresponding cognitive measure assignments are listed in TABLE2. For the scans whose cognitive scores cannot fit into any of the three cases in TABLE 2, we simply assign them with the closest relative attributes.

4.1.2 Result and Discussion



(a) Structural MRI Top 50



(b) Structural MRI Top 100

Fig. 3 Performance comparison on structural MRI scans (accuracy measured by average NDCG score)

The initial training set contains 100 pairs of structural MRI scans and another 100 pairs are gradually added in. The resulting NDCGs are averaged over both 30 random trials and the 3 relative attributes, where Fig. 3(a) shows

the accuracy at rank 50 and Fig. 3(b) shows the performance at rank 100. Firstly, we can observe that while `RankSVM` and `Relative Attributes` achieve nearly identical ranking accuracy, the proposed multi-attribute approach performs significantly better, even considerably better than its single-attribute version which is also nonlinear. A plausible explanation is that by exploiting the relatedness amongst the four cognitive measures (they are highly correlated in this case), a training pair on one attribute could be concurrently utilized by other attributes. This indicates multiple attributes learning is especially useful when attributes are highly correlated. Besides, we can see that the performance of the methods using random querying is significantly weaker than our query strategy. This reveals the effectiveness of our leveraged active query selection scheme.

4.2 Experiment 2 - Shoes

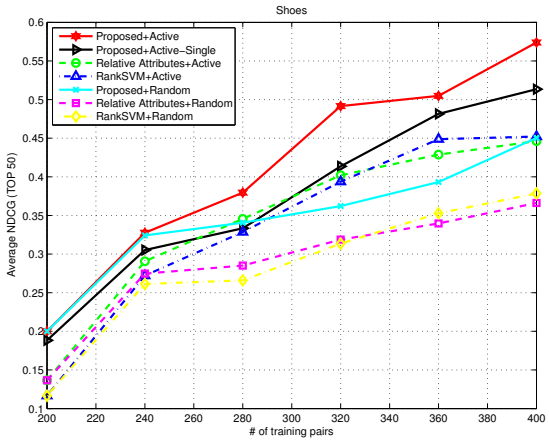
4.2.1 Dataset and Experimental Setting

Relative Attributes	Strengths of Categorical Attributes
pointy-at-the-front	SN≺A≺C≺R≺F≺B≺W≺ST≺P≺H
open	R≺B≺A≺SN≺F≺P≺H≺C≺ST≺W
bright-in-color	B≺C≺P≺H≺W≺A≺SN≺F≺ST≺R
covered-with-ornaments	R≺W≺SN≺A≺F≺C≺P≺H≺B≺ST
shiny	SN≺A≺F≺C≺P≺H≺W≺R≺B≺ST
high-at-the-heel	F≺SN≺R≺A≺C≺B≺W≺P≺H≺ST
long-on-the-leg	W≺C≺F≺ST≺P≺H≺A≺SN≺B≺R
formal	R≺SN≺A≺C≺ST≺B≺F≺P≺H≺W
sporty	ST≺W≺P≺H≺B≺C≺F≺R≺SN≺A
feminine	A≺SN≺R≺C≺F≺B≺W≺ST≺P≺H

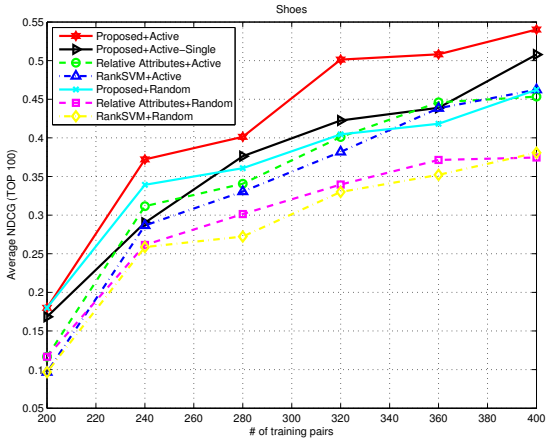
TABLE 3 Relative attribute assignments used on Shoes dataset. The relative attributes are listed in the left column, and the corresponding strength assignments of categorical attributes are shown in the right column. The ten categorical attributes are as follows: (A) athletic shoes, (B) boots, (C) clogs, (F) flats, (H) high heels, (P) pumps, (R) rain boots, (SN) sneakers, (ST) stiletto, and (W) wedding shoes. $\prec$ denotes less important than, which in practice means ranking score less than.

Shoes dataset is part of Attribute Discovery Dataset [49], and is mainly collected from like.com, which is a shopping website that aggregates product data from a wide range of e-commerce sources. In our experiment we use a randomly selected subset of the shoes data, which consists of $3,000$ shoe images in total, and each image is represented using a 990-dimensional vector, including a 960-dimensional Gist descriptor [50] and a 30-dimensional color histogram. The shoes fall into ten categories, i.e., athletic shoes, boots, clogs, flats, high heels, pumps, rain boots, sneakers, stiletto, and wedding shoes. Instead of classifying the shoes into these binary categories, people tend to describe the shoes using more semantic languages, such as "a pair of shiny shoes", "a pair of shoes that are pointy at the front", and "a pair of formal shoes". For this goal, we first learn ten ranking functions simultaneously to retrieve the ten binary categories of shoes, and then based on the produced ranking scores (by the ten learnt ranking functions) to assign each shoe image with a relative attribute using the rules listed in TABLE 3. The ground truth pairwise orderings (to generate the pairwise constraints) and binary ratings (to calculate the NDCG scores) are obtained from [10], and the ten semantic shoe attributes are generated using the same relative attribute assignments as presented in [10]. The ten semantic/relative attributes of shoes and the corresponding categorical strength assignments are listed in TABLE 3. For the shoe images whose categorical attribute strengths cannot fit into any of the ten assignments in TABLE 3, we simply assign them with the closest relative attributes.

4.2.2 Result and Discussion



(a) Shoes Top 50



(b) Shoes Top 100

Fig. 4 Performance comparison on Shoes dataset (accuracy measured by average NDCG score)

We start the training with 200 shoe image pairs, and then we gradually add in another 200 image pairs to the training set using our propose query selection scheme and random selection. The experiment were repeated for 20 times with random initial training pairs. The performance comparison of the six methods on Shoes dataset is reported in Fig. 4(a) (at rank 50) and Fig. 4(b)

(at rank 100), where the result is the average over both the ten relative attributes and the 20 times random trials. We see that, regardless of the query selection strategy, our proposed ranking model significantly outperforms the other two ranking models. This demonstrates that our nonlinear ranking function performs better than the linear methods when handling difficult problems. It also can be observed that the multi-attribute model achieves higher ranking accuracy compared with the single-attribute ones, which indicates that learning related attribute simultaneously is beneficial. With respect to the same ranking model, we can see that the performance of the ones using our proposed query strategy increases faster than the ones using random queries. The result confirms our motivation of learning related attributes together, and validates the effectiveness of our leveraged query selection scheme.

4.3 Experiment 3 - Outdoor Scene Recognition

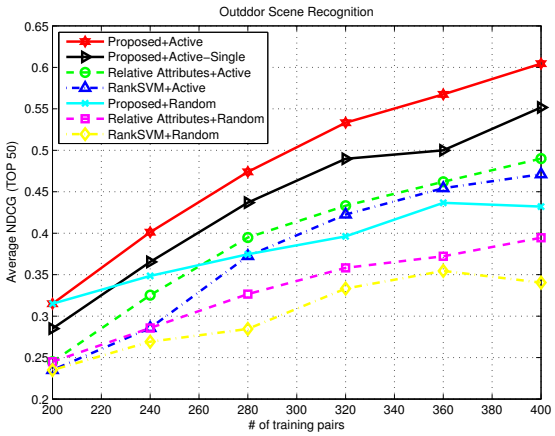
4.3.1 Dataset and Experimental Setting

Relative Attributes	Strengths of Categorical Attributes
natural	T≺I≺S≺H≺C≺O≺M≺F
open	T≺F≺I≺S≺M≺H≺C≺O
perspective	O≺C≺M≺F≺H≺I≺S≺T
large-objects	F≺O≺M≺I≺S≺H≺C≺T
diagonal-plane	F≺O≺M≺C≺I≺S≺H≺T
close-depth	C≺M≺O≺T≺I≺S≺H≺F

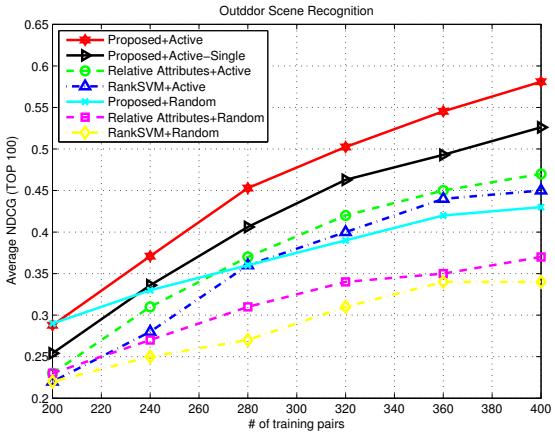
TABLE 4 Relative attribute assignments used in outdoor scene recognition. The relative attributes are listed in the left column, and the corresponding strength assignments of categorical attributes are shown in the right column. The eight categorical attributes are as follows: (C) coast, (M) mountain, (F) forest, (O)open country, (S) street, (I) inside city, (T) tall buildings and (H) highways. $\prec$ denotes less important than, which in practice means ranking score less than.

Outdoor Scene Recognition Dataset [50] consists of 2,688 outdoor scene images (256 $\times$ 256 pixels). Feature extraction is performed on the raw images using 512 dimensional gist descriptor [50]. The images fall in to eight categories, i.e., coast, mountain, forest, open country, street, inside city, tall buildings, and highways. Instead of classifying the images into these binary categories, it would be more natural and meaningful to describe them using semantic language, such as "this is a natural image", "that image is about large-objects". Therefore, the goal here is to learn eight ranking functions for the eight binary categories simultaneously, and then use the produced ranking scores (by the eight learnt ranking functions using Eq.(11)) to assign each image with an relative attribute based on the relative strengths of the eight binary categories. The rule used to generate relative attributes from categorical attributes are listed in TABLE 4. The ground truth pairwise orderings (to generate the pairwise constraints) and binary ratings (to calculate the NDCG scores) are obtained from [9], and the six semantic attributes are generated using the same attribute assignments as presented in [9]. For the images whose categorical attribute strengths cannot fit into any of the six assignments in TABLE 4, we simply assign them with the closest relative attributes.

4.3.2 Result and Discussion



(a) Outdoor Scene Top 50



(b) Outdoor Scene Top 100

Fig. 5 Performance comparison on Outdoor Scene dataset (accuracy measured by average NDCG)

In each trial, for each attribute we randomly select 200 image pairs as the initial training data. In order to evaluate the performance of our active query scheme, we then gradually add 200 additional pairwise constraints to the training set (can be either R_t or S_t) using both the proposed query selection strategy or random querying. The experiment are repeated for 30 times, and the average NDCGs (averaged over both the number of trials and the six relative attributes) are reported in Fig. 5(a) (at rank 50) and Fig. 5(b) (at rank 100). Among the three ranking models, it can be observed that our proposed model significantly outperforms the other two, which demonstrates the effectiveness of our ranking model. Comparing the performance of our proposed approach and its single-attributes version, it can be observed that

our multi-attribute learning achieves higher accuracy, which shows the benefits of simultaneous learning of multiple attributes. We see that, compared to the methods using random querying, as the number of training pairs increases the ones using our active query scheme improve the retrieval accuracy noticeably faster. This validates the efficiency of our query selection strategy and the use of $\mathcal{LS}$ and $\mathcal{GS}$, and confirms the motivation and necessity of active query since asking random questions cannot efficiently improve the performance.

4.4 Experiment 4 - Public Figure Face

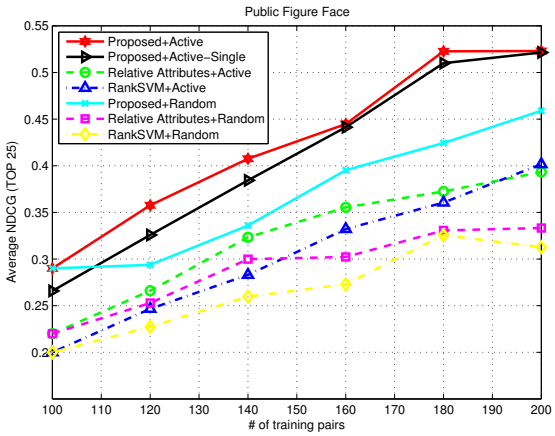
4.4.1 Dataset and Experimental Setting

Relative Attributes	Strengths of Categorical Attributes
masculine-looking	S≺M≺Z≺V≺J≺A≺H≺C
white	A≺C≺H≺Z≺J≺S≺M≺V
young	V≺H≺C≺J≺A≺S≺Z≺M
smile	J≺V≺H≺A≺C≺S≺Z≺M
chubby	V≺J≺H≺C≺Z≺M≺S≺A
visible-forehead	J≺Z≺M≺S≺A≺C≺H≺V
bushy-eyebrows	M≺S≺Z≺V≺H≺A≺C≺J
narrow-eyes	M≺J≺S≺A≺H≺C≺V≺Z
pointy-nose	A≺C≺J≺M≺V≺S≺Z≺H
big-lips	H≺J≺V≺Z≺C≺M≺A≺S
round-face	H≺V≺J≺C≺Z≺A≺S≺M

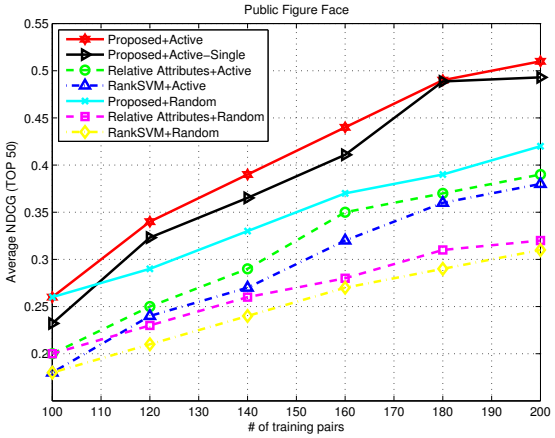
TABLE 5 Relative identity assignments used on public figure face dataset. The relative attributes are generated based on visual similarity of faces. The relative attributes are shown in the left column, and the corresponding similarity assignments of identity of faces are listed in the right column. The eight persons are: (A) Alex Rodriguez, (C) Clive Owen, (H) Hugh Laurie, (J) Jared Leto, (M) Miley Cyrus, (S) Scarlett Johanson, (V) Viggo Mortensen and (Z) Zac Efron. $\prec$ denotes less important than, which in practice means ranking score less than.

PubFig Face Database [26] used in our experiment is a subset of the original data [26]. It contains 772 face images (crop and re-size to 256 $\times$ 256) from eight persons. Each image is represented as a 542-dimensional vector, including gist descriptor and a 45-dimensional Lab color histogram. Rather than classifying the face images by the identity of the eight persons, people are more likely to describe a face using natural language, such as "A young guy with pointy nose", "A smiling girl with big lips". For this goal, we first learn eight ranking functions simultaneously to retrieve the eight persons, and then use the produced ranking scores (by the eight learnt ranking functions) to assign each face image with a relative attribute based on the rules listed in TABLE 5. The ground truth pairwise orderings (to generate the pairwise constraints) and binary ratings (to calculate the NDCG scores) are obtained from [9], and the eleven semantic face attributes are generated using the same attribute assignments as presented in [9]. For the images whose categorical attribute strengths cannot fit into any of the eleven assignments in TABLE 5, we simply assign them with the closest relative attributes.

4.4.2 Result and Discussion



(a) Public Figures Face Top 25



(b) Public Figures Face Top 50

Fig. 6 Performance comparison on PubFig dataset (accuracy measured by average NDCG))

The training starts with 100 pairs and then 100 additional pairwise constraints are added. The average NDCGs (averaged over 30 random trials and the 11 relative attributes) are reported in Fig. 6(a) (at rank 25) and Fig. 6(b) (at rank 50). It can be observed, under the same query scheme, our multi-attribute learning approach provides better retrieval results than the two baselines. Although in this dataset images of the eight persons are not very related, our multi-attribute learning performs on par with its single-attribute version. This shows that our multi-attribute ranking model performs well even when attributes are little related. We can also see that for all the three ranking models, the performance of our query strategy is significantly better than the ones using random querying. This demonstrates the success of our active scheme as it efficiently improves the retrieval accuracy across different learning models.

4.5 Discussion

The fourth question we address is: *(4) How does our method perform when the multiple attributes are little related?*.

To investigate this we test our approach on four different types of image datasets, where the attributes are related at different degrees: Structural MRI dataset in which the attributes are highly related, Shoes [49] and Outdoor Scene datasets [50] in which the attributes are moderately related, and PubFig Face dataset [26] in which the attributes are least related. From the empirical results we see that our multi-attribute method significantly outperforms single-attribute methods when attributes are highly related, and performs on par with the single ones when attributes are little related. The result is consistent with our intuition that the proposed method aims to exploit the dependence amongst attributes, and it performs similarly to a single attribute learning method when there is little dependence to make use of.

Query Selection Runtime. In the experiment, we found that the query selection dominates the total runtime of the proposed algorithm. The training of a ranking function only involves hundreds of pairwise constraints, and can be solved within a second on a regular machine. However, the query selection for active learning requires to go through all possible pairs of images in the dataset, which is quadratic to the number of images. Since most runtime is consumed by the query selection, we only focus on the query selection time in this subsection. Table 6 presents the runtime (averaged over all random trials) of query selection on the four datasets. The reported time is calculated on desktops equipped with two Intel i7 2.67 GHz 64-bit CPUs, and 12 GB memory.

Dataset	Mean Query Selection Runtime
Structural MRI	1.23 seconds
Shoes	25.11 seconds
Outdoor Scene Recognition	20.07 seconds
Public Figure Face	1.75 seconds

TABLE 6 Mean running time of query selection.

From the empirical results shown in Table 6, we see that the runtime of query selection is acceptable. But this runtime would increase quadratically along with the number of images. Therefore, when apply the proposed method to large dataset, hashing based fast query selection (similar to [39]) can be adopted to accelerate the computation. The training time of ranking function is generally not a issue, since the query selection consumes most runtime, especially in active learning where the number of training data is usually limited. But in the cases requiring to further speed up the training procedure, the stochastic gradient descent method (such as [51]) can make the learning of ranking functions faster.

Parameter Selection. To explore the parametrical stability of our method, we evaluate its performance under a series of varying parameter settings. There are two main parameters in our method. (i) λ penalizes the variation vector v_t. A small λ encourages the multiple ranking functions to be dissimilar. (ii) p controls the magnitude of $\mathcal{GS}$. A large p emphasize the $\mathcal{GS}$ in query selection. The experiment is conducted on Outdoor Scene Recognition dataset. In each random trial, we select 200 image pairs as the initial training set, and then add 200 more pairs to it. The experiment is repeated for 30 times, and the averaged performance under different parameter settings is reported in Fig. 7. In the experiment, λ ranges from 0.3 to 10, and p ranges from 0.1 to 3. From the result, we see that the proposed method is not very sensitive to the scale of the parameters λ and p, since there is no sharp jump or drop in the performance as the parameters change. Therefore, when users implement our method in practice, the parameters can be estimated using a rough grid search (possibly cross validation) of parameters.

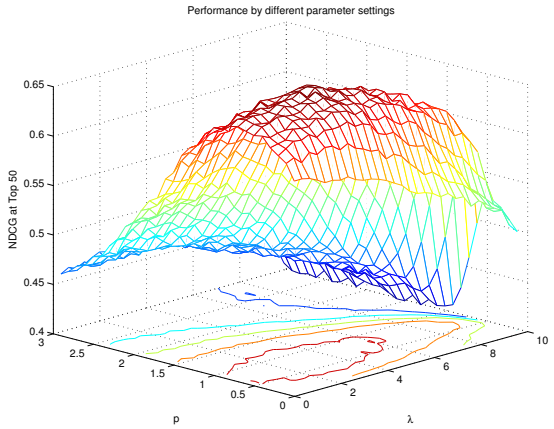


Fig. 7 Performance w.r.t. to different parameter settings on Outdoor Scene Recognition dataset.

5 Conclusion

In this paper we present an active learning paradigm for training multiple relative attributes of images under the supervision of pairwise orderings. The novel contributions of the proposed formulation include discovering the relatedness of multiple attributes to improve the overall retrieval performance, and exploring an active query selection in the setting of learning multiple attributes. The key properties of the proposed approach include allowing multiple attributes to share a common ranking basis, considering both *strongly* and *weakly* ordered pairs of images, and balancing the *local* and *global* significance to make the query selection more stable. The promising results of our empirical study demonstrate the effectiveness of our multiple attributes learning approach in solving a wide range of image retrieval problems including those in neuro-science. our future directions include employing more efficient learning algorithms [52], [53] to cope with large scale applications.

Acknowledgments

The authors gratefully acknowledge support of this research from ONR grants N00014-09-1-0712, N00014-11-1-0108 and NSF Grant NSF IIS-0801528.

REFERENCES

[1] R. Datta, D. Joshi, J. Li, and J. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Computing Surveys (CSUR)*, vol. 40, no. 2, p. 5, 2008.

[2] Y. Chen, J. Z. Wang, and R. Krovetz, "Clue: cluster-based retrieval of images by unsupervised learning." *IEEE Transactions on Image Processing*, vol. 14, no. 8, pp. 1187–1201, 2005.

[3] D. Xu, S. Yan, D. Tao, S. Lin, and H.-J. Zhang, "Marginal fisher analysis and its variants for human gait recognition and content-based image retrieval." *IEEE Transactions on Image Processing*, vol. 16, no. 11, pp. 2811–2821, 2007.

[4] J. He, M. Li, H. Zhang, H. Tong, and C. Zhang, "Generalized manifold-ranking-based image retrieval," *IEEE Transactions on Image Processing*, vol. 15, no. 10, pp. 3170–3177, 2006.

[5] S. Tong and E. Chang, "Support vector machine active learning for image retrieval," in *Proceedings of the ninth ACM international conference on Multimedia*, ser. MULTIMEDIA '01, New York, NY, USA, 2001, pp. 107–118.

[6] J. Li, N. M. Allinson, D. Tao, and X. Li, "Multitraining support vector machine for image retrieval," *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3597–3601, 2006.

[7] W. Jiang, G. Er, Q. Dai, and J. Gu, "Similarity-based online feature selection in content-based image retrieval." *IEEE Transactions on Image Processing*, vol. 15, no. 3, pp. 702–712, 2006.

[8] F. Jing, M. Li, H. Zhang, and B. Zhang, "A unified framework for image retrieval using keyword and visual features," *IEEE Transactions on Image Processing*, vol. 14, no. 7, pp. 979–989, 2005.

[9] D. Parikh and K. Grauman, "Relative attributes," in *ICCV*, 2011, pp. 503–510.

[10] A. Kovashka, D. Parikh, and K. Grauman, "Whittlesearch: Image Search with Relative Attribute Feedback," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2012.

[11] M. Rastegari, A. Diba, D. Parikh, and A. Farhadi, "Multi-attribute queries: To merge or not to merge?" in *CVPR*, 2013, pp. 3310–3317.

[12] A. Parkash and D. Parikh, "Attributes for classifier feedback," in *Proceedings of the 12th European Conference on Computer Vision - Volume Part III*, ser. ECCV'12. Berlin, Heidelberg: Springer-Verlag, 2012, pp. 354–368.

[13] A. Biswas and D. Parikh, "Simultaneous active learning of classifiers & attributes via relative feedback," in *CVPR*, 2013, pp. 644–651.

[14] V. Ferrari and A. Zisserman, "Learning visual attributes," in *Advances in Neural Information Processing Systems 20*, J. Platt, D. Koller, Y. Singer, and S. Roweis, Eds. Cambridge, MA: MIT Press, 2008, pp. 433–440.

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

[16] Y. Wang and G. Mori, "A discriminative latent model of object classes and attributes," in *Proceedings of the 11th European conference on Computer vision: Part V*, ser. ECCV'10, Berlin, Heidelberg, 2010, pp. 155–168.

[17] A. Farhadi, I. Endres, D. Hoiem, and D. A. Forsyth, "Describing objects by their attributes," in *CVPR*, 2009, pp. 1778–1785.

[18] G. Wang and D. A. Forsyth, "Joint learning of visual attributes, object classes and visual saliency." in *ICCV*. IEEE, 2009, pp. 537–544.

[19] A. Farhadi, I. Endres, and D. Hoiem, "Attribute-centric recognition for cross-category generalization," in *CVPR*, 2010, pp. 2352–2359.

[20] S. Chang, G. Qi, J. Tang, Q. Tian, Y. Rui, and T. S. Huang, "Multimedia LEGO: learning structured model by probabilistic logic ontology tree," in *2013 IEEE 13th International Conference on Data Mining, Dallas, TX, USA, December 7-10, 2013*, 2013, pp. 979–984.

[21] A. Farhadi, S. M. M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. A. Forsyth, "Every picture tells a story: Generating sentences from images," in *ECCV (4)*, 2010, pp. 15–29.

[22] T. Kamishima, H. Kazawa, and S. Akaho, "A survey and empirical comparison of object ranking methods," in *Preference Learning*. Springer-Verlag, 2010, pp. 181–201.

[23] M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych, and B. Schiele, "What helps where - and why? semantic relatedness for knowledge transfer." in *CVPR*. IEEE, 2010, pp. 910–917.

[24] J. Wang, K. Markert, and M. Everingham, "Learning models for object recognition from natural language descriptions," in *Proceedings of the British Machine Vision Conference*. BMVA Press, 2009, pp. 2.1–2.11.

[25] T. L. Berg, A. C. Berg, and J. Shih, "Automatic attribute discovery and characterization from noisy web data," in *ECCV (1)*, 2010, pp. 663–676.

[26] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and Simile Classifiers for Face Verification," in *IEEE International Conference on Computer Vision (ICCV)*, Oct 2009.

[27] G. Wang, D. Forsyth, and D. Hoiem, "Comparative object similarity for improved recognition with few or no examples," in *CVPR*. IEEE, 2010, pp. 3525–3532.

[28] A. Shrivastava, S. Singh, and A. Gupta, "Constrained semi-supervised learning using attributes and comparative attributes," in *Proceedings of the 12th European conference on Computer Vision*, ser. ECCV'12, 2012, pp. 369–383.

[29] B. Long, O. Chapelle, Y. Zhang, Y. Chang, Z. Zheng, and B. Tseng, "Active learning for ranking through expected loss optimization," in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '10, 2010.

[30] N. Ailon, "An active learning algorithm for ranking from pairwise preferences with an almost optimal query complexity," *JMLR*, vol. 13, pp. 137–164, Mar. 2012.

[31] A. Kapoor, K. Grauman, R. Urtasun, and T. Darrell, "Gaussian processes for object categorization," *Int. J. Comput. Vision*, vol. 88, no. 2, pp. 169–188, Jun. 2010.

[32] A. Holub, P. Perona, and M. C. Burl, "Entropy-based active learning for object recognition," in *Second IEEE Workshop on Online Learning for Classification. Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8.

[33] S. Branson, C. Wah, B. Babenko, F. Schroff, P. Welinder, P. Perona, and S. Belongie, "Visual recognition with humans in the loop," in *Proceedings of the European Conference on Computer Vision*, Sept. 2010.

[34] Y. Rui and T. S. Huang, "A novel relevance feedback technique in image retrieval," in *Proceedings of the seventh ACM international conference on Multimedia (Part 2)*, ser. MULTIMEDIA '99. New York, NY, USA: ACM, 1999, pp. 67–70.

[35] D. Parikh and K. Grauman, "Interactively building a discriminative vocabulary of nameable attributes." in *CVPR*. IEEE, 2011, pp. 1681–1688.

[36] A. Biswas and D. Parikh, "Simultaneous active learning of classifiers & attributes via relative feedback," *2013 IEEE Conference on Computer Vision and Pattern Recognition*, vol. 0, pp. 644–651, 2013.

[37] B. Qian, H. Li, J. Wang, X. Wang, and I. Davidson, "Active learning to rank using pairwise supervision," in *SDM*, 2013, pp. 297–305.

[38] B. Qian, X. Wang, F. Wang, H. Li, J. Ye, and I. Davidson, "Active learning from relative queries," in *IJCAI*, 2013.

[39] B. Qian, X. Wang, J. Wang, H. Li, N. Cao, W. Zhi, and I. Davidson, "Fast pairwise query selection for large-scale active learning to rank," in *ICDM*, 2013, pp. 607–616.

[40] T. Evgeniou and M. Pontil, "Regularized multi–task learning," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. KDD '04. New York, NY, USA: ACM, 2004, pp. 109–117. [Online]. Available: http://doi.acm.org/10.1145/1014052.1014067

[41] T. Joachims, "Optimizing search engines using clickthrough data," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. KDD '02, 2002, pp. 133–142.

[42] ——, "Making large-scale SVM learning practical," in *Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf, C. Burges, and A. Smola, Eds. Cambridge, MA: MIT Press, 1999, ch. 11, pp. 169–184.

[43] O. Chapelle, "Training a support vector machine in the primal," *Neural Comput.*, vol. 19, no. 5, pp. 1155–1178, May 2007. [Online]. Available: http://dx.doi.org/10.1162/neco.2007.19.5.1155

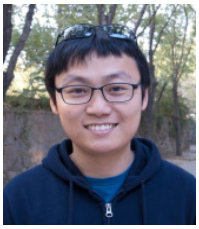
[44] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.

[45] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 1–27, 2011.

[46] I. CVX Research, "CVX: Matlab software for disciplined convex programming, version 2.0," aug 2012.

[47] M. Grant and S. Boyd, "Graph implementations for nonsmooth convex programs," in *Recent Advances in Learning and Control*, ser. Lecture Notes in Control and Information Sciences, V. Blondel, S. Boyd, and H. Kimura, Eds. Springer-Verlag Limited, 2008, pp. 95–110.
[48] O. Chapelle and S. S. Keerthi, "Efficient algorithms for ranking with svms," *Inf. Retr.*, vol. 13, no. 3, pp. 201–215, 2010.
[49] T. L. Berg, A. C. Berg, and J. Shih, "Automatic attribute discovery and characterization from noisy web data," in *Proceedings of the 11th European Conference on Computer Vision: Part I*, ser. ECCV'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 663–676. [Online]. Available: http://dl.acm.org/citation.cfm?id=1886063.1886114
[50] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vision*, vol. 42, no. 3, pp. 145–175, May 2001. [Online]. Available: http://dx.doi.org/10.1023/A:1011139631724
[51] D. Sculley and G. Inc, "Large scale learning to rank," in *In NIPS 2009 Workshop on Advances in Ranking*, 2009.
[52] D. Sculley, "Large scale learning to rank," in *NIPS Workshop on Advances in Ranking*, 2009, pp. 1–6.
[53] J. Weston, S. Bengio, and N. Usunier, "Large scale image annotation: learning to rank with joint word-image embeddings," *Machine learning*, vol. 81, no. 1, pp. 21–35, 2010.

Buyue Qian is currently a research scientist at IBM T. J. Watson Research. He received his PhD in 2013 from Computer Science Department, University of California at Davis. Before that, he received Master of Science (2009) from Columbia University, and BS in Information Engineering (2007) from Xi'an Jiaotong University. The major awards he received include Yahoo! Research Award, IBM Eminence & Excellence Award, and SIAM Data Mining'13 Best Research Paper Runner Up.



Xiang Wang is currently a research scientist at IBM T. J. Watson Research. He received his PhD in 2013 from Computer Science Department, University of California at Davis. Before that, he received Master of Software Engineering (2008) and BS in Mathematics (2004) from Tsinghua University. The major awards he received include SIAM Data Mining'13 Best Research Paper Runner Up, IBM Invention Achievement Award, and UC Davis Best Graduate Researcher Award.

Nan Cao is a research staff member at IBM T. J. Watson Research Center and was a Microsoft MVP (Most Valuable Professional). He graduated from the Hong Kong University of Science and Technology (HKUST) at Aug 2012. Before that, he was a Staff Researcher at IBM China Research Lab from May 2006 to Jan 2010. He joined IBM in the programme of IBM Extreme Blue and led many projects related to graph and text visualizations in domains of social media data analysis and healthcare data analysis. Nan's primary expertise and research interests are information visualization and visual analysis. He is specialized in graph visualization and layout design, visual analysis of networks, social media data, multidimensional data, and text documents.

Yu-Gang Jiang received the Ph.D. degree in Computer Science from the City University of Hong Kong, Kowloon, Hong Kong, in 2009. During 2008-2011, he was with the Department of Electrical Engineering, Columbia University, New York. He is currently an Associate Professor of Computer Science with Fudan University, Shanghai, China. His research interests include multimedia retrieval and computer vision. Dr. Jiang is an active participant of the Annual U.S. NIST TRECVID Evaluation and has designed a few top-performing video analytic systems over the years. He is also an organizer of the annual THUMOS action recognition challenge and the Violent Scenes Detection Task of the MediaEval benchmark. He has served on the committees of numerous conferences and will serve as a Program Chair for ACM ICMR 2015. His work has led to several awards including the 2013 ACM Shanghai Distinguished Young Scientist Award.

Ian Davidson grew up in Australia and obtained his Ph.D. at Monash University, Melbourne. He joined the University of California Davis in 2007 and is now a Professor of Computer Science. He directs the Knowledge Discovery and Data Mining Lab and his work is supported by grants from NSF, ONR, OSD Google and Yahoo!. He is a senior member of the IEEE.