Learning Models for Object Recognition from Natural Language Descriptions

Josiah Wang
scs6jwks@comp.leeds.ac.uk
Katja Markert
markert@comp.leeds.ac.uk
Mark Everingham
me@comp.leeds.ac.uk

School of Computing
University of Leeds
Leeds, UK

Abstract

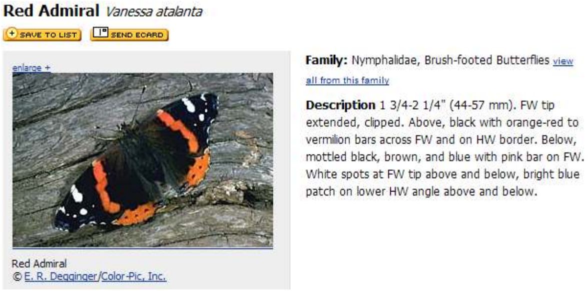
We investigate the task of learning models for visual object recognition from natural language descriptions alone. The approach contributes to the recognition of fine-grain object categories, such as animal and plant species, where it may be difficult to collect many images for training, but where textual descriptions of visual attributes are readily available. As an example we tackle recognition of butterfly species, learning models from descriptions in an online nature guide. We propose natural language processing methods for extracting salient visual attributes from these descriptions to use as 'templates' for the object categories, and apply vision methods to extract corresponding attributes from test images. A generative model is used to connect textual terms in the learnt templates to visual attributes. We report experiments comparing the performance of humans and the proposed method on a dataset of ten butterfly categories.

1 Introduction

Recent years have seen great advances in object category recognition by the use of machine learning approaches rather than hand-built models. However, progress in this area is limited by the onerous task of manually collecting and labelling large training sets. The problem is compounded by the need to collect a training set for each new category to be recognised, making it difficult to scale up from the 10-100's of categories currently tackled [8, 12] to the 1000's which are needed to cover the space of object categories recognised by humans. These challenges have led to much interest in exploiting other existing sources of image annotation, for example keywords [5] or image captions on web pages [3, 10, 17, 19]. While most previous work has either assumed the availability of pre-processed keywords [5] or used only simple text processing methods ignoring syntax or high-level semantics of text, *e.g*. searching Google Images for "car" or "voiture" to find example car images [10, 17, 19], in this paper we consider the use of Natural Language Processing (NLP) methods to extract information beyond keywords from *naturally occurring* English text.

For 'top-level' object categories (person, car, cat, dog, *etc*.) it is difficult to find images with corresponding textual descriptions of *visual* properties, for example we can find many images annotated "my car" but few annotated "my Ford car which can be recognised by it being blue, having two wheels visible and a red stripe down the side". However, for *fine-grain*

BMVC 2009 doi:10.5244/C.23.2



Red Admiral *Vanessa atalanta*

Family: Nymphalidae, Brush-footed Butterflies view all from this family

Description 1 3/4-2 1/4" (44-57 mm). FW tip extended, clipped. Above, black with orange-red to vermilion bars across FW and on HW border. Below, mottled black, brown, and blue with pink bar on FW. White spots at FW tip above and below, bright blue patch on lower HW angle above and below.

Red Admiral
© E. R. Degginger/Color-Pic, Inc.

Figure 1: Example visual description from eNature for the Red Admiral butterfly *Vanessa atalanta*. The question we investigate in this paper is whether a computer can learn to recognise this species of butterfly from the textual description alone, and indeed can humans?

categories such as animal or plant species, detailed *visual* descriptions are readily available in the form of online nature guides. An example of the level of description available in such sources is shown in Figure 1. These descriptions are different from image captions or conventional dictionary definitions in providing definitions of *visual* rather than *semantic* properties of objects. Properties described are often both detailed and discriminative, including aspects such as colours, shapes and patterns which are not found in 'casual' annotation of images where the content is clear from the image itself.

We investigate the task of learning to recognise fine-grain categories, using species of butterflies as an example. Addressing such categories is important in the aim of increasing object recognition abilities beyond the current number of categories, and is particularly salient as it becomes harder to find many example images as we move from coarse to fine-grain categories. The method we propose here uses NLP methods to automatically extract salient visual properties from naturally occurring English descriptions taken from an online nature guide [6]. We investigate two main questions: (i) can a computer learn a model to recognise each category from these textual descriptions alone? and (ii) can humans? We learn models to recognise ten categories (species) of butterflies with *no* example images. We compare our method against human performance given the same textual descriptions, and obtain accuracy comparable to that of a non-native English speaker.

Related work. A number of authors have investigated using single keyword searches to find example training images from the world wide web [2, 10, 17, 19], and refining results using vision. Duygulu *et al*. [5] investigated learning from images annotated with a *set* of keywords, posing the task as one of machine translation between 'visual' and textual words. Gupta and Davis [13] have recently investigated using prepositions and comparative adjectives *e.g*. "sky above road" in a machine translation approach. While certainly an advance over using only keywords (nouns), as noted such annotation is not readily available, and indeed the authors manually annotated their training images [13].

The use of text has also been investigated for learning recognition of individuals [3, 7]. Berg *et al*. [3] use linguistic cues including part-of-speech (PoS) tagging to find proper nouns in text from news pages accompanying images. Everingham *et al*. [7] extract annotation from automatically-aligned video subtitles and scripts, but use no linguistic processing. Work by Laptev *et al*. [15] has also exploited script text, training classifiers to identify 'actions' in the text to find training data for activity recognition.

Two pieces of work concurrent to ours [9, 14] have investigated recognition of object

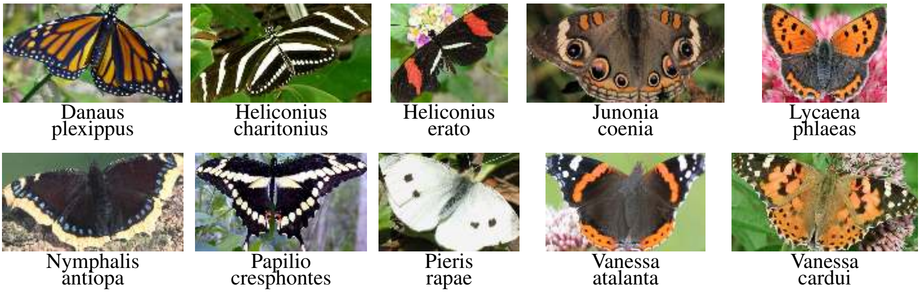


Figure 2: Ten species of butterflies used in the experiments.

categories from named ‘attributes’ *e.g.* “black, furry, hooves”, supporting learning of new object categories without example images. Farhadi *et al.* [9] train binary classifiers for a set of 64 attributes and use these predicted attributes as features for object category recognition. Lampert *et al.* [14] recognise new animal categories without training images by sharing attributes learnt from other categories. Our work differs in that we extract attributes directly from text, rather than requiring these to be specified manually [9, 14], and in that we investigate recognition of fine-grain categories, for which such text is readily available.

Overview. In Section 2 we describe the textual/image datasets collected for training and testing respectively. Section 3 describes the NLP components of our method, which extract a ‘template’ model of a category from a textual description. Section 4 covers the extraction of visual attributes from a test image, and Section 5 the model used to ‘score’ an image against the learnt template models. In Section 6 we report results of our proposed method and human performance, and Section 7 offers conclusions.

2 Dataset

For the experiments reported here we selected ten categories (species) of butterflies. Butterflies are a salient domain to investigate since visually they have both distinctive and common properties (see Figure 2). Compared to ‘top-level’ categories (person, car, bicycle, *etc.*) there are few ‘global’ features, *e.g.* configuration of parts, which can be used to discriminate between species. To our knowledge, one previous work [16] has investigated visual recognition of butterfly species, by matching local invariant features to training images. We collected natural text descriptions for each category from the eNature online nature guide [6]; examples of the descriptions can be found in Figures 1 & 3 and are discussed further in Section 3.

For each of the ten categories we collected images to use as a *test* set from Google Images by querying with the Latin name of the species *e.g.* “*Danaus plexippus*”. The images returned were manually filtered for those actually depicting the butterfly of interest. The dataset comprises 832 images in total, with the distribution ranging from 55 to 100 images per category. Figure 2 shows an example image for each category. As can be seen, several categories are challenging to distinguish, sharing orange/black colours. There is also considerable variation in illumination and pose.

3 Natural language processing

Our NLP approach extracts models from the textual descriptions obtained from eNature (Figure 1). Although full natural language understanding (such as full summarisation of hu-

Heliconius charitonius: 3-3 3/8" (76-78 mm). Wings long and narrow. Jet-black above, banded with lemon-yellow (sometimes pale yellow). Beneath similar; bases of wings have crimson spots.

Junonia coenia: 2-2 1/2" (51-63 mm). Wings scalloped and rounded except at drawn-out FW tip. Highly variable. Above, tawny-brown to dark brown; 2 orange bars in FW cell, orange submarginal band on HW, white band diagonally crossing FW. 2 bright eyespots on each wing above: on FW, 1 very small near tip and 1 large eyespot in white FW bar; on HW, 1 large eyespot near upper margin and 1 small eyespot below it. Eyespots black, yellow-rimmed, with iridescent blue and lilac irises. Beneath, FW resembles above in lighter shades; HW eyespots tiny or absent, rose-brown to tan, with vague crescent-shaped markings.

Figure 3: Example textual descriptions obtained from eNature. Textual descriptions range from brief to elaborate descriptions.

Description

1 3/4-2 1/4" (44-57 mm). FW tip extended, clipped. Above, black with orange-red to vermilion bars across FW and on HW border. Below, mottled black, brown, and blue with pink bar on FW. White spots at FW tip above and below, bright blue patch on lower HW angle above and below.

Ground truth template		**Learnt template**	
above fw colour	black	above fw colour	black
above fw pattern	[red to vermilion] bars	above fw pattern	[red to vermilion] bars
above fwm colour		above fwm colour	
above fwm pattern	[white] spots	above fwm pattern	[white] spots
above hw colour	**black**	**above hw colour**	
above hw pattern		above hw pattern	
above hwm colour		**above hwm colour**	**black**
above hwm pattern	[red to vermilion] bars; [blue] patch	above hwm pattern	[red to vermilion] bars; [blue] patch
below fw colour	black brown blue	below fw colour	black brown blue
below fw pattern	[pink] bar	below fw pattern	[pink] bar
below fwm colour		below fwm colour	
below fwm pattern	[white] spots	below fwm pattern	[white] spots
below hw colour		below hw colour	
below hw pattern		below hw pattern	
below hwm colour		below hwm colour	
below hwm pattern	[blue] patch	below hwm pattern	[blue] patch

Figure 4: An example template, for *Vanessa atalanta*. *fw* and *hw* stand for Forewing and Hindwing respectively, and *fwm* and *hwm* for Forewing Margin and Hindwing Margin respectively. The left column shows the manually-filled ground truth template, while the right shows the template automatically filled by our method. The disagreements between entries in the ground truth and automatically filled templates are shown in bold.

man argumentation, for example) is beyond current NLP capabilities, considerable progress has been made in the more limited task of *information extraction*, which turns unstructured but local and factual information in text into structured data, so-called *templates*. As the eNature texts tend to describe standard butterfly properties (such as colours and patterns) in relatively stereotypical fashion, it is possible to both design templates describing the information needed and to automatically fill these templates.

Templates. The templates used act as models of the visual attributes of a butterfly from textual descriptions. They contain slots for various butterfly attributes including colours, patterns and their location (such as the forewing or hindwing of the butterfly). An example template is shown in Figure 4. Once the templates are filled automatically from the textual descriptions, they can be used as models for our proposed classifier.

Template filling. We propose a framework to extract salient attributes from the butterfly descriptions and to fill our templates automatically. To ensure resilience across different descriptions from different sources, the text is processed with a generic and modular pipeline as shown in Figure 5. The text is first tokenised into a list of word tokens. The sequence of

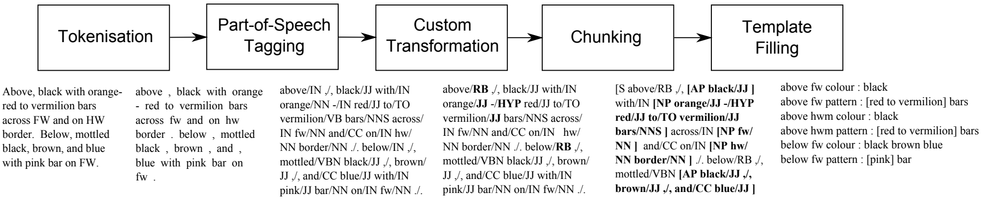

Tokenisation
Part-of-Speech Tagging
Custom Transformation
Chunking
Template Filling
Above, black with orange-red to vermilion bars across FW and on HW border. Below, mottled black, brown, and blue with pink bar on FW.
above , black with orange - red to vermilion bars across fw and on hw border . below , mottled black , brown , and , blue with pink bar on fw .
above/IN ,/, black/JJ with/IN orange/NN -/IN red/JJ to/TO vermilion/VB bars/NNS across/IN fw/NN and/CC on/IN hw/NN border/NN ./. below/IN ,/, mottled/VBN black/JJ ,/, brown/JJ ,/, and/CC blue/JJ with/IN pink/JJ bar/NN on/IN fw/NN ./.
above/**RB** ,/, black/JJ with/IN orange/**JJ** -/**HYP** red/JJ to/TO vermilion/**JJ** bars/NNS across/IN fw/NN and/CC on/IN hw/NN border/NN ./. below/**RB** ,/, mottled/VBN black/JJ ,/, brown/JJ ,/, and/CC blue/JJ with/IN pink/JJ bar/NN on/IN fw/NN ./.
[S above/RB ,/, **[AP black/JJ]** with/IN **[NP orange/JJ -/HYP red/JJ to/TO vermilion/JJ bars/NNS]** across/IN **[NP fw/NN]** and/CC on/IN **[NP hw/NN border/NN]** ./. below/RB ,/, mottled/VBN **[AP black/JJ ,/, brown/JJ ,/, and/CC blue/JJ]**
above fw colour : black
above fw pattern : [red to vermilion] bars
above hwm colour : black
above hwm pattern : [red to vermilion] bars
below fw colour : black brown blue
below fw pattern : [pink] bar

Figure 5: Pipeline for converting textual descriptions into templates. The input text is first divided into tokens, then a Part-of-Speech (PoS) tagger computes PoS tags for each token. The tags are modified by a list of rules to adapt to the specific style of the eNature descriptions. Chunking is then performed to extract noun phrases (NP) and adjective phrases (AP). Finally a template is filled by matching the resulting 'chunks' against a list of colours, patterns and location terms.

tokens is then tagged with a part-of-speech (PoS) tagger, which labels each token in a given text with a PoS tag, *e.g.* nouns (NN), adjectives (JJ) or verbs (VB). We used the freely available C&C tagger [4], which is a state-of-the-art tagger trained and used for newspaper texts (on which most current taggers operate). The eNature butterfly descriptions, however, have some genre-specific properties, such as a tendency to suppress subjects ("*Dark with pale margins.*") and verbs ("*White spots at FW tip.*"). As no tagged corpus of nature guides exists which would allow us to retrain the tagger for that new genre, a few custom transformations adapt the output of the tagger to correct any known mistakes. These include handling of butterfly-specific terms such as forewings and hindwings, changing all tokens matching a predefined list of colours to adjectives and tagging all "above" and "below" as adverbs ("*Above, black with red bars*") unless they occur before an adjective or a noun where they are tagged as prepositions ("*red above border*").

Partial parsing is performed on the tagged text by chunking [1], which extracts 'chunks' of text matching a pre-specified tag sequence. We extract noun phrases ("*wing* has *blue spots*") and adjective phrases ("wings are *black*"). These extracted phrases are then filtered through a list of colours and patterns, and mapped onto one or more template slots using simple heuristics. Thus, the sentence containing the extracted attribute is searched for a predefined list of location terms to decide on its location (top or bottom wings). Similarly, a list of location terms such as "margin" is matched with surrounding words within the clause containing the attribute to determine whether the attribute is located on the forewing or hindwing, and whether it is located on either wing's margin. Figure 4 shows an example output template filled by our method and the comparison to the corresponding ground truth template filled by hand. Over the ten categories the template-filling gives recall of 83% and precision of 78%, where an entry is considered correct only if the location (above/below and forewing/hindwing), colour and pattern terms match exactly.

4 Visual processing

This section describes the visual processing component of our approach, which extracts visual attributes of an image for matching against the models learnt from text. Our method bases recognition on two simple visual attributes determined salient from the textual descriptions: (i) dominant (wing) colour; (ii) coloured spots.

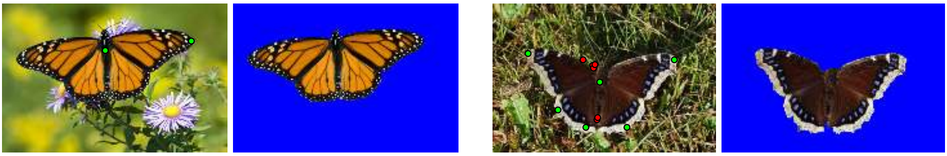


Figure 6: Examples of semi-automatic segmentation: input image (left) and segmentation result (right). Green and red points on the input image indicate the foreground/background points specified by the operator. The operator clicked 2/0 and 6/6 foreground/background points for the left and right images respectively.

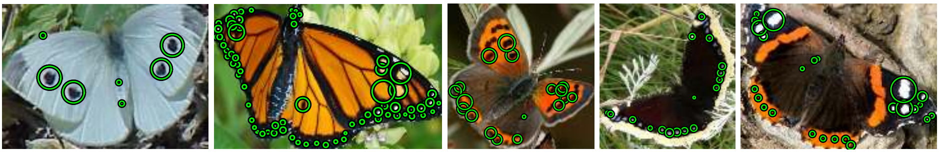


Figure 7: Examples of spot detection. For clarity, detected spots are shown at 3 times the scale of detection. Several false positives are visible, for example on the abdomen in the leftmost image.

Image segmentation. In early experiments we found that significant challenges are posed by variation in the background of the image, which are hard to overcome without training images. Therefore as a first processing step the butterfly is segmented from the background. As a pragmatic decision we use a semi-automated approach – learning a generic butterfly model to support fully automatic segmentation would require many training images, and is left for future work. We use the 'star shape' graph-cut approach proposed by Veksler [21]. In this method, pixels are assigned to foreground/background by solving a graph-cut constrained such that the resulting segmentation is a 'star shape' – from a given centre point the foreground shape can be described by a set of rays from that point which cross the foreground/background boundary exactly once; butterflies approximately satisfy this shape property. To segment the image, the operator must specify the centre of the butterfly, and can add additional points constraining parts of the image to be foreground or background. Figure 6 shows example segmentations obtained with the constraint points marked (total of 2 and 12 for each image respectively). Over the entire dataset a median of 8 points per image were marked to obtain high quality segmentations like the ones shown. We consider this an acceptable level of supervision for applications *e.g.* a mobile nature field guide, and the semi-automated approach avoids having to replicate humans' prodigious experience and ability in segmentation.

Spot detection. The presence of spots of a particular colour is a strong cue to identifying butterfly species. Our method detects spots using a two step approach. First, candidate image regions likely to be spots are extracted by applying the Difference-of-Gaussians (DoG) interest point operator [18] to the image at multiple scales. In a second stage, image descriptors are extracted around each candidate spot, and classified as 'spot' or 'non-spot'. As descriptors we use the SIFT descriptor [18] computed at three consecutive octave scales around the interest point. A linear classifier trained using logistic regression is used. Figure 7 shows example spot detections; by setting a high threshold on the classifier few false positive detections are observed, at the cost of some missed detections.

In order to train the spot classifier, training images are required. One approach [11] would be to use images obtained from a web search *e.g.* using the keyword “spots” [11]. However, we find that even a concept as simple as ‘spot’ is domain-specific, for example spots on butterflies little resemble the images returned by Google Images which include graphic designs, skin conditions and sunspots. We therefore trained a ‘butterfly spot’ detector using hand-marked butterfly images. Note, however, that no *class* information *i.e.* the category of butterflies was used in this training process.

Colour modelling. Our learnt templates contain the *names* of dominant (wing) and spot colours for a given butterfly category. In order to connect these names to image observations, models for each colour name are required. As in the case of spots, we found that querying Google Images for colour names [20] in order to learn colour models gave poor results, and so we instead learnt colour models specific to the butterfly domain. Differences observed include ‘yellow’ which is used for anything from off-white to orange colours in butterfly descriptions, compared to the canonical ‘buttercup’ hue. For each colour name c_i a probability distribution $p(\mathbf{z}|c_i)$ was learnt from training images, where $\mathbf{z}$ is a pixel colour observation in the perceptually uniform $L*a*b*$ space. The distribution is modelled using a simple Parzen density estimator with a Gaussian kernel. Note that as in the spot model, no class information is used to learn the colour models.

5 Generative model

Given an input image the task of predicting the category of butterfly depicted is cast as one of Bayesian inference, using a generative model for each of the ten butterfly categories. Denoting the image I, the probability of observing that image given the butterfly category B_i is defined as a product over spot observations S and ‘wing’ (other) colour observations W:

$$p(I|B_i) = p(S|B_i)p(W|B_i) \tag{1}$$

We assume that spots are generated independently with colour according to a prior over colour names specific to category B_i:

$$p(S|B_i) = \prod_j \sum_k p(\mathbf{z}_j^s|c_k^s)P(c_k^s|B_i) \tag{2}$$

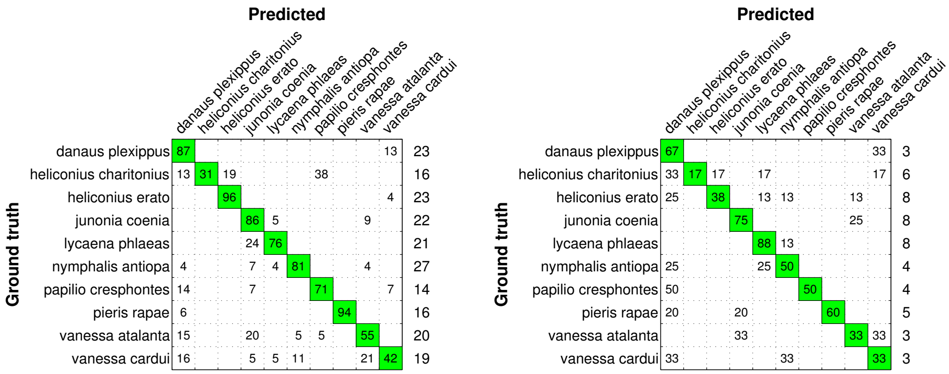
where $\mathbf{z}_j^s$ is the observed $L*a*b*$ colour of spot j, and $p(\mathbf{z}|c)$ is one of the models relating measurements to colour names, as detailed above. Note that we marginalise over the category-specific spot colour name prior $P(c_k^s|B_i)$. The *dominant* colour of the butterfly is captured by assuming non-spot pixels of the image to be generated from a second category-specific colour name distribution $P(c_k^w|B_i)$:

$$p(W|B_i) = \prod_j \sum_k p(\mathbf{z}_j^w|c_k^w)P(c_k^w|B_i) \tag{3}$$

where $\mathbf{z}_j^w$ are $L*a*b*$ observations of non-spot pixels, and again we marginalise over the colour name distribution.

To specify the generative model requires two prior distributions $P(c_k^s|B_i)$ and $P(c_k^w|B_i)$ over spot and dominant colour names respectively. These are specified by converting the templates learnt from the textual descriptions (Section 3).

Spot colour name prior. For the spot model, the set of colour names in the template is converted to a prior by assigning equal probability to each colour in the template. As an example, if the template contains white spots and black spots, we assume that the probability of each is 0.5. The prior is regularised to avoid zero probabilities by adding a small constant.



	danaus plexippus	heliconius charitonius	heliconius erato	junonia coenia	lycaena phlaeas	nymphalis antiopa	papilio cresphontes	pieris rapae	vanessa atalanta	vanessa cardui	
danaus plexippus	87									13	23
heliconius charitonius	13	31	19				38				16
heliconius erato			96							4	23
junonia coenia				86	5				9		22
lycaena phlaeas				24	76						21
nymphalis antiopa	4			7	4	81			4		27
papilio cresphontes	14			7			71			7	14
pieris rapae	6							94			16
vanessa atalanta	15			20		5	5		55		20
vanessa cardui	16			5	5	11			21	42	19

(a) Native English speakers

	danaus plexippus	heliconius charitonius	heliconius erato	junonia coenia	lycaena phlaeas	nymphalis antiopa	papilio cresphontes	pieris rapae	vanessa atalanta	vanessa cardui	
danaus plexippus	67									33	3
heliconius charitonius	33	17	17		17					17	6
heliconius erato	25		38		13	13			13		8
junonia coenia				75					25		8
lycaena phlaeas					88	13					8
nymphalis antiopa	25				25	50					4
papilio cresphontes	50						50				4
pieris rapae	20			20				60			5
vanessa atalanta				33					33	33	3
vanessa cardui	33					33				33	3

(b) Non-native English speakers

Figure 8: Results of the human experiment. Confusion matrices are shown for (a) native English speakers and (b) non-native English speakers. Numbers in cells are the percentage classification/misclassification rates. The overall balanced accuracy is 72% for native English speakers and 51% for non-native English speakers.

Dominant colour name prior. To model the dominant colour attribute in our learnt templates, we need to convert this to a prior over colour names representing the concept that we expect 'most of' the butterfly pixels to be generated by this colour model. This is captured by defining the prior $P(c_k^w|B_i)$ as a mixture of two components:

$$P(c_k^w|B_i) = \alpha P(c_k^w|\Theta_i^d) + (1-\alpha)P(c_k^w|\Theta_i^o) \qquad (4)$$

where $P(c_k^w|\Theta_i^d)$ denotes the prior over colour names for the dominant colour, which is set to 1 for the corresponding template colour, and zero for all others; $P(c_k^w|\Theta_i^o)$ denotes the prior over colour names for 'other' colours and is set uniformly for all other colour names appearing in the template. As an example, if the template contains a dominant colour of 'orange' and other colours 'black' and 'white' then $P(c_k^w|\Theta_i^d)$ is 1 for orange and zero for all other colours, and $P(c_k^w|\Theta_i^o)$ is 0.5 for black and white.

The parameter α controls how much of the image is expected to be explained by the dominant colour. Rather than setting this to an arbitrary value we define a hyper-prior over its value and marginalise. We use a Beta distribution with parameters $a = 8$ and $b = 4$. This captures that we expect on average around $2/3$ of the image to be the dominant colour, and with 90% probability at least 50% of the image.

Classification. For each butterfly category, the corresponding learnt template is converted to prior distributions over colour names as described above. By evaluating the likelihood of the image $p(I|B_i)$ under each model, the image is then classified by assigning it the category B_i which maximises the likelihood – lacking further information, we assume that the prior over butterfly categories $P(B_i)$ is uniform.

6 Experimental results

We report here the results of two sets of experiments: (i) performance of humans on the task of recognising butterflies from textual descriptions; (ii) performance of our proposed method.

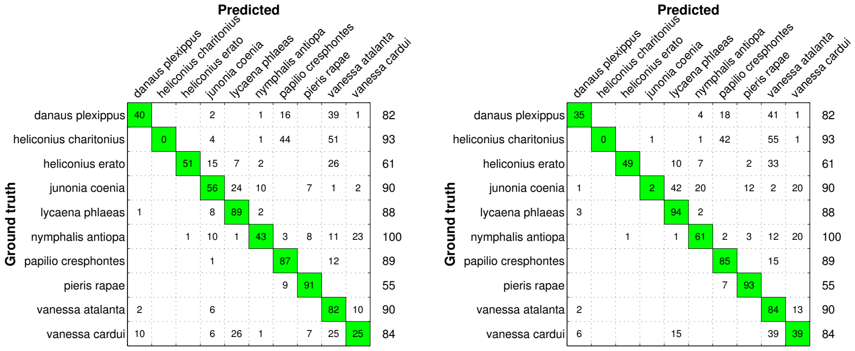
		Balanced Accuracy
Human	Non-native English speakers	51.0%
	Native English speakers	72.0%
Ground truth templates	Spot colours only	39.1%
	Dominant colours only	40.0%
	Spots + Dominant colours	**56.3%**
Learnt templates	Spot colours only	39.1%
	Dominant colours only	35.3%
	Spots + Dominant colours	**54.4%**

Table 1: Results of (i) use of ground truth *vs.* automatically filled templates; (ii) combination of visual attributes. The results show moderate reduction in accuracy arising from errors in the automatic template filling. The individual visual attributes (spots and dominant colour) perform similarly to each other, with the combination improving accuracy substantially.

Human performance. Human performance was measured in order to establish the difficulty of recognising butterflies from textual descriptions, since humans might reasonably be considered to exhibit 'upper-bound' performance on this task. The experiment was conducted via a web-page, with participants mainly drawn from Computer Science staff and students at the University of Leeds. The web-page displayed (i) the description of one of the ten categories, picked at random; (ii) a set of ten images of butterflies, one per category, again picked at random. Participants were requested to select the image of the butterfly described. In addition, participants were asked to indicate whether they were native English speakers, and if they considered themselves an 'expert' on butterflies. Each participant was limited to a *single* trial to prevent learning from the images. Participants took 60–350 seconds (10th/90th percentiles) to complete the task, reflecting its challenging nature.

There were 253 participants in total, comprising 201 native and 52 non-native English speakers. Butterfly 'experts' were excluded from the experiment. Figure 8 shows the confusion matrices obtained across all participants. The balanced accuracy for native English speakers is 72% while the accuracy for non-native English speakers is much lower at 51%. The results show the difficulty of the task at hand even for humans. The category *Heliconius charitonius* (see Figure 2) proved most difficult to recognise from the provided description for both native (31% accuracy) and non-native English speakers (17% accuracy). In this case the poor accuracy can be accounted for by two deficiencies in the description (Figure 3, top): (i) the description is brief, including little discriminatory information; (ii) the "lemon-yellow" bands described in the text most often appear white in the image, and are confused with the white spotted bands of *Papilio cresphontes* (see Figure 2).

Proposed method. The proposed method was evaluated on the same butterfly dataset used in the human experiment. We first compare classification results using models built from the learnt templates and ground truth templates respectively, to examine the effect of errors in the templates arising from the imperfect NLP methods. Figure 9 shows the confusion matrices for both schemes. The average balanced accuracy using models from learnt templates is 54.4%, substantially better than chance (10%), and comparable to the performance of non-native English speakers (51%). The accuracy exceeded 80% for four out of the ten categories. As in the human experiment the category *Heliconius charitonius* proved problematic, with no images of this category being correctly categorised. The reason is similar to the human confusion for this category – the 'yellow' bands are misclassified as white, and the spots clearly visible in the images (see Figure 2) are not mentioned in the description (Figure 3,



	danaus plexippus	heliconius charitonius	heliconius erato	junonia coenia	lycaena phlaeas	nymphalis antiopa	papilio cresphontes	pieris rapae	vanessa atalanta	vanessa cardui	
danaus plexippus	40			2		1	16		39	1	82
heliconius charitonius		0		4		1	44		51		93
heliconius erato			51	15	7	2			26		61
junonia coenia				56	24	10		7	1	2	90
lycaena phlaeas	1			8	89	2					88
nymphalis antiopa			1	10	1	43	3	8	11	23	100
papilio cresphontes				1			87		12		89
pieris rapae							9	91			55
vanessa atalanta	2			6					82	10	90
vanessa cardui	10			6	26	1		7	25	25	84

(a) ground truth templates

	danaus plexippus	heliconius charitonius	heliconius erato	junonia coenia	lycaena phlaeas	nymphalis antiopa	papilio cresphontes	pieris rapae	vanessa atalanta	vanessa cardui	
danaus plexippus	35					4	18		41	1	82
heliconius charitonius		0		1		1	42		55	1	93
heliconius erato			49		10	7		2	33		61
junonia coenia	1			2	42	20		12	2	20	90
lycaena phlaeas	3				94	2					88
nymphalis antiopa			1		1	61	2	3	12	20	100
papilio cresphontes							85		15		89
pieris rapae							7	93			55
vanessa atalanta	2								84	13	90
vanessa cardui	6				15				39	39	84

(b) learnt templates

Figure 9: Results of the proposed method. The confusion matrices for classifiers using (a) ground truth templates and (b) automatically learnt templates are shown. The balanced accuracy is 56.3% and 54.4% respectively. Accuracy varies greatly over the categories – see text for discussion.

top), causing additional confusion with the white spots of *Vanessa atalanta*.

Errors in the automatically learnt templates have a modest effect on the overall accuracy: 56.3% (ground truth) *vs.* 54.4% (automatic). An example error in the learnt templates occurred for *Junonia coenia* (Figure 3, bottom). The description "Eyespots black, yellow-rimmed" is incorrectly parsed as a pair of dominant colours black and yellow, resulting in an incorrect colour model.

Table 1 summarises the mean accuracy across all categories for each experiment, additionally comparing the accuracy of the proposed method using either of the visual attributes (spot and dominant colours) in isolation, and in combination. Each attribute in isolation performs similarly – 39.1% for spots and 35.3% for dominant colour. The two attributes prove complementary, with the combined accuracy being 54.4%. This suggests that it will be fruitful to add further attributes, for example patterns, by extending the visual processing to other properties of the learnt templates.

7 Conclusions and future work

We have proposed methods for learning models of object categories from readily available natural language descriptions in online nature guides, and showed that using NLP methods allows extraction of visually salient attributes from such text. Our method achieved moderate accuracy for 10 categories of butterflies, interestingly comparable to a non-native English speaker in our human experiment. Future work will concentrate on four aspects: (i) improved information extraction from text; (ii) combining information from multiple texts and more general corpora; (iii) more complete vision models learnt with weaker supervision; (iv) effectively combining text information with a *few* images. We believe the latter – a judicious combination of rich information from text with a small training set will always outperform learning of models from text alone.

Acknowledgements. We are grateful for financial support from the ORSAS scheme for Josiah Wang, and an RCUK Academic Fellowship supporting Mark Everingham.

References

[1] S. Abney. Parsing by chunks. *Principle-Based Parsing*, 1991.

[2] T. L. Berg and D. A. Forsyth. Animals on the web. In *Proc. CVPR*, 2006.

[3] T. L. Berg, A. C. Berg, J. Edwards, and D. A. Forsyth. Who's in the picture? In *NIPS*, 2004.

[4] J. R. Curran and S. Clark. Investigating GIS and smoothing for maximum entropy taggers. In *Proc. EACL*, pages 91–98, 2003.

[5] P. Duygulu, K. Barnard, J. F. G. de Freitas, and D. A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proc. ECCV*, 2002.

[6] eNature. http://www.enature.com/fieldguides/.

[7] M. Everingham, J. Sivic, and A. Zisserman. "Hello! My name is... Buffy" – automatic naming of characters in TV video. In *Proc. BMVC*, 2006.

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results. http://www.pascal-network.org/challenges/VOC/voc2008/workshop/index.html, 2008.

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

[10] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from Google's image search. In *Proc. ICCV*, 2005.

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

[12] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007.

[13] A. Gupta and L. Davis. Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In *Proc. ECCV*, 2008.

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

[15] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Proc. CVPR*, 2008.

[16] S. Lazebnik, C. Schmid, and J. Ponce. Semi-local affine parts for object recognition. In *Proc. BMVC*, 2004.

[17] L.-J. Li, G. Wang, and L. Fei-Fei. OPTIMOL: automatic Object Picture collecTion via Incremental MOdel Learning. In *Proc. CVPR*, 2007.

[18] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2), 2004.

[19] F. Schroff, A. Criminisi, and A. Zisserman. Harvesting image databases from the web. In *Proc. ICCV*, 2007.

[20] J. van de Weijer, C. Schmid, and J. Verbeek. Learning color names from real-world images. In *Proc. CVPR*, 2007.

[21] O. Veksler. Star shape prior for graph-cut image segmentation. In *Proc. ECCV*, 2008.