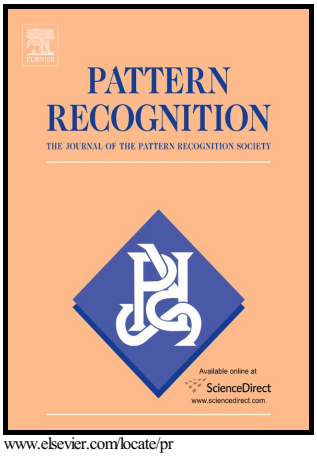
What is the Best Way for Extracting Meaningful Attributes from Pictures?

Liangchen Liu, Arnold Wiliem, Shaokang Chen, Brian C. Lovell



www.elsevier.com/locate/pr

PII: S0031-3203(16)30349-1
DOI: http://dx.doi.org/10.1016/j.patcog.2016.10.034
Reference: PR5942

To appear in: *Pattern Recognition*

Received date: 8 April 2016
Revised date: 17 October 2016
Accepted date: 31 October 2016

Cite this article as: Liangchen Liu, Arnold Wiliem, Shaokang Chen and Brian C Lovell, What is the Best Way for Extracting Meaningful Attributes from Pictures?, *Pattern Recognition,* http://dx.doi.org/10.1016/j.patcog.2016.10.034

What is the Best Way for Extracting Meaningful Attributes from Pictures?

Liangchen Liu, Arnold Wiliem, Shaokang Chen, Brian C. Lovell

The University of Queensland, School of ITEE
QLD 4072, Australia

Abstract

Automatic attribute discovery methods have gained in popularity to extract sets of visual attributes from images or videos for various tasks. Despite their good performance in some classification tasks, it is difficult to evaluate whether the attributes discovered by these methods are meaningful and which methods are the most appropriate to discover attributes for visual descriptions. In its simplest form, such an evaluation can be performed by manually verifying whether there is any consistent identifiable visual concept distinguishing between positive and negative exemplars labelled by an attribute. This manual checking is tedious, expensive and labour intensive. In addition, comparisons between different methods could also be problematic as it is not clear how one could quantitatively decide which attribute is more meaningful than the others. In this paper, we propose a novel attribute meaningfulness metric to address this challenging problem. With this metric, automatic quantitative evaluation can be performed on the attribute sets; thus, reducing the enormous effort to perform manual evaluation. The proposed metric is applied to some recent automatic attribute discovery and hashing methods on four attribute-labelled datasets. To further validate the efficacy of the proposed method, we conducted a user study. In addition, we also compared our metric with a semi-supervised attribute discover method using the mixture of probabilistic PCA. In our evaluation, we gleaned several insights that could be beneficial in developing new automatic attribute discovery methods.

Email addresses: l.liu9@uq.edu.au (Liangchen Liu), a.wiliem@uq.edu.au (Arnold Wiliem), shaokangchenuq@gmail.com (Shaokang Chen), lovell@itee.uq.edu.au (Brian C. Lovell)

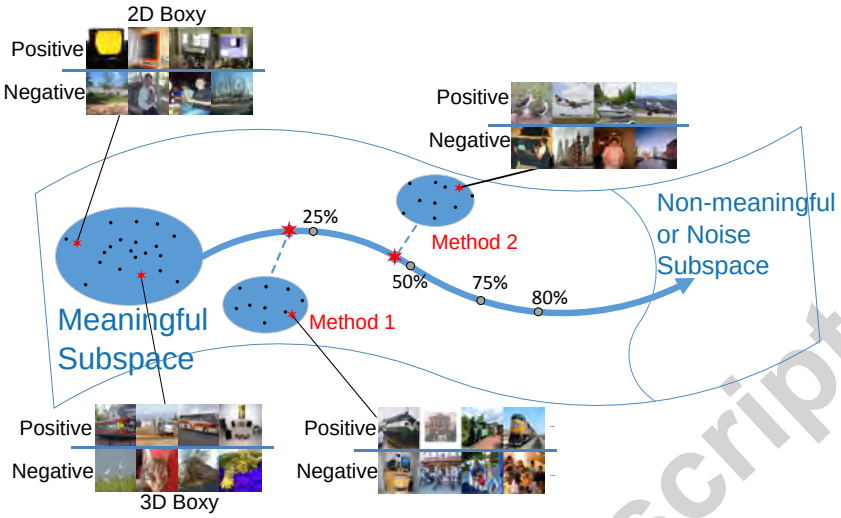


Figure 1: An illustration of the proposed attribute meaningfulness metric. Each individual attribute is represented as the outcome of the corresponding attribute classifier tested on a set of images. Inspired by [31] we propose an approach to measure the distance between a set of discovered attributes and the meaningful subspace. The metric score is derived using a subspace interpolation between Meaningful Subspace and Non-Meaningful/Noise Subspace. The score indicates how many meaningful attributes are contained in the set of discovered attributes.

Keywords: Visual Attribute, Meaningfulness Metric, Attribute Discovering, Semantic Content

1. Introduction

Language is one of the most important factors in communication. We would not have been able to write this paper if there was not any language! Human language has been used for solving computer vision problems such as scene understanding [32] and image or video description[24, 34] and image retrieval [27]. The language model helps us to make an effective transfer of domain expert knowledge into machines. People often say that "a picture is worth a thousand words." Turning this around we can also say that "a thousand words/visual attributes are required to explain a picture." The latter form can be quite powerful to address many computer vision problems [11, 20, 12, 8, 29]. For instance the active learning framework proposed in [19] employs human knowledge to learn better models. Visual attributes are extremely useful as they are: (1) human understandable; (2) machine computable and (3) shared across classes. For these reasons, recently

many attribute discovery methods have been proposed to extract visual/image attributes [3, 33, 18]. Broadly, several more concrete visual attributes such as color and texture have attracted great attentions [50, 9]. To that end, color saliency analysis [50] and shearlets-based texture learning [9] are proposed and have achieved promising performance.

One of the biggest challenges in using attribute descriptors is that an enormous amount of training images with attribute labels is required to train attribute classifiers. It is extremely tedious, time-consuming and expensive to label each individual image for every attribute (*e.g.,* if there are 64 attributes, then each image should have at least 64 attribute labels). Furthermore, in some specialized domains such as *Ornithology* [45], *Entomology* [43] and cell pathology [46], the human labelling task could be immensely expensive as only a few highly trained experts could conduct such a task.

To reduce the workload, automatic attribute discovery methods have been developed [3, 33, 37, 46, 48, 49]. The primary aim of these works is to learn a function that maps the original image feature space into a binary code space wherein each individual bit represents the presence/absence of a visual attribute. These attribute discovery methods are also closely related to hashing methods [14, 22, 44]. The difference is that unlike automatic attribute discovery approaches, hashing methods focus more on how to significantly reduce the storage demand and computational complexity whilst maintaining system accuracy.

Despite great strides that have been made in this field, there are still some important open questions: left unaddressed: 1) Given the set of attributes/binary codes discovered by a method, are these attributes or binary codes really meaningful? 2) Can we compare these methods by directly observing the discovered attributes? By exploring these questions, we can begin to glean some insights on mechanisms required to extract meaningful attributes/binary codes. We note that the aim of this work is not to propose a new method to discover attributes. Instead, we propose a novel meaningfulness metric and use this tool to study the existing methods.

Gauging "how meaningful" for a given attribute can be an ill-posed problem as there is no obvious *yardstick* for measuring this. Fortunately, it is pointed out by Parikh and Grauman that meaningful attributes may have a *shared structure* [30, 31]. This means, given the attribute feature space, meaningful attributes are likely to be close to each other within a subspace. In [25], we further studied this shared structure and applied our findings to the task of automatic generation of surveillance video descriptions.

Inspired by these works, we propose a novel meaningfulness metric that could

become one of the *yardsticks* to measure attribute meaningfulness. More specifically, we first measure the distance between the discovered attributes and the meaningful attribute subspace. To this end, an approximate geodesic distance based on reconstruction error is proposed. As it may be difficult to perform quantitatively analysis/study using this distance directly, we then derive the meaningfulness metric based on the distance. In particular, the metric is derived by first performing a subspace interpolation between meaningful subspace and non-meaningful subspace lying on the manifold of decision boundaries. The distance on each interpolated subspace is calculated. These are then used to calibrate the distance of the discovered attributes to the meaningful subspace. Fig. 1 illustrates our main idea.

An earlier form of this work has been presented in [20]. In this work, we extend our earlier work in several aspects. We perform in-depth analysis on the proposed metric and consider additional dataset. Throughout these experiments we found that the calibration step heavily depends on the meaningful subspace spanned by the selected meaningful attributes, denoted the subspace bases. The space spanned by these bases should be maximized in order to ensure the calibration is done properly. To remedy this, we present a simple-yet-effective technique using semantic reasoning and threshold setting.

Contributions — We list our contributions as follows:

- We propose a reconstruction error based approach with two different regularizations (*i.e.,* ℓ_0 and convex hull) to approximate the geodesic distance between a given attribute set and the meaningful subspace.
- We propose the novel *attribute meaningfulness* metric that allows us to quantitatively measure the meaningfulness of a set of automatically discovered attributes. The metric score is related to "the percentage of meaningful attributes contained in the set of attributes."
- We propose an improved calibration method to avoid pathological cases where the calibration could not be performed. This method is developed based on the in-depth analysis performed in this work.
- We present extensive experiments and analysis on four popular attribute-labelled datasets to demonstrate that our proposal can really capture attribute meaningfulness. The *attribute meaningfulness* of some recent automatic attribute discovery methods and various hashing approaches are also evaluated on these datasets. A user study is conducted to further validate the effectiveness of the proposed metric. In addition, we compare the proposed metric with a metric adapted from a recent semi-supervised attribute

discovery method using the Mixture of Probabilistic Principal Component Analysis (MPPCA) [31, 39].

We continue our paper as follows. Related work is discussed in Section 2. Then we introduce our approach of evaluating attribute meaningfulness in Section 3. Our proposed metric is described in Section 4. Next we discuss the experiments and results in Section 5. Finally the main findings and future directions are presented in Section 6.

2. Related Work

Evaluation of visual attribute meaningfulness is traditionally conducted by manually checking the presence/absence of consistent identifiable visual concepts in a set of given images. This task usually requires a large-scale human labelling effort. A system such as the Amazon Mechanical Turk (AMT) [1] is able to handle this task for small datasets. However, since this process needs to be repeated whenever new attributes are discovered or novel methods are proposed, this manual process is ineffective and expensive. In our case, the AMT Human Intelligence Task (HIT) is to evaluate the meaningfulness of attributes by examining corresponding positive and negative images according to each attribute. The average time of each worker spent on this typical HIT is 2 minutes [32]. Then an AMT worker may require 320 minutes to evaluate 32 attributes discovered by 5 different methods (*i.e.,* $32 \times 5 \times 2 = 320$ minutes). The time spent could increase significantly if statistically reliable results are desired by increasing the number of AMT workers.

Unequivocally, it is more desirable to develop an automatic approach, which is more cost-effective, less labor intensive and time consuming to evaluate the meaningfulness of the set of discovered attributes. The task of measuring the *attribute meaningfulness* of discovered attributes is similar to the task presented in the Turing Test [40]. In this task, we would like to measure how much a machine could provide responses like a human being. If a machine could respond like a human being, it means that the results produced must have significant meaning. Unfortunately, the Turing Test still requires a human judge who actively engages with the machine.

To that end, several works [1, 35, 40] aim to devise an automated Turing Test that follows the framework of this famous test but replaces the human judge by

[1] www.mturk.com

another machine. A notable example is CAPTCHA [42] which is very prevalent in web security applications. This technique basically lets a machine be the judge issuing the test to determine whether the subject is a human. Generally, CAPTCHA provides a challenge in the form of an image containing numbers or characters which are difficult to be identified by current machines. The main assumption in CAPTCHA is that machine recognition will not be as good as human.

Our work can be interpreted as an instance of the automated Turing Test as follows. We are testing a set of automatic attribute discovery techniques by giving challenges in the form of images. These techniques are then giving us a set of attributes. We will automatically verify the meaningfulness through the positive and negative images generated from each attribute classifier. Note that if we have human observers performing the verification instead of machines, then this becomes an instance of the standard Turing Test. To perform the automated Turing Test, there has to be a measurement to determine which automatic attribute discovery technique is 'good' and which one is 'not that good'.

Some unsupervised semantic visual representation learning works [17, 16, 33, 46] have indicated that it is possible to discover the meaningful concepts unsupervisedly from data itself with or without side information. Such as Chen *et al.* [17] introduce a simple yet powerful unsupervised approach to learn and predict visual attributes directly from data. With the help of deep Convolutional Neural Networks (CNNs), they train to output a set of discriminative, binary attributes often with semantic meanings. Hong *et al.* [16] propose a novel algorithm to cluster and annotate a set of input images with semantic concepts jointly. They employ non-negative matrix factorization with sparsity and orthogonality constraints to learn the label-based representations with the side information (a labeled reference image database) obtaining promising results.

All of these works imply there may be some potential relations between meaningful concepts. Fortunately, the shared structure assumption among meaningful attributes proposed in [31] can serve as the foundation of the automatic measurement. Based on this assumption, Parikh and Grauman *et al.* proposed an active learning approach that uses Mixtures of Probabilistic Principal Component Analysers (MPPCA) [39] to predict how likely an attribute is nameable. Nevertheless, their work only focuses on deciding whether an attribute is nameable or not. Their work does not tackle the problem of quantitatively measuring the attribute meaningfulness. In addition, this approach requires human interaction to populate the nameability space. Thus, their method is not suitable for addressing our goal (i.e., to automatically evaluate the meaningfulness of attribute sets).

In our previous work, the shared structure assumption is utilized [25]. In

particular, the work in [25] proposed a selection approach of attribute discovery methods to assist attribute-based keywords generation for video description from surveillance systems. However, the work did not consider quantitative analysis of the meaningfulness of the discovered attributes (*e.g.,* how much meaningful content is contained in a set of automatically discovered attributes). In addition, the characteristics of the meaningfulness of attributes may vary to some extent.

3. Measuring Attribute Set Meaningfulness

We begin by describing the manifold of decision boundaries and the meaningful attribute subspace. Then, we define the distance between the automatically discovered attributes and the meaningful attribute set in the manifold space to measure the attribute meaningfulness.

3.1. Manifold of decision boundaries

Supposed there is a set of exemplars $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^{N}$, an attribute can be considered as a decision boundary which partitions the set into two subsets $\mathcal{X}^+ \cup \mathcal{X}^- = \mathcal{X}$. Here $\mathcal{X}^+$ represents the set where the attribute is present and $\mathcal{X}^-$ represents the set where the attribute is absent. Therefore, all the attributes are lying on a manifold formed by decision boundaries [31].

In this case, an attribute can also be viewed as an N-dimensional binary vector whose element represents the classification output of all exemplars $\boldsymbol{x}_i$ classified by the corresponding attribute binary classifier $\phi(\cdot) \in \mathbb{R}$. The sign of the classifier output on $\boldsymbol{x}_i$ indicates whether the exemplar belongs to the positive or negative set (*i.e.,* $\mathcal{X}^+$ or $\mathcal{X}^-$). As such, an attribute can be represented as $\boldsymbol{z}^{[\mathcal{X}]} \in \{-1, +1\}^N$ whose i-th element is $\boldsymbol{z}_{(i)}^{[\mathcal{X}]} = \text{sign}(\phi(\boldsymbol{x}_i)) \in \{-1, +1\}$. For the sake of simplicity, we drop the symbol $[\mathcal{X}]$ from $\boldsymbol{z}^{[\mathcal{X}]}$ whenever the context is clear. Thus, the manifold of decision boundaries w.r.t. $\mathcal{X}$ can be defined as $\mathcal{M}^{[\mathcal{X}]} \in \{-1, +1\}^N$ which is embedded in a N-dimensional binary space. Again, we also write $\mathcal{M}^{[\mathcal{X}]}$ as $\mathcal{M}$.

As observed from [31, 30], the meaningful attributes have shared structure wherein they lie close to each other on the manifold. In other words all the meaningful attributes are contained in a subspace on $\mathcal{M}$. In an ideal case, all possible meaningful attributes should be enumerated in this subspace. Unfortunately, it is infeasible to enumerate all of them. One intuitive solution is by relying on the existing human knowledge, that is the human labelled attributes from various datasets such as [5, 30, 31]. These attributes are all naturally meaningful since they are collected via manual human labeling process using the AMT. However, the number of available labelled attributes may not be sufficient. To this

end, based on the shared structure assumption, we thus introduce an approximation of the meaningful subspace by linear combinations of the human labelled attributes. This means, if an automatically discovered attribute is close enough to any attribute existing in the meaningful subspace, it should be considered as a meaningful attribute.

3.2. Distance of an attribute to the Meaningful Subspace

In this section, we mathematically describe the reconstruction error based distance of an attribute to the Meaningful Subspace. Given a set of N images $\mathcal{X}$, we denote $\mathcal{S} = \{\boldsymbol{h}_j\}_{j=1}^{J}, \boldsymbol{h}_j \in \{-1, +1\}^N$ as the set of meaningful attributes. We use a matrix $\boldsymbol{A} \in \mathbb{R}^{N \times J}$, in which each column vector is the representation of an attribute, to form the set $\mathcal{S}$. As the assumption in [31], meaningful attributes should be close to the meaningful subspace spanned by the set of meaningful attributes $\mathcal{S}$. For instance, the primary colors *i.e.,* red, green, blue are able to construct the set of secondary colors such as yellow, magenta and cyan. Moreover, the primary colors can provide negative information clues to describe other primary colors (*e.g.,* blue is neither green nor red). Under this assumption, we are able to define a reconstruction error based distance between an attribute and the meaningful subspace. More specifically, let $\boldsymbol{z}_k$ be an attribute and $\boldsymbol{A}$ be the representation of meaningful attributes. The distance is defined as:

$$\min_{\boldsymbol{r}} \|\boldsymbol{A}\boldsymbol{r} - \boldsymbol{z}_k\|_2^2, \tag{1}$$

where $\boldsymbol{r} \in \mathbb{R}^{J \times 1}$ is the reconstruction coefficient vector. Note that the reconstruction in (1) may not be in the manifold $\mathcal{M}$ (i.e. $\boldsymbol{A}\boldsymbol{r} \notin \mathcal{M}$). Therefore, we relax this reconstruction procedure into Euclidean space for computational simplicity. This relaxation effectively becomes an approximation of the true geodesic distance.

3.3. Distance between a set of discovered attributes and the Meaningful Subspace

Analogously, suppose there are K discovered attributes, we use matrix $\boldsymbol{B} \in \{0, 1\}^{N \times K}$ to represent the discovered attribute set $\mathcal{D}$. Then, according to the specific set of images $\mathcal{X}$, we can define the distance between the set of discovered attributes $\mathcal{D}$ and the Meaningful Subspace $\mathcal{S}$ as the average reconstruction error:

$$\delta(\mathcal{D}, \mathcal{S}; \mathcal{X}) = \frac{1}{K} \min_{\boldsymbol{R}} \|\boldsymbol{A}\boldsymbol{R} - \boldsymbol{B}\|_F^2, \tag{2}$$

where $\|\cdot\|_F$ and $\boldsymbol{R} \in \mathbb{R}^{J \times K}$ are the Frobenious norm and the reconstruction matrix respectively.

The reconstruction coefficients are preferably sparse, because generally only a few attributes can provide useful clues to reconstruct a particular attribute while most of them should stay inactive in this procedure. Similar to the example in section 3.2, only a few color attributes can reconstruct another color attribute, most of them should stay inactive (*i.e.,* their reconstruction coefficient should be 0). Unfortunately, the distances in (1) and (2) may create dense reconstruction coefficients due to the absence of a regularization term. As such, we first introduce convex hull regularization used in [25]. Moreover, according to [36], the perception mechanism of human visual systems follows the sparsity principle. That means only a few attributes will first trigger the semantic-visual connection in our brain. Desirable attribute discovery methods should also obey this principle. Hereby, we consider the sparsity-inducing ℓ_0 regularization as the second regularization alternative.

3.3.1. Convex hull regularization

Via introducing a convex hull constraint, (2) becomes:

$$\begin{aligned}\delta_{\text{cvx}}(\mathcal{D},\mathcal{S};\mathcal{X}) = \frac{1}{K}\min_{\boldsymbol{R}} \|\boldsymbol{A}\boldsymbol{R}-\boldsymbol{B}\|_F^2 \,\text{s.t.}\\ \boldsymbol{R}(i,j)\geq 0\\ \sum_{i=1}^{J}\boldsymbol{R}(i,\cdot)=1.\end{aligned} \tag{3}$$

This objective function describes the average distance between each discovered attribute $\boldsymbol{z}_k \in \mathcal{D}$ and the convex hull of $\mathcal{S}$. Its optimization can be efficiently solved using the method proposed in [7].

3.3.2. ℓ_0 regularization

As to ℓ_0 regularization, different from the convex hull regularization, a possible direct correlation between each discovered attribute $\boldsymbol{z}_k \in \mathcal{D}$ and the meaningful attribute, $\boldsymbol{h}_j \in \mathcal{S}$ is considered:

$$\begin{aligned}\delta_{\text{jp}}(\mathcal{D},\mathcal{S};\mathcal{X}) =& \frac{1}{K}\min_{\boldsymbol{R}} \|\boldsymbol{A}\boldsymbol{R}-\boldsymbol{B}\|_F^2, \,\text{s.t.}\\ &\forall k\in\{1\cdots K\}, \|\boldsymbol{R}_{\cdot,k}\|_0\leq 1,\\ &\forall j\in\{1\cdots J\}, \|\boldsymbol{R}_{j,\cdot}\|_0\leq 1.\end{aligned} \tag{4}$$

where $\boldsymbol{R}_{j,\cdot}$, $\boldsymbol{R}_{\cdot,k}$ represent the j-th row vector and the k-th column vector in matrix $\boldsymbol{R}$ respectively. The two additional ℓ_0 regularizers enforce one-to-one relationships between $\mathcal{S}$ and $\mathcal{D}$. The reconstruction matrix $\boldsymbol{R}$ correlates each discovered attribute to a particular meaningful attribute. More specifically, for each discovered attribute $\boldsymbol{z}_k \in \mathcal{D}$, the closest $\boldsymbol{h}_j \in \mathcal{S}$ is found to minimize the function. However, it could be possible that $|\mathcal{S}| > |\mathcal{D}|$. In this case, we can only match K discovered attributes in $\mathcal{S}$ and vice versa.

Unfortunately, the optimization for (4) is non-convex. As such, a greedy approach is proposed to address this through iteratively finding pairs of meaningful discovered attributes with the smallest distance. This can be converted into finding the pairs with the highest similarities (lowest distance means highest similarity).

Here we can define the similarities between a discovered attribute $\boldsymbol{z}_k$ and a meaningful attribute $\boldsymbol{h}_j$ in terms of their correlations. Let $\rho(\boldsymbol{z}_k, \boldsymbol{h}_j), \boldsymbol{z}_k \in \mathcal{D}, \boldsymbol{h}_j \in \mathcal{S}$ be the correlation between $\boldsymbol{z}_k$ and $\boldsymbol{h}_j$. Then ρ can be defined as:

$$\rho(\boldsymbol{z}_k, \boldsymbol{h}_j) = \frac{\text{count}(\boldsymbol{z}_k = \boldsymbol{h}_j)}{N}, \tag{5}$$

where count means the operation which counts how many same elements $\boldsymbol{z}_k$ shares with $\boldsymbol{h}_j$.

Thus, the function $\rho(\boldsymbol{z}_k, \boldsymbol{h}_j)$ can be computed from $\boldsymbol{A}_{\cdot,j}$ and $\boldsymbol{B}_{\cdot,k}$, where $\boldsymbol{B}_{\cdot,k}$, $\boldsymbol{A}_{\cdot,j}$ represent the discovered attribute $\boldsymbol{z}_k$ and the meaningful attribute $\boldsymbol{h}_j$ respectively. Denote $\mathcal{P}$ as the set of M pairs of $\boldsymbol{h}_j \in \mathcal{S}$ and $\boldsymbol{z}_k \in \mathcal{D}$ that have the highest correlation, $\mathcal{P} = \{(\boldsymbol{h}_j^1, \boldsymbol{z}_k^1) \cdots (\boldsymbol{h}_j^M, \boldsymbol{z}_k^M)\}$, $\boldsymbol{h}_j^i = \boldsymbol{h}_j^l$ if and only if $i = l$, $\boldsymbol{z}_k^i = \boldsymbol{z}_k^l$.

Therefore the matrix $\boldsymbol{R}^*$ that minimizes (4) is defined, after $\mathcal{P}$ is determined, via:

$$\boldsymbol{R}_{j,k}^* = \begin{cases} 1 \text{ if } (\boldsymbol{h}_j, \boldsymbol{z}_k) \in \mathcal{P} \\ 0 \text{ if } (\boldsymbol{h}_j, \boldsymbol{z}_k) \notin \mathcal{P}. \end{cases} \tag{6}$$

For the given inputs $\mathcal{D} = \{\boldsymbol{z}_k\}_{k=1}^K$, $\mathcal{S} = \{\boldsymbol{h}_j\}_{j=1}^J$ and $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^N$, Algorithm 1 elaborates the procedures of computing the set $\mathcal{P}$. Note that, $(\boldsymbol{h}_j, \cdot)$ and $(\cdot, \boldsymbol{z}_k)$ in step 3 represent all possible pairs containing $\boldsymbol{h}_j$ and $\boldsymbol{z}_k$, respectively.

4. Attribute Set Meaningfulness Metric

Attribute meaningfulness metric is designed to determine which existing automatic attribute discovery method is more likely to discover meaningful attributes. Moreover, it can provide some insights about how to devise new automatic attribute discovery methods.

Algorithm 1 The proposed greedy algorithm to solve (4)

Input: $\mathcal{D} = \{\boldsymbol{z}_k\}_{k=1}^{K}$, $\mathcal{S} = \{\boldsymbol{h}_j\}_{s=1}^{J}$ and $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^{N}$
Output: $\mathcal{P}$ that contains M pairs that have the highest correlation, where $M = \min(K, J)$.

$\mathcal{P} \leftarrow \{\}$
while $|\mathcal{P}| \leq M$ **do**
  Find the highest $\rho(\boldsymbol{h}_j, \boldsymbol{z}_k)$ where $(\boldsymbol{h}_j, \cdot) \notin \mathcal{P}$ and $(\cdot, \boldsymbol{z}_k) \notin \mathcal{P}$.
  $\mathcal{P} = \mathcal{P} \cup (\boldsymbol{h}_j, \boldsymbol{z}_k)$
end while

In this section we will introduce the Attribute Set Meaningfulness Metric. We order our discussion as follows (1) Meaningful subspace interpolation; (2) Selecting meaningful subspace representation and (3) Computing the meaningfulness metric.

By means of the distance functions δ_{jp} and δ_{cvx} described in Section 3.2, we are able to measure how far is the set of discovered attributes $\mathcal{D}$ from the Meaningful Subspace $\mathcal{S}$. The closer the distance, the more meaningful the set of attributes are. However, as the relationship between the proposed distances and meaningfulness could be non-linear, the distance may not be easy to interpret. Furthermore, it is difficult to compare the results between δ_{cvx} and δ_{jp}.

4.1. Attribute meaningful subspace interpolation

Our goal is to obtain a metric that is both easy to interpret and able to perform comparisons between various distance functions. Inspired by [15], we apply the subspace interpolation to generate a set of subspaces between Meaningful Subspace and Non-Meaningful Subspace, or Noise Subspace. Here, we use a set of evenly distributed random attributes to represent the Non-Meaningful Subspace $\mathcal{N}$.

For the purpose of subspace interpolation, the meaningful attribute set $\mathcal{S}$ is divided into two subsets:

$$\mathcal{S}^1 \cup \mathcal{S}^2 = \mathcal{S} \tag{7}$$

where we consider the set $\mathcal{S}^1$ as the representation of the Meaningful Subspace. When gradually adding random attributes $\tilde{\mathcal{N}} \in \mathcal{N}$ into $\mathcal{S}^2$, the interpolated sets of subspaces can be obtained. Here we present the proposition which guarantees that the interpolation is able to generate subspaces between the Meaningful Subspace and the Non-Meaningful Subspace.

Proposition 4.1. *Let $\tilde{\mathcal{S}} = \mathcal{S}^2 \cup \tilde{\mathcal{N}}$; when $\tilde{\mathcal{N}} = \{\}$, the distance δ^* between $\tilde{\mathcal{S}}$ and $\mathcal{S}^1$ (refers to (7)) is minimized. However, when $\tilde{\mathcal{N}} -> \mathcal{N}$, the distance between $\tilde{\mathcal{S}}$ and $\mathcal{S}^1$ is asymptotically close to $\delta^*(\mathcal{S}^1, \mathcal{N}; \mathcal{X})$, where δ^* is the distance function presented previously such as δ_{jp} and δ_{cvx}. More precisely, we can define the relationship as follows:*

$$\lim_{|\tilde{\mathcal{N}}| \to \infty} \delta^*(\tilde{\mathcal{S}}, \mathcal{S}^1; \mathcal{X}) = \delta^*(\mathcal{N}, \mathcal{S}^1; \mathcal{X}). \tag{8}$$

Remarks. Proposition 4.1 basically describes when random attributes are added into $\tilde{\mathcal{S}}$ gradually, the subspace, that is initially close to the Meaningful Subspace $\mathcal{S}^1$, will be more and more distant from $\mathcal{S}^1$. Eventually the subspace will be spanned by random attributes that is asymptotically close to the Non-Meaningful attribute subspace. While it is easy to prove the above Proposition, we present one version of the proof in the appendix.

4.2. Selecting meaningful subspace representation

As discussed in section 3.1, enumerating all the meaningful attributes to represent the meaningful attribute subspace is impossible. We thus use linear combinations of meaningful attributes to approximate the meaningful subspace.

However, the division of the meaningful attributes into two subsets as suggested in Proposition 4.1 will reduce the subspace spanned to represent the meaningful subspace. More specifically, the linear combination of attributes from $\mathcal{S}^1$ may not span the whole meaningful subspace. To remedy this, one should carefully select the attributes to form $\mathcal{S}^1$ that can maximize the space spanned by the representation.

Under our proposed approach which is based on the linear reconstruction, the selected meaningful attributes for $\mathcal{S}^1$ should form the bases of the meaningful subspace. Here, one way to maximize the space spanned is that to select independent bases.

The attribute independence with respect to the others can be evaluated by how well the attribute can be reconstructed from others. In addition, one can evaluate the attribute independence from the attribute semantic names. For instance, textural attributes such as 'metal' may be independent to the other textural attributes such as 'grass', 'wooden'. Therefore, these attributes should be included in the set to represent the meaningful subspace *i.e.,* the set $\mathcal{S}^1$.

In the light of these facts, we propose an approach to perform meaningful subspace representation selection, $\mathcal{S}^1$. First, the attribute semantic descriptions

are considered. Any attributes that are deemed to be independent will be indicated and always put in the set $\mathcal{S}^1$. On the second step, we evaluate the attribute independence by applying either d_{cvx} or d_{jp}. In particular, we use a leave-one-attribute-out scheme which calculates the distance between one attribute to the rest of the attributes. We then set the threshold α. Again, we indicate any attributes having distance more than α and always put them in $\mathcal{S}^1$. The threshold, α will be one of the parameters which will be determined during the experiments.

4.3. Computing the meaningfulness metric

After constructing the meaningful subspace, we can calibrate the attribute set meaningfulness distance by subspace interpolation based on the equivalent distance effect assumption [4]. That is, if the distance of two attribute subspaces to the meaningful subspace are the same, the amount of meaningful contents contained in these two subspaces are the same.

We denote the distance between $\tilde{\mathcal{S}}$ and the Meaningful Subspace $\mathcal{S}^1$ as $\delta^{\tilde{\mathcal{S}}}$ and the distance between $\mathcal{D}$ and the Meaningful Subspace $\mathcal{S}^1$ as $\delta^{\mathcal{D}}$. After subspace interpolation, we find the subspace $\tilde{\mathcal{S}}$ that makes $\delta^{\tilde{\mathcal{S}}} \approx \delta^{\mathcal{D}}$. Using the equivalent distance effect assumption, if $\delta^{\tilde{\mathcal{S}}} \approx \delta^{\mathcal{D}}$, the meaningfulness between $\tilde{\mathcal{S}}$ and $\mathcal{D}$ should be on par with each other. As $\tilde{\mathcal{S}}$ is defined as a set of meaningful attributes added with additional noise attributes, this representation is able to evaluate the meaningfulness of $\mathcal{D}$. We can consider this task as an optimization problem as follows:

$$g^* = \arg\min_{|\tilde{\mathcal{N}}|} \left\| \delta^*(\{\mathcal{S}^2 \cup \tilde{\mathcal{N}}\}, \mathcal{S}^1; \mathcal{X}) - \delta^*(\mathcal{D}, \mathcal{S}^1; \mathcal{X}) \right\|_2^2. \tag{9}$$

where g^* represents how many minimum number of random attributes required to be added into $\tilde{\mathcal{S}}$ to make $\delta^{\tilde{\mathcal{S}}} \approx \delta^{\mathcal{D}}$. The above optimization problem can be interpreted as searching for the furthest subspace $\tilde{\mathcal{S}}$ from the Meaningful Subspace in an open sphere with radius $\delta^{\mathcal{D}}$. The above equation can be simply solved by a curve fitting approach. In our implementation, we apply the least square approach.

Finally, we denote γ as the proposed attribute meaningfulness metric as follows.

$$\gamma(\mathcal{D}; \mathcal{X}, \mathcal{S}) = (1 - \frac{g^*}{|\mathcal{S}^2| + g^*}) \times 100. \tag{10}$$

Remarks. The equation in (10) determines how many noise/Non-Meaningful attributes are required for a set of automatically discovered attributes to have similar distance as $\delta^{\mathcal{D}}$. On the other hand, our proposed metric reflects how many

meaningful attributes are contained in the attribute set. A smaller number of Non-Meaningful attributes indicates a more meaningful attribute set overall.

Since different aspects of meaningfulness may be captured by various distance functions, we combine the metric values calculated using both proposed distance functions. For simplicity, we use a equally weighted summation in this paper: $\tilde{\gamma} = \frac{1}{2}\gamma_{\text{cvx}} + \frac{1}{2}\gamma_{\text{jp}}$, as our final metric.

5. Experiments

In this part, the efficacy of our approach to measure the meaningfulness of a set of attributes will be first evaluated. Then the proposed metric is used to evaluate meaningfulness of the attribute sets generated by various automatic attribute discovery methods such as PiCoDeS [3] and Discriminative Binary Codes (DBC) [33] as well as some recent hashing methods such as Iterative Quantization (ITQ) [14], Spectral Hashing (SPH) [44], Locality Sensitivity Hashing (LSH) [22] and Kernel-Based Supervised Hashing (KSH) [26].

The two proposed metrics γ_{jp} (4), γ_{cvx} (3) and the combined metric $\tilde{\gamma}$ are applied to compare the meaningfulness of the attributes discovered from the comparative methods on four attribute datasets: (1) a-Pascal a-Yahoo dataset (ApAy) [11]; (2) Animal with Attributes dataset (AwA) [21] and; (3) SUN Attribute dataset (ASUN) [32]; (4) Unstructured Social Activity Attribute dataset (USAA) [13]

Finally, our metric will be then compared against a user study and a metric, denoted the MPPCA metric or MPPCA, adapted from semi-supervised attribute discovery method proposed in [31].

5.1. Datasets and experiment setup

a-Pascal a-Yahoo dataset (ApAy) [11] — comprises two sources: a-Pascal and a-Yahoo. There are 12,695 cropped images in a-Pascal that are divided into 6,340 for training and 6,355 for testing with 20 categories. The a-Yahoo set has 12 categories disjoint from the a-Pascal categories. Moreover, it only has 2,644 test exemplars. There are 64 attributes provided for each cropped image. The dataset provides four features for each exemplar: local texture, HOG, edge and color descriptor. We use the training set for discovering attributes and we perform our study on the test set. More precisely, we consider the test set as the set of images $\mathcal{X}$ defined in 3.1.

Animal with Attributes dataset (AwA) [21] — the dataset contains 35,474 images of 50 animal categories with 85 attribute labels. There are six features provided in this dataset: HSV color histogram, SIFT [28], rgSIFT [41], PHOG [6],

SURF [2] and local self-similarity [38]. The AwA dataset is proposed for studying the zero-shot learning problem. As such, the training and test categories are disjoint; there are no training images for test categories and vice versa. More specifically, the dataset contains 40 training categories and 10 test categories. Similar to the ApAy dataset, we use the training set for discovering attributes and we perform the study in the test set.

SUN Attribute dataset (ASUN) [32] — ASUN is a fine-grained scene classification dataset consisting of 717 categories (20 images per category) and 14,340 images in total with 102 attributes. There are four types of features provided in this dataset: (1) GIST; (2) HOG; (3) self-similarity and (4) geometric context color histograms (See [47] for feature and kernel details). From 717 categories, we randomly select 144 categories for discovering attributes. As for our evaluation, we randomly select 1,434 images (*i.e.,* 10% of 14,340 images) from the dataset. It means, in our evaluation, some images may or may not come from the 144 categories used for discovering attributes.

Unstructured Social Activity Attribute dataset (USAA) [13] — USAA is a relatively novel benchmark attribute dataset for social activity video classification and annotation. It is manually annotated with 69 groundtruth attributes from 8 semantic class videos of Columbia Customer Video (CCV) dataset. There are 100 videos per-class for training and testing respectively. The annotated attributes can be divided into 5 broad categories: actions, objects, scenes, sounds, and camera movement. The 8 classes in the dataset are birthday party, graduation party, music performance, non-music performance, parade, wedding ceremony, wedding dance and wedding reception. The SIFT, STIP and MFCC features for all these videos are extracted in the dataset.

For each experiment, we apply the following pre-processing step described in [3]. We first lift each feature into a higher-dimensional space which is three times larger than the original space. After the features are lifted, we then apply PCA to reduce the dimensionality of the feature space by 40 percent. This pre-processing step is crucial for PiCoDeS as it uses lifted feature space to simplify their training scheme while maintaining the information preserved in the Reproducing Kernel Hilbert Space (RKHS). Therefore, the method performance will be severely affected when lifted features are not used.

Each method is trained with the training images to discover the attributes. Then we use the manifold $\mathcal{M}$ w.r.t. the test images for the evaluation. More precisely, each attribute descriptor is extracted from test images (*i.e.,* $\boldsymbol{z}_k, \boldsymbol{z}_k \in \{-1, 1\}^N$, where N is the number of test images). For each dataset, we use the attribute labels from AMT to represent the Meaningful Subspace, $\mathcal{S}$.

We adapted the MPPCA metric from the semi-supervised attribute discovery method proposed in [31]. In particular, to discover an attribute, the method in [31] progressively updates MPPCA model using human feedback. In our settings, we directly train MPPCA using attributes found from AMT for each dataset. To measure meaningfulness, we compute the posterior probability of the given discovered attribute to the MPPCA model. We train MPPCA model using five components and three dimensional subspace for ASUN dataset. As for ApAy dataset, we use three components and three dimensional subspace. This is because the number of attributes in ApAy dataset is much smaller than ASUN dataset. Unless otherwise stated, we follow the experiment settings for MPPCA as described in [31]. For instance, we employ a threshold on the posterior probability to determine whether an attribute is meaningful. The MPPCA metric is the computed by computing the percentage of the attributes deemed as meaningful over the total discovered attributes.

5.2. Do δ_{cvx} and δ_{jp} measure meaningfulness?

In this experiment, we evaluate whether the proposed approach really does measure the meaningfulness on a set of automatically discovered attributes. One of the key assumptions in our proposal is that the distance between the Meaningful Subspace and the given attribute set $\mathcal{D}$ reflects the meaningfulness of a set of attribute. More specifically, if the distance is small, it is assumed that the attribute set is potentially meaningful and vice versa. Aiming to evaluate that, we construct two sets of attributes, respectively with meaningful and non-meaningful attributes and observe their distances to the meaningful subspace.

As to the meaningful attribute set, we follow the methods used in Section 4. Providing manually labelled attribute set $\mathcal{S}$, here denoted the AMT attribute set [2], in each dataset, we follow the approach used in Section 4 to divide the set into two subsets $\mathcal{S}^1 \cup \mathcal{S}^2 = \mathcal{S}$ where $\mathcal{S}^1$ represents the Meaningful Subspace and $\mathcal{S}^2$ is considered as a set of discovered attributes (*i.e.,* $\mathcal{D} = \mathcal{S}^2$). Unequivocally, the attributes in $\mathcal{S}^2$ should be meaningful as they are manually labelled by human annotator. Thus, we name $\mathcal{S}^2$ as the *MeaningfulAttributeSet*.

For non-meaningful attribute set, we create this set by randomly generating the attributes. As described in Section 4, we generate a finite set of random attributes denoted by $\tilde{\mathcal{N}}$. We name this set as *NonMeaningfulAttributeSet* since it

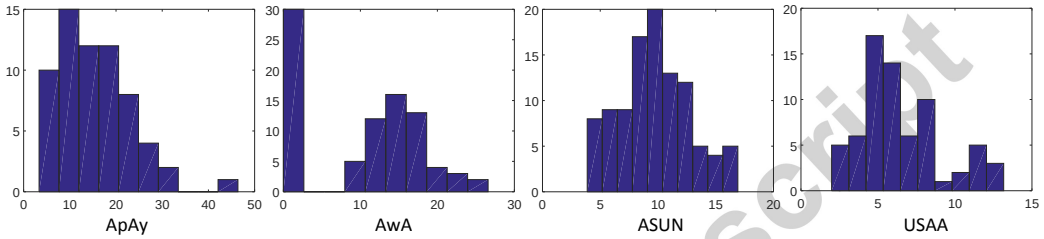
[2] As mentioned before, attributes discovered from the AMT procedure are considered meaningful. For the sake of clarity, we call these attributes AMT attributes.

is non-meaningful and should have significantly larger distance to the Meaningful Subspace.

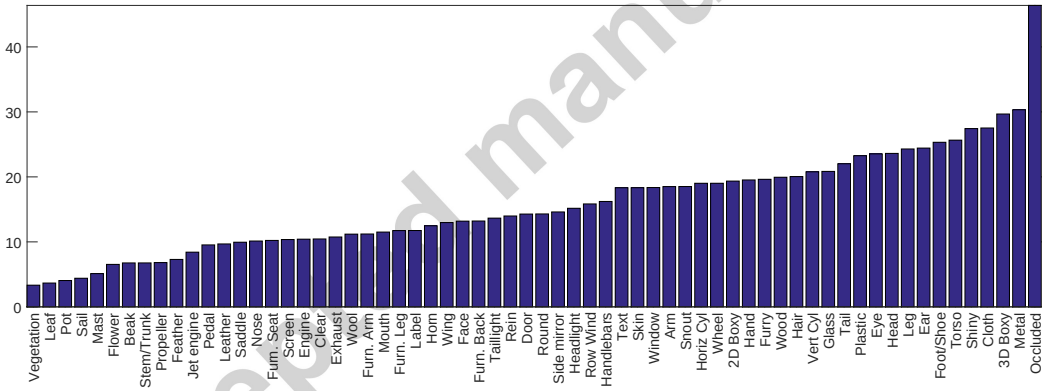
Recall that the Meaningful Subspace $\mathcal{S}^1$ needs to be carefully selected to maximize the meaningful subspace spanned. However, to show the efficacy of our proposed selection, we first randomly select $\mathcal{S}^1$. Then, on the second experiment, we apply our proposed selection approach. To perform our proposed selection approach, we must evaluate the independence of each AMT attribute via analysing its attribute name and computing its individual reconstruction error. We will always put independent attributes in $\mathcal{S}^1$. In other words, let $\hat{\mathcal{S}}^1$ be the set of AMT attributes marked as independent attributes. Then, the set $\mathcal{S}$ is divided into $\mathcal{S}^1$ and $\mathcal{S}^2$ such that, $\hat{\mathcal{S}}^1$ will be always in $\mathcal{S}^1$. In this case, we still randomly divide $\mathcal{S}$ with a constraint that the $\hat{\mathcal{S}}^1$ should always be in the set $\mathcal{S}^1$. As previously described, a leave-one-attribute-out scheme is used to determine the independence of an AMT attribute with respect to the rest of AMT attribute set. Fig. 2(a) presents the result of this analysis.

As we can see, the reconstruction error of ApAy dataset are, in general, much larger than the other datasets. We conjecture that this might caused by the fact that the other three datasets are all fine-grain classification datasets, however ApAy is proposed for addressing the general classification problem. This means the attributes provided in this dataset are more likely to be independent as they are used to describe a wide variety of classes. For further inspection, we take the ApAy dataset and present the results in the perspective of semantic reasoning of the attribute names. In Fig. 2(b), we plot the reconstruction error for each attribute in ascending order. As we can see, the attributes with low reconstruction errors are often more likely to be correlated (*i.e.,* less independent). For example, images such as the "leaf" and "pot" can reconstruct the "vegetation" attribute, "Sail" and "Mast" are often present together in the sailing scenes. However, the attributes with high reconstruction error are more likely to be independent (i.e. less correlated) such as shape-related attribute "3D Boxy", material-related attribute "Metal" and especially "Occluded". It is worthy to note that we only show the analysis using reconstruction error with convex hull regularization. The same findings are also exhibited when the ℓ_0 regularization is used.

To reasonably determine the parameter α (See 4.2), we average the highest reconstruction error scores from the other three fine-grain datasets. This gives us a value $\alpha = 18.89$. That means we consider any attributes in the datasets with error above α to be independent attributes. Thus the 22 independent attributes with highest reconstruction error can be put into the meaningful attribute subset $\mathcal{S}^1$ for better approximation of meaningful attribute subspace. The rest of the



(a)



(b)

Figure 2: The reconstruction error analysis on datasets. In (a), the horizontal axis represents the reconstruction error value, the vertical axis represents the frequency of the attributes which fall into the range of reconstruction error. In (b), the horizontal axis depicts the name of each attribute, the vertical axis represents the reconstruction error value.

attributes are still pooled and randomly selected. Table 1 shows the results with and without the proposed selection strategy. As we can see, after applying the selection, the MeaningfulAttributeSet which is always considered as meaningful, exhibits the lowest reconstruction error. Other methods almost remain the same with little random perturbation that indicates for the automatic attribute discovery methods our metric is quite stable. Again, we note that we use δ_{cvx} and similar results are also found when using δ_{jp}.

Table 1: Comparisons of reconstruction error results on ApAy dataset with and without the selection strategy.

	With selection	Without selection
PiCoDeS	12.65	12.52
DBC	48.97	49.66
ITQ	50.73	51.70
SPH	48.91	49.79
LSH	52.17	53.14
KSH	38.34	38.66
NonMeaningful	53.20	54.34
Meaningful	12.09	18.29

Now, we are ready to discuss the evaluation of our proposed approach to determine whether our approach can measure the attribute meaningfulness. In order to do that, we first perform the subspace interpolation of all the attribute set discovered by the methods. To perform the subspace interpolation, the random attributes are progressively added to the set of attributes from each method. By doing this, we can evaluate if the distance to Meaningful Subspace is enlarged when we progressively increase the number of non-meaningful attributes.

Fig. 3 presents the evaluation results.

Here we present the reconstruction error results for all 4 datasets where 16 and 32 attributes are discovered by the methods respectively. Recall that although we carefully selected $\mathcal{S}^1$, the set $\mathcal{S}$ is still randomly divided. We produced the results shown in Fig. 3 by repeating the random division of $\mathcal{S}^1$ 100 times and calculated the average distance. The detail results are also shown in Table 2. We note that no matter how $\mathcal{S}^1$ is selected, our method is relatively stable.

As we can see from the results that the *MeaningfulAttributeSet* has the closest distance to the Meaningful Subspace for both distances δ_{cvx} and δ_{jp} on all datasets. As expected, the *NonMeaningfulAttributeSet* has the largest distance compared

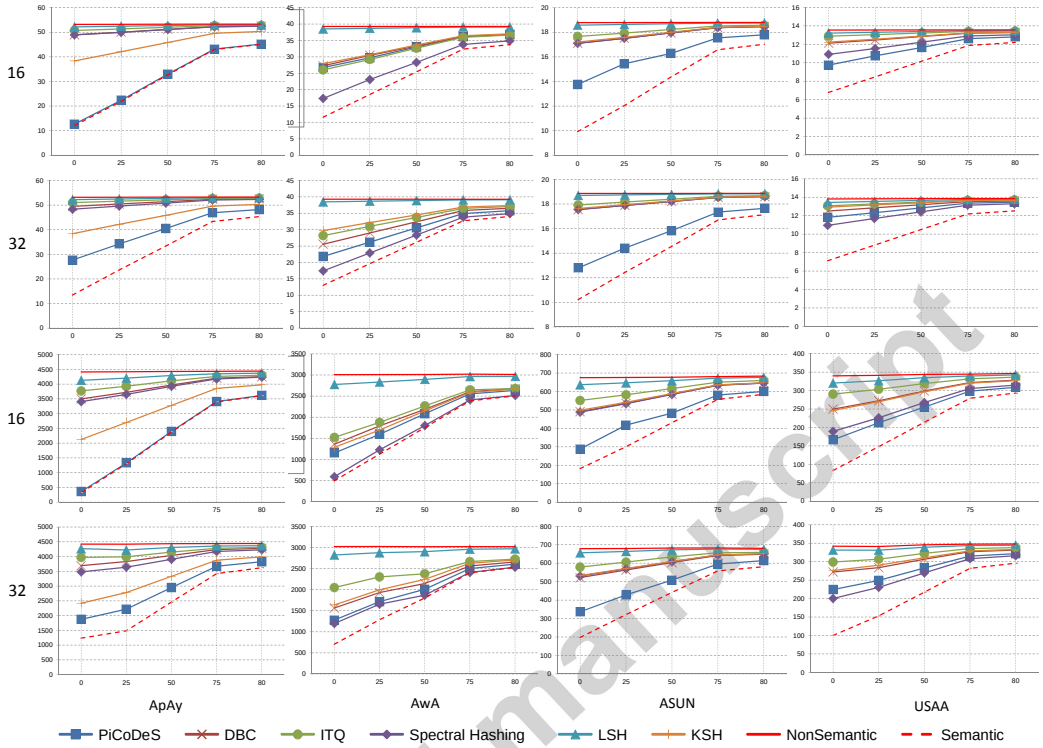


Figure 3: Validation of attribute meaningfulness measurement by reconstruction error δ_{cvx} (first and second rows) and δ_{jp} (third and fourth rows). In each subfigure, the horizontal axis represents the percentage of noise attributes, the vertical axis means the reconstruction error values. As we can see, both distances become larger when more random/non-meaningful attributes are added. MeaningfulAttributeSet has the closest distance to the Meaningful Subspace and NonMeaningfulAttributeSet always has the largest distance. Here, each method is configured to discover 16 and 32 attributes. The smaller the δ, the more meaningful the attribute set is.

Table 2: The table of values in reconstruction errors with standard deviation analysis. Letter E conventionally represents 'times ten raised to the power of'

		cvx				jp			
		ApAy	AwA	ASUN	USAA	ApAy	AwA	ASUN	USAA
16	PiCoDeS	12.65 ±9.57E-02	26.73 ±7.10E-02	13.76 ±2.33E-02	9.74 ±5.57E-02	358.61 ±1.53E+01	1150.76 ±1.71E+01	287.95 ±8.80E-01	166.57 ±3.17E+00
	DBC	48.97 ±1.78E-04	27.30 ±2.04E-01	17.10 ±2.19E-02	12.12 ±1.14E-01	3499.92 ±2.93E+00	1358.82 ±2.32E+01	491.13 ±1.01E+00	249.69 ±3.39E+00
	ITQ	50.73 ±3.15E-04	26.04 ±3.51E-01	17.66 ±4.85E-02	12.88 ±6.79E-02	3768.67 ±4.66E+00	1519.80 ±6.30E+01	551.51 ±3.13E+00	290.29 ±1.03E+00
	SPH	48.91 ±1.26E-03	17.32 ±1.12E-01	17.10 ±1.85E-02	10.91 ±3.40E-02	3406.41 ±3.22E+00	588.92 ±3.38E+01	488.34 ±9.30E-01	189.45 ±8.48E-01
	LSH	52.17 ±1.15E-09	38.51 ±7.64E-03	18.58 ±3.91E-02	13.22 ±9.34E-02	4127.82 ±6.09E+00	2775.42 ±3.59E+00	636.73 ±1.35E+00	320.39 ±1.53E+00
	KSH	38.34 ±1.53E-02	27.94 ±9.73E-02	17.21 ±1.91E-02	12.23 ±6.70E-02	2122.98 ±4.50E+00	1285.05 ±9.89E+00	498.27 ±1.48E+00	246.14 ±1.40E+00
	NonMeaningful	53.20 ±1.52E-09	39.27 ±4.32E-14	18.78 ±2.75E-02	13.55 ±1.21E-01	4411.74 ±7.45E-01	3008.71 ±8.16E-01	675.03 ±3.84E-01	339.79 ±2.52E+00
	Meaningful	**12.09 ±1.09E+00**	**11.60 ±2.08E+00**	**9.92 ±6.22E-01**	**6.78 ±6.27E-01**	**322.44 ±9.29E+01**	**492.54 ±1.06E+02**	**182.04 ±2.20E+01**	**83.54 ±1.44E+01**
32	PiCoDeS	27.67 ±1.12E-01	21.84 ±5.42E-02	12.79 ±4.39E-02	11.82 ±1.49E-01	1872.33 ±3.95E+01	1277.40 ±6.75E+01	336.71 ±4.75E+00	224.44 ±1.96E+00
	DBC	49.50 ±1.25E-04	25.56 ±2.69E-01	17.55 ±5.88E-02	12.50 ±3.02E-01	3691.58 ±7.46E+00	1563.89 ±6.90E+01	522.29 ±1.61E+00	272.04 ±2.89E+00
	ITQ	51.02 ±2.57E-04	28.18 ±4.31E-01	17.92 ±1.09E-01	13.07 ±2.09E-01	3971.37 ±1.22E+01	2048.51 ±6.32E+01	578.97 ±3.25E+00	298.99 ±1.66E+00
	SPH	48.35 ±1.62E-03	17.43 ±1.23E-01	17.54 ±4.82E-02	10.95 ±1.08E-01	3480.28 ±2.39E+01	1196.88 ±1.01E+02	530.47 ±1.45E+00	200.20 ±2.80E+00
	LSH	52.22 ±3.37E-09	38.48 ±1.12E-02	18.70 ±9.51E-02	13.43 ±2.99E-01	4268.15 ±9.82E+00	2822.03 ±1.02E+01	656.52 ±2.36E+00	331.45 ±1.54E+00
	KSH	38.37 ±3.52E-02	29.71 ±1.29E-01	17.63 ±4.34E-02	12.93 ±2.68E-01	2419.02 ±1.52E+01	1637.47 ±5.03E+01	533.15 ±1.46E+00	276.13 ±1.95E+00
	NonMeaningful	53.19 ±6.56E-10	39.28 ±1.18E-04	18.86 ±8.02E-02	13.82 ±3.73E-01	4421.49 ±1.13E+00	3020.82 ±1.69E+00	678.89 ±6.81E-01	342.10 ±1.47E+00
	Meaningful	**13.45 ±5.79E-01**	**13.04 ±1.51E+00**	**10.23 ±5.13E-01**	**7.12 ±4.13E-01**	**1234.20 ±9.14E+01**	**706.19 ±9.04E+01**	**197.50 ±1.50E+01**	**100.29 ±7.55E+00**

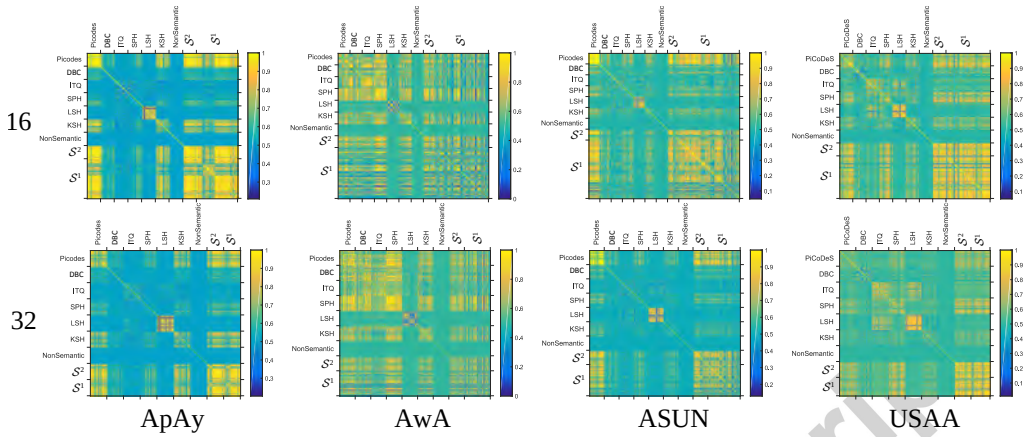


Figure 4: Visualization of co-occurrence matrix, the color bar represents the value of joint probability. The range between two ticks represents the attributes from each method. The first row shows the result for each dataset when each method is configured to discover 16 attributes. The second row shows the result for each dataset when each method is configured to discover 32 attributes.

with the others. In addition, when the random attributes are progressively added, the distance between the Meaningful Subspace and the sets of attributes discovered by each method increases. These results indicate that the proposed approach could be used to measure the meaningfulness of a set of attributes. Moreover, they also give a strong indication that there is a *shared structure* between meaningful attributes.

5.3. Attribute co-occurrence matrix analysis

For further inspection, we also perform the co-occurrence matrix analysis on the attributes discovered by each method and the AMT attributes *i.e.,* $\mathcal{S}^1$ and $\mathcal{S}^2$. The results are shown in Fig. 4.

The co-occurrence matrix figures represent the visualization of joint probability between the discovered attributes from each method and the AMT attributes, which are considered as meaningful. As we can see in the figure, almost in every dataset, the highest joint probability is achieved between $\mathcal{S}^1$ and $\mathcal{S}^2$. The trend is quite obvious in ApAy dataset and ASUN dataset and USAA dataset. However, the trend does not look apparent in AwA dataset. We conjecture this could be that there are many attributes in $\mathcal{S}$ that are independent. The attributes in AwA dataset are class-level *i.e.,* each sample in same class has the same attribute representation. Therefore, in order to guarantee the discriminative power between classes,

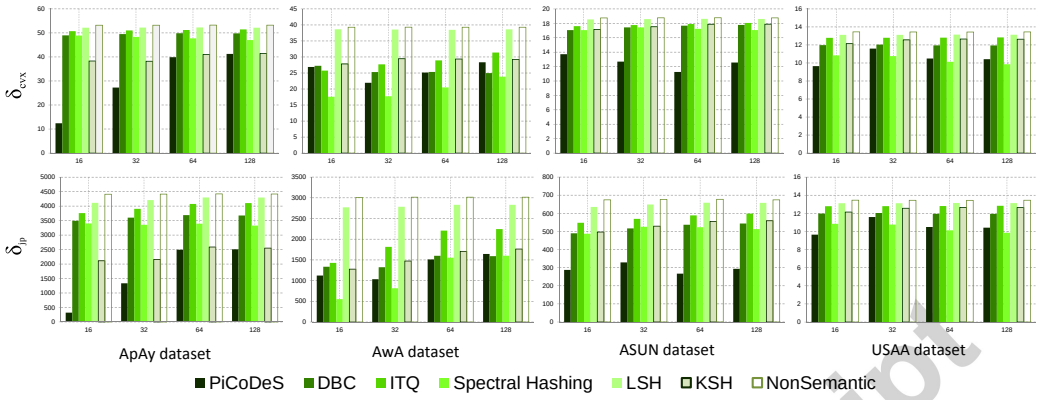


Figure 5: *Attribute meaningfulness* comparisons between different methods on variant number of discovered attributes. The first row reports the results using δ_{cvx} and the second row reports the results using δ_{jp}. The smaller the δ, the more meaningfulness.

the attributes may be chosen to reflect different aspects of classes, thus they could have lower joint probability. We note that the supervised attribute learning methods such as PiCoDeS, DBC and KSH also have comparable high probability with the AMT attributes. Another finding is that the attribute generated by the LSH method tend to have high joint probability with each other. This may be due to the simple linear projection of the data feature matrix in generating the final attribute representation. Generally, the results are consistent with the previous experiment, which further indicate the capability of our approach to capture the attribute meaningfulness.

5.4. Attribute set meaningfulness evaluation using δ_{cvx} and δ_{jp}

In this section, the meaningfulness is evaluated by δ_{cvx} and δ_{jp} for the set of attributes automatically discovered by various comparative methods in the literature. For that purpose, all manually labelled attributes from AMT in each dataset are used as the representation of the Meaningful Subspace. Then each method is configured to discover 16, 32, 64 and 128 attributes.

Fig. 5 reports the evaluation results on all datasets. It is noteworthy to mention that both the proposed distances δ_{cvx} and δ_{jp} are not calibrated and scaled; making it difficult to perform in-depth evaluation. However, we still can evaluate the results in terms of the method rank ordering (*i.e.,* which method takes first place and which comes the second).

PiCoDeS has the smallest distance in various number of attributes extracted on most of the datasets. PiCoDeS uses category labels and employs a max-margin

framework to jointly learn the category classifier and attribute descriptor in an attempt to maximize the discriminative power of the descriptor. In other words, the goal of PiCoDeS is to discover a set of attributes which can discriminate between categories.

DBC is also developed under the max-margin framework to extract meaningful attributes as PiCoDeS. However, compared with PiCoDeS, DBC discovers less meaningful attributes. We conjecture the reason could be DBC learns the whole attribute descriptors for each category simultaneously unlike PiCoDeS that learns the attribute individually. This scheme will inevitably emphasize category discrimination of attributes rather than preserving the meaningfulness of individual attribute. Note that here we do not suggest that DBC is not able to discover meaningful attributes, rather, PiCoDeS may find more meaningful attributes. Therefore, our finding does not contradict the results presented in the DBC original paper [33].

Another observation from the results of SPH indicate that it is able to discover meaningful attributes. SPH is aimed to discover binary codes via a graph embedding approach preserving the local neighborhood structure. One possible explanation could be that when two images belong to the same class, they should share more attributes indicating a shorter distance between them in the binary space, and vice versa.

Although ITQ aims to learn similarity preserving binary descriptor, it has a larger distance than SPH, DBC and PiCoDeS. The reason may be the way ITQ learns the binary descriptor which mainly relies on the global information of the data distribution. In other word, the algorithm minimizes the quantization error of the mapping data to the vertices of a zero centered binary hypercube suggesting that only global information by itself might not be sufficient to discover meaningful attributes.

As expected, the attribute sets from LSH have the largest distances to the Meaningful Subspace (*i.e.,* least meaningfulness). LSH uses random hyperplanes to project a data point into the binary space. Therefore, the consistent identifiable visual concepts are hardly presented in the positive images.

In summary, two recipes could be derived from the current results that could be significant for the future automatic attribute discovery method design: the method should attempt to preserve local neighborhood structure as well as to consider the discriminative power of attributes.

5.5. *Attribute set meaningfulness calibration using the proposed meaningfulness metric*

As described in section 5.4, the distance between attribute sets and the meaningful subspace have some limitations preventing us to perform in-depth analysis. Quantitative comparisons between different methods are more desirable in analysis of attribute meaningfulness. Here we report the meaningfulness metric results.

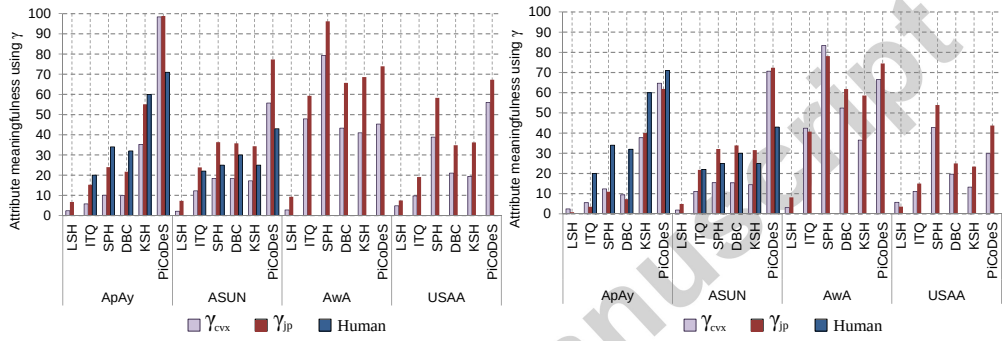


Figure 6: Comparisons of various methods using the proposed meaningfulness metric as well as human study results. Each method is set to discover 16 and 32 attributes. The higher the more meaningful. Human study is not conducted for AWA dataset as special zoology knowledge is required, nor for USAA dataset for inconvenience to display and subjectiveness bias problem. The human results for LSH method are 0 for ApAy and ASUN datasets.

As shown in (10), we apply γ_{cvx} and γ_{jp} on the datasets and for each method by calibrating the proposed distances δ_{cvx} and δ_{jp}.

The results are shown in Fig. 6 when each method is configured to discover 16 and 32 attributes. The rank orders of the methods are the almost the same with similar values in most tests by metric γ_{cvx} and γ_{jp}, with two exceptions in ASUN dataset. This can be explained to the fact that each metric captures a different aspect of attribute meaningfulness. The proposed γ_{cvx} captures a one-to-many relationship while γ_{jp} evaluates the one-to-one relationship. Then the equal weighted metric score $\tilde{\gamma}$ is applied for further analysis.

A user study is also conducted on the attributes discovered by each method. Since AwA requires experts in animal studies and USAA is a quite large video dataset whose complex social group activities are likely to cause subjectivity bias, we only use ApAy and ASUN dataset for the user study.

The study collected over 100 responses for each number of discovered attributes. In each response, there are positive and negative images presented from

8 randomly chosen discovered attributes. The user was asked whether these two set of images represent a consistent visual concept (hence meaningful). The users were university staffs and students with different knowledge background from various major fields including IT, Electronic Engineering, History, Philosophy, Religion and Classics and Chemical Engineering. The responses were averaged by considering 1 as meaningful and 0 as non-meaningful.

Table 3 illustrates the results of the proposed metric, $\tilde{\gamma}$ compared with the human study. In addition, we compare our metric against the MPPCA metric.

Table 3: The results (in percentage) of meaningfulness metric $\tilde{\gamma}$ on each dataset compared with user study and MPPCA metric on ApAy & ASUN datasets. Each method is configured to discover 32 attributes. In addition, for convenience we also report the proposed metric results on AwA and USAA datasets. The bold text indicates the top performing method in the proposed metric. The higher the more meaningful.

Methods \Datasets	ApAy			ASUN			AwA		USAA	
	$\tilde{\gamma}$	MPPCA	Human	$\tilde{\gamma}$	MPPCA	Human	$\tilde{\gamma}$	Human	$\tilde{\gamma}$	Human
LSH	1.7	0	0	3.4	0	0	5.6		4.7	
ITQ	4.5	34.4	20	16.4	31.3	22	41.6		13.1	
SPH	11.7	21.9	34	23.8	21.9	25	**80.7**	N/A	**48.3**	N/A
DBC	8.4	15.6	32	24.6	21.9	30	57.1		22.3	
KSH	38.9	37.5	60	23.0	12.5	25	47.6		18.3	
PiCoDeS	**63.3**	**56.3**	**71.0**	**71.5**	**78.1**	**43**	70.5		36.8	

Again, the attribute set discovered by LSH has the lowest meaningful content at close to 0%. Thus, LSH generates the least meaningful attribute sets. PiCoDeS and SPH generally discover more meaningful attribute sets. The methods using randomization scheme such as LSH and ITQ tend to generate lesser meaningful attribute sets with attribute meaningfulness around 1%-20%. The results indicate that the attribute meaningfulness could be significantly increased (*i.e.,* on average by 10-20 percentage points) by applying learning techniques such as PiCodes, DBC and SPH.

Compared with the results of the proposed metric $\tilde{\gamma}$, similar trends have been observed in the user study. Moreover, the user study results compared with γ_{cvx} and γ_{jp} are also shown in Fig. 6. Consistent similar trend as shown in previous experiments is visible.

As for MPPCA metric, similar results can be found such as LSH discovers the least meaningful attribute sets and PiCoDeS generally discover more meaningful attribute sets. However, result of our proposed method is closer to the human study in terms of the ranking order of attribute discovery methods. This could

indicate that the amount of AMT attributes used to train the MPPCA may not be sufficient. We note that, the MPPCA was originally designed to have human feedback in multiple iterative process to discover attributes [31]. As, in our experiment we only fed the MPPCA once with the AMT attributes, it may lack of human feedback.

We also perform two statistical analysis to compare which metric is closer to the human study. Both of the analysis are shown in Fig. 7 by applying a simple logarithmic fitting using the data from Table 3. Fig. 7(a) shows that regression line fits these data very well.

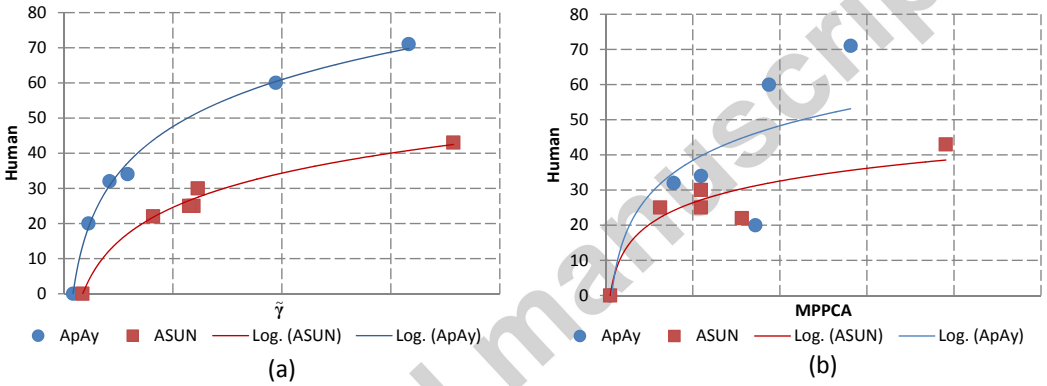


Figure 7: Demonstration of correlation analysis between user study and the proposed method $\tilde{\gamma}$ as well as MPPCA on both ApAy and Asun datasets.

The coefficients of determination R^2 [10] that indicates how well data fit a statistical model for ApAy dataset and ASUN dataset fitting are respectively 0.99 and 0.98 between the proposed metric and the human study results. The results suggest the regression line of our proposed method nearly perfectly fits the data. Fig. 7(b) shows the fitting result for MPPCA is not as good as the proposed method. The R^2 values of MPPCA metric are respectively 0.64 and 0.89 on ApAy dataset and ASUN dataset.

This demonstration further indicates to some extent, our proposed metric is able to evaluate the meaningfulness of a set of discovered attributes from comparative methods as human does.

It is noteworthy to mention that the time cost of the evaluation by our metric is significantly lower than the manual process using AMT. Recall that, the time required for a human annotator (an AMT worker) to finish one HIT is 2 minutes, an AMT worker may need 320 minutes to finish evaluating 5 methods wherein

each is configured to discover 32 attributes. Our approach only needs 105 seconds in total to evaluate all four datasets (*i.e.,* 35 seconds each); thus, leading to several orders of magnitude speedup!

6. Conclusions

In this paper, we studied a novel problem of measuring the meaningfulness of automatically discovered attribute sets. To that end, we proposed a novel metric, here called the *attribute meaningfulness* metric. We developed two distance functions for measuring the meaningfulness of a set of attributes. The distances were then calibrated by using subspace interpolation between Meaningful Subspace and Non-meaningful/Noise Subspace. We extended our previous work by proposing the meaningful attribute set selection technique that leads to a better meaningful subspace approximation. The final metric score indicates how much meaningful content is contained within the set of discovered attributes. In the extensive experiment, the proposed metrics were used to evaluate the *meaningfulness* of attributes discovered by two recent automatic attribute discovery methods and four hashing methods on four datasets. A user study on two datasets showed that the proposed metric has strong correlation to human responses. Our metric was also shown to be more correlated with the user study compared with a metric adapted from a recent semi-supervised attribute discovery method. All results suggested that there is a strong indication that the shared structure may exist among the meaningful attributes. The results also suggest that discovering attributes by optimising the attribute descriptor discrimination and/or preserving the local similarity structure could yield more meaningful attributes. In future work, we plan to explore other constraints or optimisation models [23] to capture the hierarchical structure of semantic concepts. Up to our knowledge, there are still no such works in deep learning area that have similar purpose as our work. In future, we plan to get more inspiration from the semantic learning in deep learning area and further develop our work in that direction. We also plan to perform more large-scale user studies using AMT on other datasets.

Acknowledgment: We deeply appreciate Teng Zhang's valuable discussion and efforts in proofreading and revisions. This project is partly funded by Sullivan Nicolaides Pathology and the Australian Research Council (ARC) Linkage Projects Grant LP130100230. Arnold Wiliem is funded by the Advance Queensland Early Career Research Fellowship.

Appendix A.

Proposition 4.1 Let $\tilde{\mathcal{S}} = \mathcal{S}^2 \cup \tilde{\mathcal{N}}$; when $\tilde{\mathcal{N}} = \{\}$, then the distance δ^* between $\tilde{\mathcal{S}}$ and $\mathcal{S}^1$ (refers to (7)) is minimized. However, when $\tilde{\mathcal{N}} \to \mathcal{N}$, then the distance between $\tilde{\mathcal{S}}$ and $\mathcal{S}^1$ is asymptotically close to $\delta^*(\mathcal{N}, \mathcal{S}^1; \mathcal{X})$, where δ^* is the distance function presented previously such as δ_{jp} and δ_{cvx}. More precisely, we can define the relationship as follows:

$$\lim_{|\tilde{\mathcal{N}}| \to \infty} \delta^*(\tilde{\mathcal{S}}, \mathcal{S}^1; \mathcal{X}) = \delta^*(\mathcal{N}, \mathcal{S}^1; \mathcal{X}). \tag{A.1}$$

Proof. Let $\boldsymbol{R}^*$ be the solution for the distance δ^*. The distance δ^* can be computed as follows:

$$\begin{aligned}
\delta^*(\mathcal{S}^2 \cup \tilde{\mathcal{N}}, \mathcal{S}^1; \mathcal{X}) &= \frac{1}{|\mathcal{S}^2 \cup \tilde{\mathcal{N}}|} \|\boldsymbol{A}\boldsymbol{R}^* - \boldsymbol{B}\|_F^2 \\
&= \frac{1}{|\mathcal{S}^2 \cup \tilde{\mathcal{N}}|} \sum_{\boldsymbol{b}_i \in \mathcal{S}^2 \cup \tilde{\mathcal{N}}} \|\boldsymbol{A}\boldsymbol{r}_i^* - \boldsymbol{b}_i\|_2^2 \\
&= \frac{1}{|\mathcal{S}^2 \cup \tilde{\mathcal{N}}|} \left\{ \sum_{\boldsymbol{b}_j \in \mathcal{S}^2} \|\boldsymbol{A}\boldsymbol{r}_j^* - \boldsymbol{b}_j\|_2^2 + \sum_{\boldsymbol{b}_l \in \tilde{\mathcal{N}}} \|\boldsymbol{A}\boldsymbol{r}_l^* - \boldsymbol{b}_l\|_2^2 \right\}
\end{aligned} \tag{A.2}$$

As $\mathcal{S}^2$ is assumed to be meaningful and $\tilde{\mathcal{N}}$ is not, then adding attribute $\boldsymbol{b}_l$, a member of $\tilde{\mathcal{N}}$, should increase the average distance. Thus, we have the following inequality:

$$\frac{1}{|\mathcal{S}^2 \cup \tilde{\mathcal{N}}|} \left\{ \sum_{\boldsymbol{b}_j \in \mathcal{S}^2} \|\boldsymbol{A}\boldsymbol{r}_j^* - \boldsymbol{b}_j\|_2^2 + \sum_{\boldsymbol{b}_l \in \tilde{\mathcal{N}}} \|\boldsymbol{A}\boldsymbol{r}_l^* - \boldsymbol{b}_l\|_2^2 \right\} \geq \frac{1}{|\mathcal{S}^2|} \sum_{\boldsymbol{b}_j \in \mathcal{S}^2} \|\boldsymbol{A}\boldsymbol{r}_j^* - \boldsymbol{b}_j\|_2^2 \tag{A.3}$$

It means that the distance between $\tilde{\mathcal{S}} = \mathcal{S}^2 \cup \tilde{\mathcal{N}}$ and $\mathcal{S}^1$ can only be minimized when $\tilde{\mathcal{N}}$ is an empty set. On the other hand, when we keep increasing the size of $\tilde{\mathcal{N}}$, the contribution of the second term in (A.2) becomes more significant than the first term. Thus, (A.2) is approximately close to:

$$\approx \frac{1}{|\tilde{\mathcal{N}}|} \left\{ \sum_{\boldsymbol{b}_l \in \tilde{\mathcal{N}}} \|\boldsymbol{A}\boldsymbol{r}_l^* - \boldsymbol{b}_l\|_2^2 \right\}, \text{ as } |\tilde{\mathcal{N}}| >> |\mathcal{S}^2| \tag{A.4}$$

In addition, as $|\tilde{\mathcal{N}}| \to \infty$, $\tilde{\mathcal{N}}$ will be close to $\mathcal{N}$. Thus, the above equation is approximately close to:

$$\approx \frac{1}{|\mathcal{N}|}\left\{\sum_{\boldsymbol{b}_l \in \mathcal{N}} \|\boldsymbol{A}\boldsymbol{r}_l^* - \boldsymbol{b}_l\|_2^2\right\}$$
$$= \delta^*(\mathcal{N}, \mathcal{S}^1; \mathcal{X}) \tag{A.5}$$

□

[1] Baird, H. S., Coates, A. L., Fateman, R. J., 2003. Pessimalprint: a reverse turing test. International Journal on Document Analysis and Recognition 5 (2-3), 158–163. 5

[2] Bay, H., Ess, A., Tuytelaars, T., Van Gool, L., 2008. Speeded-up robust features (surf). Computer Vision and Image Understanding 110 (3), 346–359. 15

[3] Bergamo, A., Torresani, L., Fitzgibbon, A. W., 2011. Picodes: Learning a compact code for novel-category recognition. In: NIPS. 3, 14, 15

[4] Bishop, R., Goldberg, S., 2012. Tensor Analysis on Manifolds. Dover Publications. 13

[5] Biswas, A., Parikh, D., 2013. Simultaneous active learning of classifiers & attributes via relative feedback. In: CVPR. 7

[6] Bosch, A., Zisserman, A., Munoz, X., 2007. Representing shape with a spatial pyramid kernel. In: Proceedings of the 6th ACM international conference on Image and video retrieval. pp. 401–408. 14

[7] Cevikalp, H., Triggs, B., 2010. Face recognition based on image sets. In: CVPR. 9

[8] Chang, X., Yang, Y., Long, G., Zhang, C., Hauptmann, A. G., 2016. Dynamic concept composition for zero-example event detection. In: AAAI. 2

[9] Dong, Y., Tao, D., Li, X., Ma, J., Pu, J., 2015. Texture classification and retrieval using shearlets and linear regression. IEEE Transactions on Cybernetics 45 (3), 358–369. 3

[10] Draper, N. R., Smith, H., 2014. Applied regression analysis. John Wiley & Sons. 26

[11] Farhadi, A., Endres, I., Hoiem, D., Forsyth, D., 2009. Describing objects by their attributes. In: CVPR. 2, 14

[12] Feng, J., Jegelka, S., Yan, S., Darrell, T., 2014. Learning scalable discriminative dictionary with sample relatedness. In: CVPR. 2

[13] Fu, Y., Hospedales, T. M., Xiang, T., Gong, S., 2012. Attribute learning for understanding unstructured social activity. In: ECCV. 14, 15

[14] Gong, Y., Lazebnik, S., 2011. Iterative quantization: A procrustean approach to learning binary codes. In: CVPR. 3, 14

[15] Gopalan, R., Li, R., Chellappa, R., 2014. Unsupervised adaptation across domain shifts by generating intermediate data representations. IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (11), 2288–2302. 11

[16] Hong, S., Choi, J., Feyereisl, J., Han, B., Davis, L. S., 2016. Joint image clustering and labeling by matrix factorization. IEEE Transactions on Pattern Analysis and Machine Intelligence 38 (7), 1411–1424. 6

[17] Huang, C., Change Loy, C., Tang, X., 2016. Unsupervised learning of discriminative attributes and visual representations. In: CVPR. 6

[18] Kovashka, A., Grauman, K., 2015. Discovering attribute shades of meaning with the crowd. International Journal of Computer Vision 114 (1), 56–73. 3

[19] Kovashka, A., Parikh, D., Grauman, K., 2015. Whittlesearch: Interactive image search with relative attribute feedback. International Journal of Computer Vision 115 (2), 185–210. 2

[20] Kumar, N., Berg, A. C., Belhumeur, P. N., Nayar, S. K., 2009. Attribute and simile classifiers for face verification. In: ICCV. 2

[21] Lampert, C. H., Nickisch, H., Harmeling, S., 2014. Attribute-based classification for zero-shot visual object categorization. IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (3), 453–465. 14

[22] Leskovec, J., Rajaraman, A., Ullman, J., 2013. Mining of Massive Datasets. Cambridge university press. 3, 14

[23] Li, C., Yu, X., Huang, T., Chen, G., He, X., 2016. A generalized hopfield network for nonsmooth constrained convex optimization: Lie derivative approach. IEEE transactions on neural networks and learning systems 27 (2), 308–321. 27

[24] Liu, J., Yu, Q., Javed, O., Ali, S., Tamrakar, A., Divakaran, A., Cheng, H., Sawhney, H. S., 2013. Video event recognition using concept attributes. In: WACV. pp. 339–346. 2

[25] Liu, L., Wiliem, A., Chen, S., Zhao, K., Lovell, B. C., 2016. Determining the best attributes for surveillance video keywords generation. In: The IEEE International Conference on Identity, Security and Behavior Analysis (ISBA). 3, 6, 7, 9

[26] Liu, W., Wang, J., Ji, R., Jiang, Y.-G., Chang, S.-F., 2012. Supervised hashing with kernels. In: CVPR. 14

[27] Liu, Y., Zhang, D., Lu, G., Ma, W.-Y., 2007. A survey of content-based image retrieval with high-level semantics. Pattern Recognition 40 (1), 262–282. 2

[28] Lowe, D. G., 2004. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60 (2), 91–110. 14

[29] Ma, Z., Yang, Y., Xu, Z., Yan, S., Sebe, N., Hauptmann, A. G., 2013. Complex event detection via multi-source video attributes. In: CVPR. 2

[30] Parikh, D., Grauman, K., 2011. Interactive discovery of task-specific nameable attributes. In: Workshop on Fine-Grained Visual Categorization, CVPR. 3, 7

[31] Parikh, D., Grauman, K., 2011. Interactively building a discriminative vocabulary of nameable attributes. In: CVPR. 2, 3, 5, 6, 7, 8, 14, 16, 26

[32] Patterson, G., Hays, J., 2012. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In: CVPR. 2, 5, 14, 15

[33] Rastegari, M., Farhadi, A., Forsyth, D., 2012. Attribute discovery via predictable discriminative binary codes. In: ECCV. 3, 6, 14, 23

[34] Rohrbach, A., Rohrbach, M., Qiu, W., Friedrich, A., Pinkal, M., Schiele, B., 2014. Coherent multi-sentence video description with variable level of detail. In: Pattern Recognition (GCPR). 2

[35] Rui, Y., Liu, Z., 2004. Artifacial: Automated reverse turing test using facial features. Multimedia Systems 9 (6), 493–502. 5

[36] Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., Poggio, T., 2007. Robust object recognition with cortex-like mechanisms. IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (3), 411–426. 9

[37] Sharmanska, V., Quadrianto, N., Lampert, C. H., 2012. Augmented attribute representations. In: ECCV. 3

[38] Shechtman, E., Irani, M., 2007. Matching local self-similarities across images and videos. In: CVPR. 15

[39] Tipping, M., Bishop, C., 1999. Mixtures of probabilistic principal component analyzers. Neural Computation 11 (2), 443–482. 5, 6

[40] Turing, A. M., 1950. Computing machinery and intelligence. Mind, 433–460. 5

[41] Van De Sande, K. E., Gevers, T., Snoek, C. G., 2010. Evaluating color descriptors for object and scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (9), 1582–1596. 14

[42] Von Ahn, L., Blum, M., Hopper, N. J., Langford, J., 2003. Captcha: Using hard ai problems for security. In: Advances in Cryptology EUROCRYPT 2003. pp. 294–311. 6

[43] Wang, J., Markert, K., Everingham, M., 2009. Learning models for object recognition from natural language descriptions. In: BMVC. 3

[44] Weiss, Y., Torralba, A., Fergus, R., 2009. Spectral hashing. In: NIPS. 3, 14

[45] Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., Perona, P., 2010. Caltech-UCSD Birds 200. Tech. Rep. CNS-TR-2010-001, California Institute of Technology. 3

[46] Wiliem, A., Hobson, P., Lovell, B. C., 2014. Discovering discriminative cell attributes for hep-2 specimen image classification. In: WACV. 3, 6

[47] Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., Torralba, A., 2010. Sun database: Large-scale scene recognition from abbey to zoo. In: CVPR. 15

[48] Yu, F., Cao, L., Feris, R., Smith, J., Chang, S.-F., 2013. Designing category-level attributes for discriminative visual recognition. In: CVPR. 3

[49] Zhang, D., Islam, M. M., Lu, G., 2012. A review on automatic image annotation techniques. Pattern Recognition 45 (1), 346–362. 3

[50] Zhang, J., Wang, M., Zhang, S., Li, X., Wu, X., 2016. Spatiochromatic context modeling for color saliency analysis. IEEE Transactions on Neural Networks and Learning Systems 27 (6), 1177–1189. 3