Zero-Shot Object Recognition System Based on Topic Model

Wai Lam Hoo, *Student Member, IEEE*, and Chee Seng Chan, *Senior Member, IEEE*

Abstract—Object recognition systems usually require fully complete manually labeled training data to train classifier. In this paper, we study the problem of object recognition, where the training samples are missing during the classifier learning stage, a task also known as zero-shot learning. We propose a novel zero-shot learning strategy that utilizes the topic model and hierarchical class concept. Our proposed method advanced where cumbersome human annotation stage (i.e., attribute-based classification) is eliminated. We achieve comparable performance with state-of-the-art algorithms in four public datasets: PubFig (67.09%), Cifar-100 (54.85%), Caltech-256 (52.14%), and Animals with Attributes (49.65%), when unseen classes exist in the classification task.

Index Terms—Image understanding, object recognition, topic model, zero-shot learning.

I. INTRODUCTION

OBJECT classification from natural images is useful in content-based image retrieval, video surveillance, robot localization, and image understanding. According to Lampert *et al.* [1], humans are able to distinguish between at least 30 000 relevant classes. However, training conventional object detectors for all these classes would require millions of well-labeled training images and is likely out of reach for years to come.

As such, the zero-shot learning paradigm [1]–[7] is motivated from the human ability to learn and abstract from examples and the capability to describe completely unseen classes (i.e., training classes are not available during training of the object detector) from existing (known) classes. For instance, the authors in [1]–[4] recognize a set of unseen objects using a list of high-level attributes that serve as an intermediate layer in the classifier cascade. The attributes enable those systems to recognize the object classes, even without a single training example. Others, e.g., authors in [5] and [6], use semantic relationships from different reference classes to predict the unseen classes. Although promising results were obtained, all these aforementioned approaches require either extensive human supervision to build the attributes or a tight semantic relationship between the unseen classes and the training classes.

In this paper, we propose 1) a *topic model* to replace the attributes [1]–[4] so that extensive human supervision is no longer required, and 2) the *Hierarchical Class (HiC) concept* to relate the unseen classes to the existing seen classes. The HiC concept has a loose relationship image hierarchy compared with [5] and [6]. Our framework starts with building a bag-of-word (BoW) model using the image features from the (small amount of) seen (available) classes. Herein, the HiC concept is utilized to build the codebook. A topic model [here, we employ the probabilistic latent semantic analysis (pLSA)] is learned using the generated BoW model. Based on the learned pLSA model and the HiC concept, signature topics for both the seen and unseen classes are deduced (i.e., we cluster similar object classes that share visual similarity). Finally, object classification is performed using the deduced signature topics representation. Experimental results using four publicly available datasets, namely the Pub-Fig, Cifar-100, Caltech-256, and AwA datasets, have shown the effectiveness of the proposed method.

The remainder of this paper is structured as follows. Section II presents the related work. Section III details the proposed methodology. Section IV shows the experimental results, and Section V presents the discussion.

II. RELATED WORK

Palatucci *et al.* [17] showed that the attribute description of an instance or category is useful as a semantically meaningful intermediate representation to bridge the gap between low-level features and high-level classes. Thus, the attributes facilitate transfer and zero-shot learning to alleviate issues of the lack of labeled training data by expressing classes in terms of well-known attributes. This is followed by Lampert *et al.* [1], [13] who extended the work to animal categorization by introducing direct attributes prediction (DAP) and indirect attributes prediction (IAP).

Unlike [1], [13], and [17], Parikh and Grauman [2] introduced relative attributes to perform zero-shot learning. This approach captures the relationships between images and objects in terms of human-nameable visual properties. For example, the models capture that animal A is "taller" than animal B, or subject X is "happier" than subject B. This allows a richer language of supervision and description than the commonly used categorical (binary) attributes. Although relative attributes seem efficient for zero-shot learning, the dataset needs to be intraclass (i.e., the images in the dataset must belong to a set of object classes that are visually similar). In addition, a binary or relative relationship between all classes needs to be defined beforehand. Such a process will require extensive human supervision efforts, and the decision is always subjective.
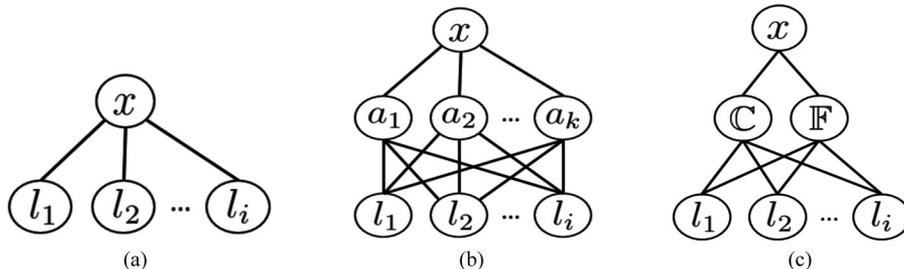
Fig. 1. Comparison between (a) weakly supervised learning as conventional learning algorithms, (b) learning classifier by including association between images with attributes, or tags, respectively, and (c) the HiC concept.

In our proposed strategy, we replace the attributes with a topic model in order to reduce the human supervision needed. Others who use topic models in zero-shot learning are in [14] and [15]. They propose a hybrid attribute-topic model to deal with group social activities. Specifically, they define three unique attributes: user-defined, latent class-conditional, and latent generalized free attributes. These attributes are learned jointly in a semilatent attribute space and as the multimodal latent attribute topic model (M2LATM). The motivation is to reduce the annotation effort through the introduction of the latent attributes in their proposed framework. In contrast, our focus in this paper is on object recognition that learns the topic model directly from the BoW representations and infers the unseen classes using the proposed HiC concept. We eliminate the time-consuming human annotation process by replacing the attributes with topic models. Instead of learning the topic models on top of user-defined and latent attributes [14], [15], we choose the pLSA as our topic model because it does not require prior comparison with the Latent Dirichlet Allocation model. We further extend the topic model representation as a mapping algorithm to object classes so that zero-shot learning would be possible.

Fig. 1 shows conventional solutions that associate each image with a class label [8]–[11] or further describe the image content with the association of attributes [1]–[3], [12]–[15] or image tags [16]. These are insufficient in zero-shot learning because these attributes and tags can be redundant and not useful when too many of them are introduced. Yet, there is no specific evaluation method on "what is an effective attribute or tag." Therefore, we introduce a new codebook learning method, i.e., the HiC concept that utilizes the HiC characteristics during the codebook learning stage. This concept is inspired by [5], [6], and [18], where a set of common objects are clustered into different classes in order to deduce the relationship among them. Specifically, we integrate two different levels of image class labels, namely the Coarse Class, $\mathbb{C}$ and Fine Class, $\mathbb{F}$. Then, this class hierarchy is learned in the topic model to identify the significant differences among the classes and improve the model prediction capability. Such an approach is better than attribute-based classification [1], [13], which is commonly applicable in interclass problems only. The HiC concept manages to deal with both interclass and intraclass problems.

Similar work that employed the HiC strategy in zero-shot learning paradigm includes Rohrbach *et al.* [4] and Frome *et al.* [5]. In both approaches, a set of frameworks on how to in-

corporate the semantic information from a language model/set to assist in the zero-shot learning is studied. Rohrbach *et al.* [4] employed WordNet and Wikipedia as the language model and learned a similarity measure to represent the hierarchy/attributes/objectness measure between the object classes. Frome *et al.* [5] extended the idea to learn the class relationship directly from the unannotated data (i.e., visual-semantic relationship between object classes from millions of documents in Wikipedia) using the Deep Visual-Semantic Embedding Model.

In another approach, Mensink *et al.* [6] used a different concept where a distance metric from a set of seen classes (e.g., 800 seen classes) and errors for both seen and unseen classes (e.g., 800 seen classes and 200 unseen classes, result in 1000-way classification) are learned. In order to classify the object classes, a nearest class mean classifier is employed. This approach does not require the semantic relationship and manages to generalize the unseen classes in near to zero computational cost. For our proposed framework, although it is similar to the hierarchy-based knowledge transfer in [4], we do not need a language model to build the hierarchy. Instead, the HiC concept relates the unseen classes to the seen classes. In addition, we use learnt topic model to perform the zero-shot learning, which is different from the attribute-based or direct similarity-based knowledge transfer in [4] that uses attributes or objectness measure and [6] that uses metric learning.

III. APPROACH

In this section, we first discuss the prerequisites of the proposed framework: BoW model and topic model. Second, we explain the HiC concept and detail how to perform zero-shot learning in pLSA with the HiC concept. Finally, we show the inference method for image classification purposes.

A. Codebook Representation

To build the BoW model, we engaged the random forest (RF) algorithm [11], [19], where a random decision tree is constructed using a random subset of the training data with replacement. The *labeled* training images at a particular node $I_{\mathrm{node}} = \{x_i, l_i\}$ are recursively split into left node I_{left} and right node I_{right} subsets, according to a threshold $t \in T$ and a split function f

$$I_{\mathrm{left}} = \{x_i \in I_{\mathrm{node}} | f(x_i) < t\}, I_{\mathrm{right}} = I_{\mathrm{node}} \setminus I_{\mathrm{left}} \quad (1)$$

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

HOO AND CHAN: ZERO-SHOT OBJECT RECOGNITION SYSTEM BASED ON TOPIC MODEL   3
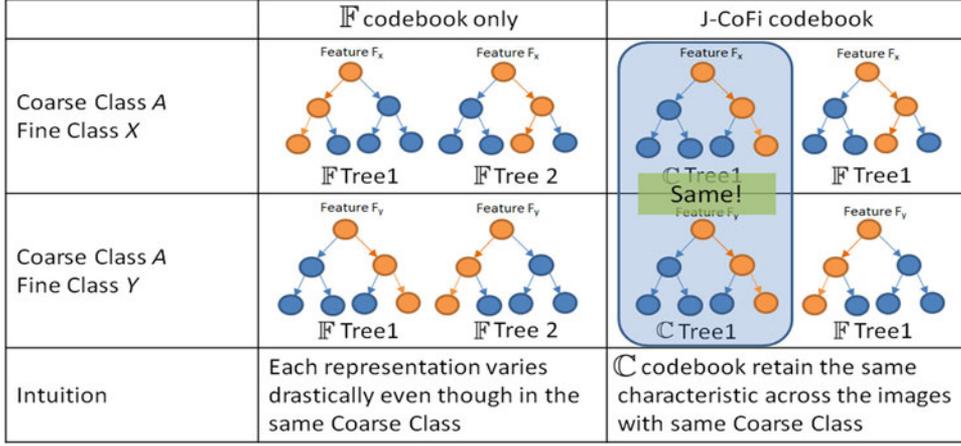
Fig. 2.   Effects of Coarse Class ($\mathbb{C}$) on J-CoFi codebook compared with ordinary $\mathbb{F}$-based codebook. Two examples include building a RF with two trees. One example has $\mathbb{F}$ trees only. Another example has one $\mathbb{C}$ tree and one $\mathbb{F}$ tree. After the trees are built, we fit feature F_x and F_y that have the similar $\mathbb{C}$, but different $\mathbb{F}$ into the trees. We notice significant differences in the leafnode (orange nodes indicate the path choice of the feature F) for an $\mathbb{F}$-based codebook. However, with the help of $\mathbb{C}$ in the J-CoFi codebook, the feature F with the same $\mathbb{C}$ will have similar path choice in the $\mathbb{C}$ Tree, but a different one in the $\mathbb{F}$ tree. With this, we can build a BoW model that retains similarities for the images that belong to the same $\mathbb{C}$.

TABLE I
EXAMPLES OF THE $\mathbb{C}$ AND $\mathbb{F}$ RELATIONSHIP IN THE HiC CONCEPT

$\mathbb{C}$	Electrical Devices	Building	Water Spot
$\mathbb{F}$	- Television - Refrigerator - Washing Machine	- House - Apartment - Tall Building	- Coast - Beach - Underwater

where x_i are the feature vectors from the training images, and l_i are the associated class labels. At each split node, random subsets of features are generated and compared with T. In this process, $t \in T$ that maximizes the expected information gain $\triangle E$ is selected:

$$\triangle E = E(I_{\text{node}}) - \sum_{p=\text{left,right}} \frac{\mid I_p \mid}{\mid I_{\text{node}}\mid} \mid E(I_p) \tag{2}$$

where $E(I) = p(l_i) \log p(l_i)$, and $E(I)$ is the Shannon entropy of the probability class histogram $p(l_i)$. As such, the leafnodes of all trees in the RF form a codebook. Then, the codebook is used to quantize I into BoW representation, by passing x_i to each tree and count the occurrence of each leafnode.

B. Topic Model

Our model is based on a latent topic model, in particular, the pLSA model. We briefly introduce it using the terminology in our context. Suppose we are given a collection of images $D = \{d_1, \ldots, d_N\}$. Each image d is represented by a collection of features $W = \{w_1, \ldots, w_V\}$, where it shows how frequent a particular w_v is used in d. A word is the basic item from a codebook indexed by $\{1, 2, \ldots, V\}$. A joint probability model $p(w, d)$ over $V \times N$ can be defined as

$$p(w, d) = \sum p(z) p(w|z) p(z|d) \tag{3}$$

where $z \in Z = \{z_1, \ldots, z_K\}$ is a latent variable. We can further derive the document-specific word distribution $p(w|d)$ as

$$p(w|d) = \sum p(w|z) p(z|d). \tag{4}$$

However, at the current setting, (3) and (4) could not infer the unseen classes as the algorithm needs prior knowledge about which z belongs to which c [20] or a set of labeled training image $\{x_i, l_i\}$ in learning the model. In the zero-shot paradigm, such information is simply not available. In order to handle this issue, we proposed the HiC concept (discussed next) so that we can infer the unseen classes to perform zero-shot learning using the pLSA model.

C. Hierarchical Class Concept

We introduced the HiC concept—a nested class concept as illustrated in Fig. 1(c), where one image consists of two class labels (semantically related), HiC = $\{\mathbb{C}, \mathbb{F}\}$. One has a broader visual concept, namely the Coarse Class, while the other class labels have a narrow visual concept, namely the Fine Class. Table I shows some examples of the HiC concept.

Definition 3.1: Coarse Class $\mathbb{C}$ is a large concept class (parent) that shares a conceptual similarity, either physical or biological, within its own Fine Class.

Definition 3.2: Fine Class $\mathbb{F}$ is a specific object class and is a subset to one of the Coarse Class (child).

1) Codebook Representation in HiC concept: Using the HiC concept, we have three new codebook representations, that are 1) Coarse ($\mathbb{C}$) or Fine ($\mathbb{F}$); 2) Joint Coarse-Fine (J-CoFi); and 3) CoarseFine (CoFi). We next explain their properties.

Property 3.1 (Coarse ($\mathbb{C}$) or Fine ($\mathbb{F}$)). The $\mathbb{C}$ and $\mathbb{F}$ codebooks are similar to the initial RF learning described in Section III-A, except that we substitute l_i in the Shannon entropy with $\mathbb{C}_i$ or $\mathbb{F}_i$, respectively. We illustrate in Fig. 2 that utilizing only the $\mathbb{F}$ codebook is not an optimum setting as each of the codebook representations varies drastically although they

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

4   IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS

TABLE II
COMPARISON BETWEEN [COARSE ($\mathbb{C}$) OR FINE ($\mathbb{F}$)] VERSUS [JOINT
COARSE-FINE (J-COFI)] CODEBOOK LEARNING STRATEGY

Codebook Type	$\mathbb{F}$	$\mathbb{C}$	J-CoFi	Shannon Entropy ($E(I)$)
$\mathbb{C}$ Tree	No	Yes	Yes	$p(\mathbb{C})\log p(\mathbb{C})$
$\mathbb{F}$ Tree	Yes	No	Yes	$p(\mathbb{F})\log p(\mathbb{F})$

belong to the same $\mathbb{C}$. Therefore, we built a variant, namely the J-CoFi.

Property 3.2 (Joint Coarse-Fine (J-CoFi)). The J-CoFi codebook strategy adapts both $\mathbb{C}$ and $\mathbb{F}$ information during the RF learning. Specifically, we denote the total number of trees as R. If one uses r of $\mathbb{C}$ trees that govern the similarity between x_i with the same $\mathbb{C}_i$ and $R - r$ of $\mathbb{F}$ trees that distinguish those x_i within its associated $\mathbb{C}$, this will result in a BoW model that has a similar histogram shape for codebook bins that are created by $\mathbb{C}$ trees. Hence, it eliminates the limitations in Property 3.1.

Table II summarizes the difference between Property 3.1 and 3.2. There still exist limitations in the Property 3.1 and 3.2 when (2) is employed to compute $\triangle E$. That is, at one time, one could only optimize either $\mathbb{C}_i$ or $\mathbb{F}_i$ during the RF tree node splitting, and therefore, we introduce the CoFi codebook (see Property 3.3) to handle this limitation.

Property 3.3 (CoarseFine (CoFi)). The CoFi is proposed to learn the trees in such a way that utilizes both the $\mathbb{C}_i$ and $\mathbb{F}_i$, simultaneously in the RF tree node splitting. Specifically, we modified (2); therefore, for each CoFi tree, we consider the total maximum $\triangle E$ from $\mathbb{C}_i$ and $\mathbb{F}_i$ simultaneously for each split node as $\triangle E_{\text{total}}$:

$$\triangle E_{\text{total}} = \sum_{\mathbb{F}, \mathbb{C} \in c} \left[E(I_c) - \sum_{p=\text{left,right}} \frac{|I_{c_p}|}{|I_n|} E(I_p) \right]. \quad (5)$$

and the splits that maximize the $\triangle E_{\text{total}}$ will be selected.

D. Zero-Shot Learning in Probabilistic Latent Semantic Analysis With the Hierarchical Class Concept

In order to perform the zero-shot learning using the HiC concept, we denote a seen class as $s \in S$ and an unseen class as $u \in U$, where $\{S, U\} \subset C$. As such, we collect a set of seen classes pair α_u for each u that associate u to a pair of seen classes s which belongs to the same $\mathbb{C}$:

$$\alpha_u = \{(g, h) \in \mathbb{F}, \mathbb{F} \subset \mathbb{C} | g \sim u \sim h\} \quad (6)$$

where $\{g, h\} \in S$ and $\sim$ indicates conceptual similarity between c (i.e., as described in Definition 3.1 and in [2]). In the pLSA model, we introduce a novel mapping algorithm namely topic sets $\mathbb{T}$ that indicate index of z. Each c_m will be associated with specific $\mathbb{T}_m$, which creates a relationship between z and c_m. Our idea is that the unseen class u that could be related to a pair of unseen classes s (i.e., in this case are g and h) will have high similarity for their respective $\mathbb{T}$. Therefore, we could relate u by defining $\mathbb{T}_u$ that satisfies the conditions of $\mathbb{T}_g \sim \mathbb{T}_u \sim \mathbb{T}_h$ and $(g, h) \in S$. We denote $\mathbb{T}_s$ as the signature topic set for the

seen class s as

$$\mathbb{T}_s = \arg\max_{\mathbb{T}_m} \sum_{k \in \mathbb{T}_m} p(z_k|d_m) \quad (7)$$

where the size of M is 2^K, and $p(z_k|d_m)$ is a class-specific topic distribution that is used to determine $\mathbb{T}_m$ for every c_m:

$$p(z_k|d_m) = \frac{\sum_{n \subset m} p(z_k|d_n)}{\sum_m p(z_k|d_m)} \quad (8)$$

where $\mathbb{T}_u$ is inferred as the union of the $\mathbb{T}_s$ pairs ($\mathbb{T}_g$ and $\mathbb{T}_h$) to achieve zero-shot learning. Taking $K = 3$ as an example, the size of M is 8 ([0 0 1], [0 1 0], [1 0 0], [0 1 1], [1 0 0], [1 0 1], [1 1 0], [1 1 1]), where 1 indicates the signature topic(s) and *vice versa*. Ideally, if $\mathbb{T}_g$ is [0 0 1] and $\mathbb{T}_h$ is [1 0 0], then $\mathbb{T}_u$ is [1 0 1].

Finally, given a test class c'_m, it can be predicted by evaluating

$$p(c'_m|d_{\text{test}}) = \frac{\sum_{k \in \mathbb{T}_{c'_m}} p(z_k|d_{\text{test}})}{\sum_m p(c_m|d_{\text{test}})}. \quad (9)$$

Algorithm 1 summarizes the proposed framework.

Algorithm 1. Proposed Framework.

Require: A set of labeled training images $\{x_i, l_i\}$, HiC concept, identify unseen classes U and seen classes S.
Ensure: All parameters are set: number of trees R, number of leafnodes per tree, number of topics K and number of unseen class q.
1. Learn RF codebook using $\{x_i, l_i\}$, either using $\mathbb{F}$ codebook, J-CoFi codebook or CoFi codebook (**Property 3.1–3.3**).
2. Build BoW histogram based on the codebook in Step 1.
3. Learn pLSA model using the BoW histogram.
4. Find α_u for each U based on (6).
5. Calculate the signature topic sets $\mathbb{T}_s$ for each S as to (7).
6. Randomly pick α_u to relate u with s in terms of $\mathbb{T}_g \sim \mathbb{T}_u \sim \mathbb{T}_h, (g, h) \in S$.
7. Calculate the signature topic sets $\mathbb{T}_u$ for each U as to (7).
8. Classification for test class c'_m using (9).

IV. RESULTS

In the experiments, we employed four public datasets—PubFig [3], Cifar-100 [21], Caltech-256 [9], and Animals with Attributes (AwA) [1]. These datasets are designed to pose different visual challenges in terms of illumination effects, scales, and viewpoints, as well as support more than 120 000 objects.

Implementation details: In order to evaluate $p(c'_m|d_{\text{test}})$, 1-versus-all classification is performed. Unless specified, the PubFig, Cifar-100, and Caltech-256 dataset features are extracted using the pyramid histogram of gradient (PHOG) with three pyramid levels, $180°$ angle, and 20 bins. Specifically, we use the PHOG from [10] and [22]. However, we did not concatenate all the PHOG descriptors found. Instead, we put all these features in a codebook learning mechanism using the RF algorithm [11], [19]. Therefore, we can obtain a set of HOG descriptors that

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

HOO AND CHAN: ZERO-SHOT OBJECT RECOGNITION SYSTEM BASED ON TOPIC MODEL

5

TABLE III
PubFig Dataset: Performance Evaluation (%) of the Proposed Method in Different Numbers of Unseen Class, q, and Comparison With State-of-the-Art Methods

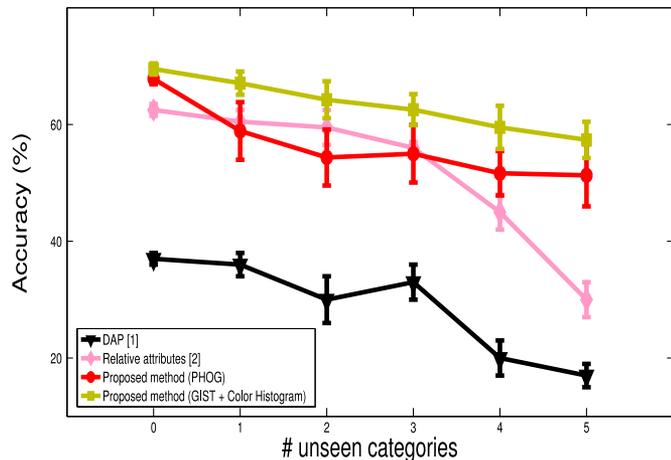| Features	Our Proposed Method Number of Unseen Class, q						Binary Attributes	Relative Attributes
	0	1	2	3	4	5	[1]	[2]
PHOG	**67.83**	58.89	54.99	54.35	51.65	51.30	N/A	N/A
GIST+color histogram	**69.52**	67.09	64.25	62.55	59.52	57.37	37.00	62.00

Fig. 3. Consistency Test: Comparison of the proposed method and the state-of-the-art solutions [1], [2] in the PubFig dataset with different number of unseen categories, q.

TABLE IV
Cifar-100: Comparison of the Proposed Method and the State-of-the-Art Methods in Terms of Accuracy (%)

	Our Proposed Method Number of Unseen Class, q						
	0						
without	with HiC concept				Sparse	Beyond Spatial
HiC concept	J-CoFi	**CoFi**	1	2	Coding [23]	Pyramid [24]
58.13	57.79	**58.21**	56.84	54.85	53.70	54.80

TABLE V
Caltech-256 Dataset: Performance Evaluation (%) of the Proposed Method in Different Numbers of Unseen Class, q

	Number of Unseen Class, q							
	0							
without	with HiC concept							
HiC concept	J-CoFi	CoFi	1	2	3	4	5
67.72	64.60	65.65	52.14	51.49	51.86	52.13	51.32

quantize shape information locally and globally by the nature of the PHOG. The RF codebook can learn image shapes as a whole, as well as the local patch characteristic. For the RF codebook, it is learned using ten trees and 100 leafnodes.

A. PubFig

The PubFig or Public Figures Face Database has a total of 58 797 images of 200 celebrities faces. We used identical subsets as in [2], where eight random identities are extracted with each class of 100 images. The pLSA model is built using $K = 11$, similar to the number of attributes in [2]. In addition to the PHOG features, we also reimplement our framework using features identical to [2], which is a combination of GIST features and color histograms. We employ the class relationship as in [2] to find the $\mathbb{T}_u$. However, the optimum nearest seen classes pair between the unseen classes are chosen, and we assume the ($\succ$) relationship in [2] is similar to our ($\sim$) relationship.

Table III shows that our proposed method has better accuracy (PHOG: 67.83%; GIST + color histogram: 69.52%) compared with that of Lampert *et al.* [1] that uses the binary attributes, and Parikh and Grauman [2] that uses the relative attributes. Our results are achieved without the annotation required in [1] and [2]. When the number of unseen classes q is increased, there is a consistent drop in the system accuracy from 67.83% to 51.30% for PHOG features, and from 69.52% to 57.37% for GIST + color histogram features. This is expected as when the number

of unseen classes increases, the system accuracy decreases due to the tradeoffs between computational complexity and system accuracy.

We performed a consistency test where we tested the accuracy of our proposed method and the method in [1] and [2] across different q. Fig. 3 shows that the proposed method has a better consistency (PHOG: $\pm 7\%$; GIST + color histogram: $\pm 10\%$) in comparison with [1] ($\pm 17\%$) and [2] ($\pm 23\%$). In addition, Lampert *et al.* [1] performed the worst in terms of accuracy, while Parikh and Grauman [2] performed the worst in terms of consistency. Such results have shown the effectiveness and consistency of our proposed algorithm to handle the intraclass variation problem as opposed to the extensive attributes annotation in [1] and [2].

B. Cifar-100

The Cifar-100 [21] dataset has 100 classes, and each class contains 600 images with 32×32 resolutions. The 100 classes are further grouped into 20 Coarse Class. Each $\mathbb{C}$ has five $\mathbb{F}$, where q of them is (are) unseen. Thus, we have a total

TABLE VI
COARSE CLASS ℂ FOR SELECTED CALTECH-256 DATASET

Coarse Class, ℂ	Caltech-256 class (Fine Class, 𝔽)
household electrical devices	binoculars, boom-box, bread maker, calculator, cd, computer keyboard, computer monitor, computer mouse, floppy-disk, head-phones, iPod, joystick, laptop, light bulb, megaphone, microwave, palm-pilot, paper-shredder, PCI-card, photocopier, refrigerator, rotary-phone, toasters, treadmill, tripod, VCR, video-projector, washing machine
household furniture	bathtub, chandelier, chess-board, desk-globe, doorknob, ewer, flashlight, hammock, hot-tub, hourglass, mailbox, mattress, menorah, picnic table
large man-made outdoor things	Buddha, Eiffel-tower, golden-gate-bridge, light-house, minaret, pyramid, skyscraper, smokestack, teepee, tower-Pisa, windmill
medium mammals	dog, duck, elk, goat, goose, llama, minotaur, penguin, porcupine, raccoon, skunk, swan, unicorn, zebra, greyhound
vehicles	blimp, bulldozer, cannon, canoe, car-tire, covered-wagon, fighting-jet, fire-truck, helicopter, hot-air-ballon, kayak, ketch, license-plate, motorbikes, mountain-bike, pram, school-bus, segway, self-propelled-lawn-mower, snowmobile, speedboat, steering-wheel, touring-bike, tricycles, wheelbarrow, airplanes, car-side
household daily items	beer-mug, chopsticks, coffee-mug, knife, spoon, stained-glass, paperclip, paper-shredder, coins, dice, drinking-straw, dumb-bell, fire-extinguisher, frying-pan, ladder, pez-dispenser, playing-card, roulette-wheel, screwdriver, Swiss-army-knife, tweezer, umbrella
sports	baseball-bat, baseball-glove, baseball-hoop, billiards, bowling-ball, bowling-pin, boxing-glove, football-helmet, Frisbee, golf-ball, skateboard, soccer-ball, tennis-ball, tennis-court, tennis-racket, yo-yo
wears	cowboy-hat, diamond-ring, eyeglasses, football-helmet, necktie, sneaker, socks, top-hat, t-shirt, human-wear, wielding-mask, yarmulke, tennis-shoes, saddle, stirrups
musical instruments	electric-guitar, French-horn, grand-piano, guitar-pick, harmonica, harp, harpsichord, mandolin, sheet-music, tambourine, tuning-fork, xylophone

TABLE VII
COARSE CLASS ℂ FOR AWA DATASET

Coarse Class, ℂ	AwA class (Fine Class, 𝔽)
hooves	antelope, horse, moose, ox, sheep, rhinoceros, giraffe, buffalo, zebra, deer, **pig**, cow
weak	Siamese cat, **Persian cat**, skunk, mole, sheep, hamster, rabbit, bat, chihuahua, mouse
grazer	antelope, horse, moose, spider monkey, elephant, ox, sheep, hamster, rhinoceros, rabbit, giraffe, buffalo, zebra, **giant panda**, deer, mouse, cow
stalker	grizzly bear, German shepherd, Siamese cat, tiger, **leopard**, fox, wolf, bobcat, lion, polar bear
flippers	killer whale, blue whale, **humpback whale**, **seal**, otter, walrus, dolphin
strainteeth	killer whale, beaver, blue whale, **hippopotamus**, **humpback whale**, walrus
hibernate	grizzly bear, beaver, skunk, mole, fox, hamster, squirrel, bat, **rat**, bobcat, mouse, polar bear, **raccoon**
bipedal	grizzly bear, spider monkey, gorilla, **chimpanzee**, squirrel, bat, **giant panda**, polar bear

The object classes in **bold** are the predefined unseen classes in [1].

$s = q \times 20$. We picked 30 training images randomly, and the rest are used for testing. In this dataset, we use $K = 10$, as ten major semantic topics exist in ℂ, i.e., mammals, size, trees, vehicles, food, household, insects, reptiles, people, and flowers. The dataset is challenging due to its limited resolution, and therefore, we only use two pyramid levels for PHOG features and 50 codewords per tree in codebook learning.

Table IV shows that our proposed method with or without the HiC concept performed much better as compared with [23] and [24]. Our approach also outperformed [23], [24], when $q = 2$. When $q = 2$, there is a total of 40 unseen 𝔽 when training the classifier. However, our approach was still able to achieve 54.85% accuracy in comparison with 53.70% [23] and 54.80% [24] where in both approaches, $q = 0$ (no unseen classes). In addition, the computational cost of our proposed method is lower, as we only employed a small number of training images.

Similar to the PubFig dataset, we also observed that when using fewer seen classes in the learning process, the accuracy drops. However, the accuracy differences between $q = 1$ and $q = 2$ only differ by a fraction of $\pm 2\%$ even when the difference number of u is large (the total unseen class here is $q \times 20$). This indicates that our proposed method is robust as it is ca-

pable to handle the Cifar-100 dataset with very tiny (30×30) images that causes the collected features vector to be very similar. Besides, in comparison with the three different codebook learning strategies, the CoFi codebook method performs the best as it utilized both ℂ$_i$ and 𝔽$_i$ simultaneously in the RF tree node splitting.

C. Caltech-256

The Caltech-256 dataset [9] consists of 30 607 images grouped into 256 object classes and a background class. Unfortunately, it does not provide any ℂ concepts in the dataset. Therefore, we group the classes manually to ℂ similar to Cifar-100, except for some specific classes where we introduce new ℂ. In Table VI, we show the distribution of the selected Caltech-256 classes with five existing ℂ as in Cifar-100 and four newly introduced ℂ. Only 158 of the total Caltech-256 classes are grouped because some object categories belong to a ℂ that had very few 𝔽 members. For this dataset, the total u are $q \times 9$.

Table V shows minor fluctuations compared with the PubFig and Cifar-100 results when different q values are employed. For classification settings that have $q = 0$, interestingly, the

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

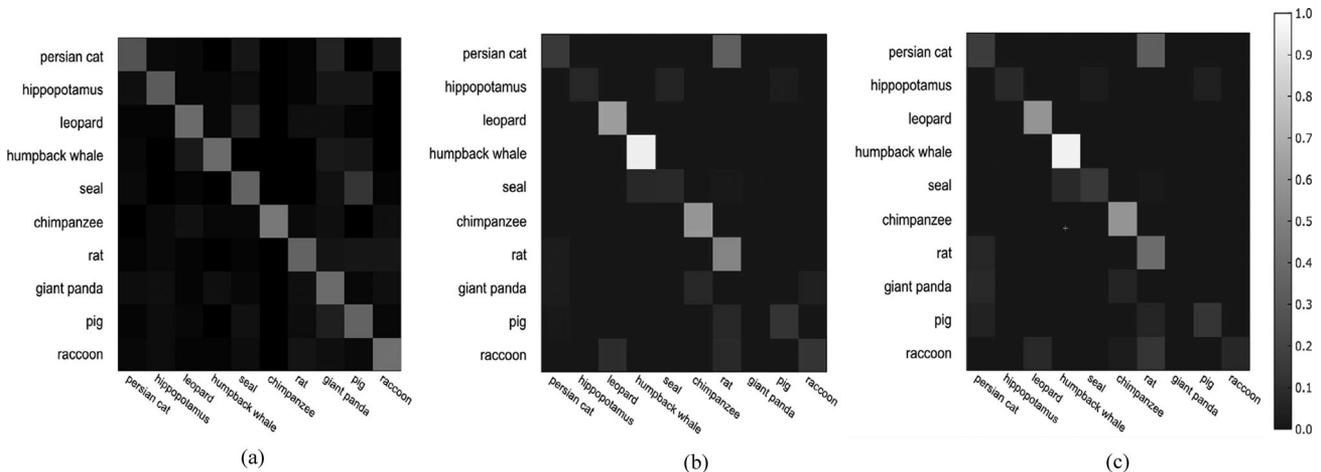HOO AND CHAN: ZERO-SHOT OBJECT RECOGNITION SYSTEM BASED ON TOPIC MODEL 7

Fig. 4. AwA Dataset: Confusion matrix between our proposed method, IAP [13], and DAP [13].

proposed method performs better without applying the HiC concept. We found that this may be due to 1) $\mathbb{F}$ in some $\mathbb{C}$ are semantically related but have low visual similarity (i.e., in this context, the visual similarity is referring to the visual appearance of the object class), e.g., "computer keyboard," "computer monitor," and "computer mouse," which belong to the $\mathbb{C}$ = "household electrical devices"; 2) introducing $\mathbb{C}$ tree to the codebook did not help in boosting the codebook discriminating power, which might be due to the low visual similarity among $\mathbb{F}$ in some $\mathbb{C}$ as well; and 3) the complexity of the objects in Caltech-256. However, the zero-shot learning still provides reasonable results. For this dataset, we did not perform any comparisons as there are only 158 classes extracted.

D. Animal With Attributes

AwA is an object dataset of animal classes with corresponding attributes attached to each class. There are a total of 50 animal classes and 85 attributes in the dataset. We use similar experimental settings as in [1] and [13], where same features and partitions of seen (i.e., 40) and unseen (i.e., 10) classes were employed. To build the $\mathbb{C}$ and $\mathbb{F}$ relationships, we adopt the attributes relationships and pick the attributes that have $q \leq 2$ and have the lowest number of $\mathbb{F}$ possible. As a result, we grouped eight $\mathbb{C}$ and each $\mathbb{C}$ has 6 to 12 $\mathbb{F}$, as shown in Table VII.

Herein, our proposed method achieved an accuracy of 49.65%. This is better as compared with the DAP and IAP [13], which achieved 41.4% and 42.2%, respectively, to M2LATM [15] that obtain 41.3%, and attribute/hierarchical label embedding [25] that achieved 43.5%. We also show the confusion matrix of the ten test classes in Fig. 4. We observe that our proposed method has better average classification results compared with DAP and IAP [13]. Although our proposed method does not predict the "humpback whale" class as well as DAP and IAP, but we achieve better accuracy in the "giant panda" class, which leads to better overall accuracy. These results benefit from the HiC concept defined for AwA, where "giant panda" class is the only u in $\mathbb{C}$ = {"grazer"}. Note that the "humpback whale" class resides in $\mathbb{C}$ = {"flippers," "strainteeth"},

and both $\mathbb{C}$ contains more than one u. Therefore, we observe the accuracy drops in the "humpback whale" class. The same situation applies to the "rat" class and "raccoon" class as they share the same $\mathbb{C}$ = {"hibernate"} .

V. DISCUSSION

In this paper, we compared our proposed method with four public datasets and achieved better performance compared with state-of-the-art methods for zero-shot learning. Even in the conventional classification problem where training images for all object classes are available, we still manage to get state-of-the-art accuracy in PubFig and Cifar-100 datasets.

In the conducted experiments, there are some cases where the predicted $\mathbb{T}_u$ is redundant. That is, if a lower number of K is chosen, the numbers of possible M will be reduced as well, and hence, there is a possibility to obtain similar $\mathbb{T}_u$ for different u, which is a redundant representation. In order to handle this issue, we employ a large number of K in the experiments to minimize the probability of $\mathbb{T}_c$ to be redundant.

Based on the experiments in Caltech-256 dataset, we are aware that the classification accuracy is fluctuating due to the quality of $\mathbb{F}$ collection under each $\mathbb{C}$. Although the $\mathbb{F}$ within $\mathbb{C}$ is grouped based on the semantic relationship, these $\mathbb{F}$ might be visually dissimilar. This limitation is likely to be solved by introducing a middle-level class group to extend the $\mathbb{F}$ within the $\mathbb{C}$ to some high-visual similarity group, e.g., we can group and extend $\mathbb{F}$: "head-phones," "rotary-phones," and "megaphone" in $\mathbb{C}$: "household electrical devices" to $\mathbb{C}$: "phones." When we pick the random α_u to model u = {"megaphone"}, the "head-phones" and "rotary-phones" will have priority as the related s. Nonetheless, our future work includes introducing tighter relationship between the Fine Class and the Coarse Class so that better performance can be achieved.

REFERENCES

[1] C. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogit.*, 2009, pp. 951–958.

[2] D. Parikh and K. Grauman, "Relative attributes," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 503–510.

[3] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 365–372.

[4] M. Rohrbach, M. Stark, and B. Schiele, "Evaluating knowledge transfer and zero-shot learning in a large-scale setting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 1641–1648.

[5] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, M. Ranzato, and T. Mikolov, "Devise: A deep visual-semantic embedding model," in *Proc. Adv. Neural Inform. Process. Syst.*, 2013, pp. 2121–2129.

[6] T. Mensink, J. Verbeek, F. Perronnin, and G. Csurka, "Metric learning for large scale image classification: Generalizing to new classes at near-zero cost," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 488–501.

[7] W. L. Hoo, and C. S. Chan, "PLSA-based zero-shot learning," in *Proc. 20th IEEE Int. Conf. Image Process.*, 2013, pp. 4297–4301.

[8] L. Fei Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, vol. 2, pp. 524–531.

[9] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," California Inst. Technol., Pasadena, CA, USA, Caltech Tech. Rep. CNS-TR-2007-001., 2007.

[10] A. Bosch, A. Zisserman, and X. Muoz, "Image classification using random forests and ferns," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

[11] F. Moosmann, E. Nowak, and F. Jurie, "Randomized clustering forests for image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 9, pp. 1632–1646, Oct. 2008.

[12] V. Ferrari and A. Zisserman, "Learning visual attributes," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 433–440.

[13] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 453–465, Mar. 2014.

[14] Y. Fu, T. Hospedales, T. Xiang, and S. Gong, "Attribute learning for understanding unstructured social activity," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 530–543.

[15] Y. Fu, T. Hospedales, T. Xiang, and S. Gong, "Learning multimodal latent attributes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 303–316, Feb. 2014.

[16] M. Guillaumin, J. Verbeek, and C. Schmid, "Multimodal semi-supervised learning for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit..*, 2010, pp. 902–909.

[17] M. Palatucci, D. Pomerleau, G. Hinton, and T. Mitchell, "Zero-shot learning with semantic output codes," in *Proc. Advances Neural Informat. Process. Syst.*, 2009, pp. 1410–1418.

[18] C. Silberer, V. Ferrari, and M. Lapata, "Models of semantic representation with visual attributes," in *Proc. 51th Annu. Meet. Assoc. Comput. Linguistics*, 2013, pp. 572–582.

[19] W. L. Hoo, T.-K. Kim, Y. Pei, and C. S. Chan, "Enhanced random forest with image/patch-level learning for image understanding," in *Proc. 22nd Int. Conf. Pattern Recognit.*, 2014, pp. 3434–3439.

[20] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman, "Discovering objects and their location in images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 370–377.

[21] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Comput. Sci. Dept., Univ. Toronto, Tech. Rep., 2009.

[22] A. Bosch, A. Zisserman, and X. Munoz, "Representing shape with a spatial pyramid kernel," in *Proc. ACM Int. Conf. Image Video Retrieval*, 2007, pp. 401–408.

[23] I. Goodfellow, A. Courville, and Y. Bengio, "Large-scale feature learning with spike-and-slab sparse coding," in *Proc. Int. Conf. Mach. Learn.*, 2012, pp. 1439–1446.

[24] Y. Jia, C. Huang, and T. Darrell, "Beyond spatial pyramids: Receptive field learning for pooled image features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3370–3377.

[25] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Label-embedding for attribute-based classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 819–826.

Wai Lam Hoo (S'11) received the B.Sc. (Hons.) degree in computer science, majoring in artificial intelligence from the University of Malaya, Kuala Lumpur, Malaysia, in 2010, where he is currently working toward the Ph.D. degree.

He was with the Imperial College, University of London, London, U.K., as a Visiting Researcher. His research interests include computer vision, image processing, and scene understanding.

Chee Seng Chan (S'05–M'09–SM'14) received the Ph.D. degree from the University of Portsmouth, Portsmouth, U.K., in 2008.

He is currently a Senior Lecturer with the Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia. In general, his major research interests include fuzzy qualitative reasoning and computer vision, with a focus on image/video content analysis and human–robot interaction.