Automatic Metadata Annotation of Images via a Two-Level Learning Framework

Chih-Fong Tsai, Ken McGarry, John Tait
School of Computing, Engineering and Technology
University of Sunderland, Sunderland SR6 0DD, UK
+44 919 515 3555 ext. 3291, 3283, 2712
{Chih-Fong.Tsai, Ken.McGarry, John.Tait}@sunderland.ac.uk

ABSTRACT
A key problem for handling multimedia data in the semantic web is finding a way to associate concepts from ontologies to multimedia data items at an acceptable cost. This paper describes experiments with a system to assign automatically keyword metadata descriptors to unlabelled images. Learning to automatically match low level image features, like colour or texture to high level concepts (the so called *semantic-gap* problem) is very challenging. The usual approach is to design a learning machine or classifier to learn low-level feature vectors for high-level concept classification as a *single-step* (*direct*) *mapping* function. These systems often do not perform well for large numbers of classes. We present a two-level supervised learning framework for effective image annotation. In the first level induction stage, colour and texture feature vectors are classified individually into their corresponding outputs, i.e. colour and texture terms. Then, the colour and texture terms as middle-level features are classified into the target high-level conceptual classes during the second level induction stage. Three experimental studies are described in this paper. Experimental results using vocabularies of 60 and 150 keywords are reported, based on single step Support Vector Machines, two step Support Vector Machines, and k Nearest Neighbour. In the final experiment a comparison between human and automatic metadata annotation is described. Results show promise that the techniques will scale and perform acceptably for practical retrieval.

Categories and Subject Descriptors
H.3.1 [**Content Analysis and Indexing**]: Indexing methods; I.5.2 [**Design Methodology**]: classifier design and evaluation;

General Terms
Design, Experimentation

Keywords
Content-based image retrieval, image annotation/classification, machine learning, support vector machines, *k*-nearest neighbour

1. INTRODUCTION
At the core of the problem posed by the semantic web paper [4] is the connection of the worlds of human perception and action, and of human communication with the kinds of unambiguous representation normally manipulated by computer software agents.

This task, of connecting different forms of representation of perception, is highly challenging when dealing with hypertext on the web, but is still more challenging when considering other forms of data like images.

One way to make progress in manipulating unknown still images in the semantic web would be to automatically assign keyword or concept terms to each image and then manipulate them within the text based infrastructure.

This paper describes recent progress in constructing a system to perform such automatic keywording of still images. The heart of the approach adopted is to reformulate the keywording process from a process of identifying whether a keyword (say "lion") applies to an image to a process of identifying that class of images to which a user is likely to find it acceptable to return the class in response to the keyword query, with some degree of liklihood. In this way the problems of full blown object recognition are replaced with a simpler image classification task. Furthermore the automatic keywording problem is now formulated in a way which makes it amenable to supervised learning, given a collection of images which have acceptable performance when preexisting keywords are used to query the image collection.

A content-based image retrieval (CBIR) system which indexes and retrieves images by low-level image features cannot fully capture high-level concepts in humans' minds. This gives rise the so called *semantic-gap* problem in which the CBIR does not deal with the contents of images as viewed by the human searcher. This is a very challenging problem [31]. However, image annotation or classification could provide a solution to bridge the gap between low-level features and high-level concepts.

The literature shows a number of applications using machine learning techniques for this problem, such as Bayesian and probabilistic approaches [3, 8, 10, 20, 35], *k*-nearest neighbour (*k*-NN) [32], neural networks [19], support vector machines [8, 29, 34], combinations of different learning models for segmentation and classification processes respectively [6, 18, 33] and so on. Some of them classify each image into one category and some multiple categories per image. For this image *keywording* problem, we consider each image should belong to multiple categories, i.e. each image has multiple keywords assigned. Note

that as it has been addressed by supervised and unsupervised learning perspectives in which the former has better performances than the latter [25]. In addition, we are interested in extracting high-level concepts of images per se rather than extracting textual information associated with images, and furthermore trying to do this in a way which is compatible with proposals for the semantic web infrastructure like [14, 15].

1.1 Challenges

In general, a learning machine/classifier is designed to recognise/classify images by learning the low-level features directly. That is, image classification is approached by representing different low-level features of an image by one feature vector that is fed to a single classifier [2]. However, many existing classifiers described above only solve small scale problems, i.e. small numbers of classes, where discrimination between those conceptual classes is usually high and thus helps them perform well. Very few works in image databases and retrieval have dealt with larger numbers of classes by using machine learning techniques. Moreover, as these classifiers learn the combination of different low-level feature vectors directly to recognise/classify images into conceptual classes via single-step learning as *direct mapping*, the following problems occur:

- It is difficult to construct a metric which is simultaneously optimal for different features since features in colour, texture, and shape are extracted by different computational methods and thus may require different similarity measurements [32] and it is hard to lump several feature vectors together due to their diversified forms [16].
- The direct relationship between low-level features and high-level concepts usually does not exist because images which have similar visual contents may have different concepts, i.e. the feature vectors of some semantically dissimilar images may be located very close in the feature space [30].
- The combination of different feature vectors into a higher dimensional space may introduce the *curse of dimensionality* problem [5].

Therefore, computers have difficulty in learning a certain scale of high-level concepts by learning low-level features and need a more sophisticated learning strategy, such as active learning for relevance feedback [13, 34] for further improvement. However, a large number of initial training examples are required to ensure subsequent learning is efficient. The aim of this paper is to design a robust learning model to improve initial learning of keyword concepts.

1.2 Proposed Solution

Considering the above problems, we propose a two-level learning framework, namely a two-stage mapping model (TSMM) for reducing classification errors in a divide-and-conquer manner. The main idea is that colour and texture features are individually classified into colour and texture names/classes as mid-level features based on a colour and texture classifiers respectively. Either or both classifiers may sometimes classify new colour and texture vectors into incorrect (colour/texture) classes. However, we assume that to design a *second-level* classifier which learns the mapping between correct and/or error predictions of the two *first-level* classifiers and the desired outputs as high-level concepts could have a higher chance for correct classification.

The consideration of using multiple learning model(s) for one classification task is not new in pattern recognition, such as mixture of experts, stacking and meta-learning [5, 7]. One early related work focuses on solving the indoor and outdoor classification problem [32]. The idea is that colour and texture classifiers are designed to classify colour and texture features for indoor and outdoor predictions and a combiner is designed to *vote* the predictions to make the final decision. In [10, 29], they use the same concept but different learning models/classifiers to solve the indoor and outdoor and/or close-up classification problem. Iyengar et al. [19] consider audio, speech, and visual models for video annotation and a combiner is designed to decide the final annotations based on the *scores*, i.e. confidence values as mid-level features produced from the three models. Similarly, a meta-classifier is built to decide the final annotations from two confidence values generated from text- and image-based classifiers for unlabelled images [24]. In [8], an ensemble of binary-classifier is trained to give mutiple soft labels, i.e. class membership to an image and then, the most *correct* label(s) which have higer confidence values can be decided from these lables.

The novelty of our proposed approach is composed of the following components:

(1) first-level classification: the visual (colour and texture) features are decomposed for colour and texture classification instead of high-level concept classification directly,

(2) fusion of the colour and texture names/classes: a data extractor is designed to fuse the colour and texture names (mid-level features) and select candidate training examples for second-level classification instead of *voting* from the outputs (high-level concepts) of first-level classification,

(3) second-level classification: the colour and texture names/classes represented by binary feature vectors instead of confidence values are used to map into the final high-level concepts.

Therefore, this divide-and-conquer approach is new and different from the related work which designs multiple *direct mapping* classifiers for image annotation.

In this paper, we conduct two quantitative experiments to compare the propsed approach with a base (single-step) learning approach under the scale problem of 60 and 150 categories respectively as follows:

- For the 60-category problem, support vector machines (SVMs) [11, 36] are used for both approaches. This study is intended to investigate their error estimation and generalisation performances. The margin and number of support vectors of the hyperplane, and classification accuracy of these classifiers are examined. The contributions of this study are two fold. From the system performance perspective, we show that the two-level learning framework generalises better than the general single-step learning approach under 60-category classification in terms of classification accuracy and requires smaller numbers of training examples. In addition, it has better margin maximisation ability and reduces the number of support vectors for more effective classification. For the evaluation strategy using SVMs, the majority of related work only reports their classification accuracy, we further examine the margin and number of support vectors of an SVM with its relation to classification accuracy which is a simple but robust quantitative evaluation method.
- For the 150-category problem, SVMs constructed by the proposed approach are compared with a *k*-nearest neighbour (*k*-NN) classifier to examine their performances of both

classification accuracy and numbers of classes with zero rate accuracy. The contributions of this study are twofold. First, we challenge the image classification problem of 150-category that few or none of related supervised learning classifiers have tackled. Second, we further discover the reliability and extendibility of both classification techniques that SVMs perform better than *k*-NN only under smaller scale classification problems. On the contrary, *k*-NN performs stable under larger scale classification problems.

In addition, a qualitative study is conducted by asking human subjects to annotate a set of images using the 150 categories. As many image annotation systems are evaluated by some chosen ground truth dataset(s), few studies consider user-centred evaluation. Moreover, human judgments are usually based on assessing the results/outputs of a system directly. Few studies focus on comparing the system performance with human annotations.

The rest of this paper is organised as follows. Section 2 briefly describes the learning machine generation through inductive learning and the concept of support vector machines (SVMs) and *k*-nearest neighbour (*k*-NN). Section 3 presents the proposed two-level learning framework. Section 4, 5, and 6 describe the three experiments including their experimental setup, comparison methodology, and the results. Section 7 and 8 provide discussions and conclusions of this study respectively.

2. PATTERN CLASSIFICATION

2.1 Learning Model Generation

The goal of inductive learning or learning from examples is to build a general decision procedure based on a set of training examples. Given a training set with m examples, $\{(x_1, y_1), (x_2, y_2),\ldots, (x_m, y_m)\}$, for some unknown function $f(x) = y$ that x_i is represented by k number attributes (feature) vectors of x_i, i.e. $\{x_{i1}, x_{i2},\ldots, x_{ik}\}$ and each y_i represents a class label (high-level concept or keyword) associated with each x_i, the learning task is to compute a classifier or model f' that approximates f and correctly labels the training set. This can be called as the *training* stage. After the model f' is generated or trained, it is able to classify an unknown instance, i.e. low-level feature vectors, into one of the y class labels.

2.2 Support Vector Machines

Support Vector Machines (SVMs) are one of the major machine learning techniques and have been used for image database and retrieval applications such as indoor/outdoor [29] and natural scene classification [8], colour histogram classification [9], texture classification [23], relevance feedback [13, 34], etc.

An SVM is designed for binary classification. That is, to separate a set of training vectors which belong to two different classes, (x_1, y_1), $(x_2, y_2),\ldots, (x_m, y_m)$ where $x_i \in R^n$ denotes vectors in a n-dimensional feature space and $y_i \in \{-1, +1\}$ is a class label. For any feature vector $x \in R^n$, $f(x) \in \{-1, +1\}$ is the predicted label for x. During the SVM model generation, the input vectors, i.e. low-level feature vectors, such as colour and texture, are mapped into a new higher dimensional feature space. Then, an optimal separating hyperplane in the new feature space is constructed by a kernel function. There are two most used kernel functions which are Polynomial and Gaussian Radial Basis Function (RBF) kernel functions. For detailed description, please consult [11, 36].

Figure 1 shows two examples with different margins and the larger margin which is the distance between the two dashed lines is expected to provide better generalisation [26]. The larger margin can be interpreted as a 'confident' correct classification [28].

All vectors lying on one side of the hyperplane are labeled as -1, and all vectors lying on another side are labeled as +1. The training instances that lie closest to the hyperplane are called support vectors. The number of these support vectors is usually small compared to the size of the training set and they determine the margin of the hyperplane. If the optimal separating hyperplane can be constructed from a small number of support vectors, the generalisation ability will be high [11].

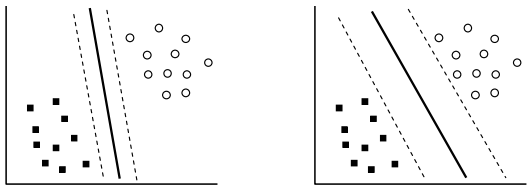


Figure 1. A separating hyperplane with a small margin shown in the left hand side and larger margin shown in the right hand side

For M-class classification problems where $M > 2$, there are two general approaches [17]. The first one is 'one-against-others' that M SVMs are constructed and each of them is to classify one positive class and M-1 negative classes. The second one is 'one-against-one' that $\frac{M(M-1)}{2}$ classifiers are constructed and each of them is to classify one positive and negative class. According to [9] the accuracies of both methods are almost the same, but Chang et al. [8] report that the former approach performs better than the latter one. Therefore, we chose the more computationally efficient method which is the 'one-against-others' approach for constructing a cascade of multiple binary SVMs.

2.3 *k*-Nearest Neighbours

In pattern recognition, the *k*-NN (*k*-nearest neighbours) classifier is a conventional nonparametric classifier [5]. It is different from the inductive learning approach described previously which needs *training* as approximating a function of mapping between input feature vectors and their corresponding class labels. Therefore, k-NN is computationally more efficient than inductive learning methods. This *k*-NN rule is assumed that a new instance belongs to the same class as its *k* nearest neighbours in the training data set (where *k* is an integer). A neighbour is deemed nearest if it has the smallest distance in the feature space. Therefore, the *k*-NN algorithm needs *searching* through all the examples of the given training set for classifying the new instance. That is, the main computation of *k*-NN is the on-line scoring of training examples to find the *k* nearest neighbours of the new instance.

3. THE TWO-LEVEL LEARNING FRAMEWORK

This section describes our two-level learning framework which aims at providing better generalisation performance for effective image classification. The first level inductive leaning stage is to generate two models which can predict colour and texture names

of an image. Then, the second level inductive learning stage is to generate a model which gives the final prediction as the label from the first level predictions.

Figure 2 depicts the first level induction framework. Given a training set P, i.e. $P = \{(x_1, y_1), (x_2, y_2),\ldots, (x_m, y_m)\}$ where m is the number of training examples and each image, x_i, is represented by a k number attribute (feature) vectors of x_i, i.e. $\{x_{i1}, x_{i2},\ldots, x_{ik}\}$ and each y_i represents a class label associated with each x_i. Next, we partition P into Pc and Pt such that that $P = Pc \cup Pt$ and $Pc \cap Pt$ = empty_set, i.e. $x_i = x_{ci} \cup x_{ti}$ and $x_{ci} \cap x_{ti}$ = empty_set where 'c' and 't' represent colour and texture features respectively. Therefore, $Pc = \{(x_{c1}, y_{c1}),(x_{c2}, y_{c2}),\ldots,(x_{cm}, y_{cm})\}$ where each y_{ci} is a label of the colour names associated with each x_{ci} and $Pt = \{(x_{t1}, y_{t1}),(x_{t2}, y_{t2}),\ldots,(x_{tm}, y_{tm})\}$ where each y_{ti} is a label of the texture names associated with each x_{ti}. While the models *f*'c and *f*'t are constructed, they are used to give level-0 predictions as colour and texture names of each x_i.

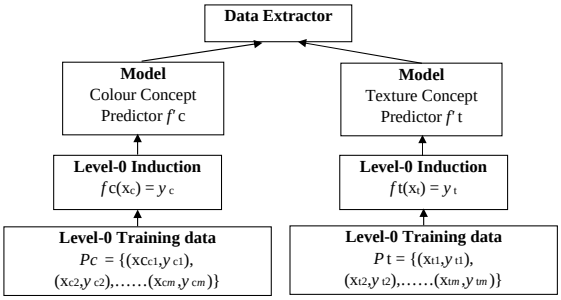


Figure 2. Level-0 model generation during first level learning

The data extractor is to generate new training examples for level-1 induction and model generation from the predictions of *f*'c and *f*'t. There is a major criterion to select candidate examples from the level-0 predictions. For example, if the predictions of colour and texture names are unique, i.e. there is just one x_i with such colour and texture predictions, then, the two predictions are chosen as the inputs/training examples for the level-1 induction. This criterion is intended to *correct* some error predictions of level-0 model(s) to map into desired predictions if any. Figure 3 shows the second level induction framework. The level-1 training data $Q = \{(z_1, y_1),(z_2, y_2),\ldots,(z_n, y_n)\}$ where $n \leq m$ and $z_i = \{c_i, t_i\}$. Then the model *f'* is constructed based on *Q*.

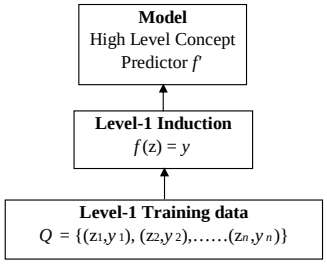


Figure 3. Level-1 model generation during second level learning

Table 1 gives an example of a *sky* class represented by c_i and t_i. The dimensionality of c_i and t_i depends on the pre-defined number of colour and texture concepts. After the training or model generation phase is done, the classification procedure (assigning keywords to images) shown in Figure 4 can be summarised as follows: given an unlabelled image which is represented by the colour and texture feature vectors. Then, the colour and texture feature vectors are first fed into the *f*'c and *f*'t models respectively. Next, the predictions (mid-level features) of both level-0 models are fed into the *f*' model for the final prediction.

Table 1. The middle-level feature vector for the *sky* class

c_i and t_i	Feature vector
white, blue, red, … colours (c_i)	{0, 1, 0, …}
sky, grass, tree, … textures (t_i)	{1, 0, 0, …}

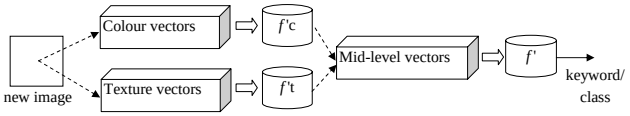


Figure 4. The classification procedure

4. EXPERIMENT I

This section describes the first experimental study including experimental setup and evaluation methodology by comparing our proposed two-level learning framework with the general single-level learning framework to construct SVMs. The margins and numbers of support vectors of hyperplanes of the SVMs are examined. In addition, some images are used to test their classification accuracies to see the relationship between the margin and number of support vectors and classification accuracies. The aim of this study is to see whether the proposed approach outperforms the general single-step learning one by using SVMs.

4.1 Experimental Setup

The Corel stock photo library is used for the dataset. Due to the lack of standard data sets for evaluation, for the first study we manually identified 60 categories for training and testing. We only chose suitable patches of images for training. For example, in Figure 5 the down-left tile is used for training the *beach* class. Note that these categories are *concrete* classes, each of which means a physical object or entity defined by WordNet [38], such as *tree*, *car*, *building*, etc. The number of examples in the training set is 1,639 in which the training examples per category range from 8 to 62.

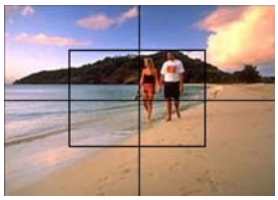


Figure 5. The tiling scheme

Each image is divided into five equal area tiles illustrated in Figure 5. The input vectors of each example are composed of colour and texture vectors, in which HSV (hue, saturation, value) and three levels of Daubechies-4 wavelet decomposition [12] are the colour and texture representations respectively.

Consequently, the general single level learning approach constructs 60 high-level concept SVMs and the proposed approach constructs 10 colour and 60 texture SVMs for the first level induction and 60 high-level concept SVMs for the second level induction by 226 training examples generated from the data extractor. All of the classifiers are produced by the Matlab

Support Vector Machine Toolbox [1]. We apply a degree 2 Polynomial and sigma 2 RBF kernel functions to design different SVMs for further comparisons and the *C* parameter is 1 as the default of the SVM toolbox. It should be noted that the ten colour names are based on the colour perceptual model of CHROMA [22] which has a fixed set of colour names, such as *white*, *blue*, *green*, etc. However, the defined colour names are sufficient to express the colour information [39] and the high-level concept SVMs will decide the final answer based on the training set *Q*. These inputs will be mapped into desired target conceptual classes. For the definition of texture names, as texture is one of the most important characteristics which has been directly used to classify and recognise objects and scenes (e.g. [1]) but it is difficult to describe in terms of a generally understood fixed set of descriptors. The number of texture classes defined here is the same as the one of conceptual classes, i.e. the texture names are only thought of as *middle-level concepts* in the first prediction stage.

4.2 Performance Evaluation

To evaluate the classification performances, i.e. error estimation and generalisation, we examined the margin of the hyperplane and its number of support vectors for each of the 60 high-level concept SVMs of the two approaches, i.e. the *f*' models. We manually selected 60 unknown images which contain all the 60 concepts as the test set. Each of these images is partitioned into five equal sized regions/zones (i.e. centre, up-left/right, and down-left/right square subimages) and thus contains five sets of unseen colour and texture vectors per image. Therefore, there are 300 sets of colour and texture future vectors and each test image will be classified into five categories, i.e. each image has five keywords assigned. We also examined the classification performances of both approaches in terms of the number of training examples. Note that the ground truth answer of the test set, i.e. five keywords per image, is manually defined by the authors since each image of Corel only belongs to one category.

4.3 Results

For the training results, Table 2 shows the averaged results of the 60 high-level concept SVMs where 'TSMM' represents the proposed approach and 'DM' represents the general direct mapping one. (For more detailed information of each of the 60 SVMs, please refer to the Appendix I).

Table 2. Average Margin and number of support vectors

	Average Margin	Average No. of Support Vectors
TSMM (Poly)	1.3	27
DM (Poly)	1.13	86
TSMM (RBF)	1.16	159
DM (RBF)	0.9	83

They indicate that the Polynomial (Poly) kernel function outperforms the Gaussian Radial Basis Function (RBF) one. In addition, the SVMs trained by the proposed learning framework is expected to have better generalisation performances than the SVMs trained by the general approach because of their larger margins and smaller numbers of support vectors. Therefore, we used the trained Polynomial SVMs of both approaches to compare their classification accuracies based on the test set. Figure 6 shows classification accuracies of both approaches for some classes. Figure 7 shows some classification results of both approaches. TSMM outperforms DM which is significant at the 0.01 level.

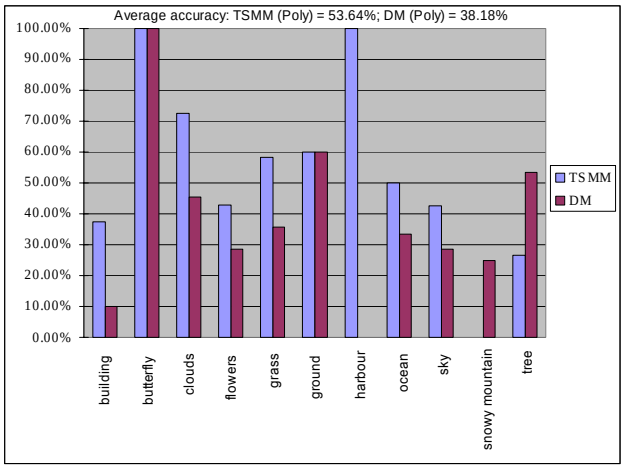
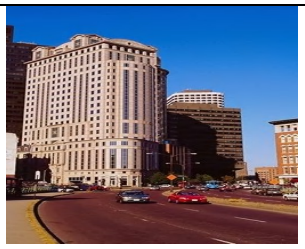


Figure 6. Classification accuracies of some classes

(a) TSMM: 'harbour, building, sky, wild cat, plane'; DM: 'trees, grass, sky, flowers, trees'

(b) TSMM: 'clouds, sand, clouds, sky, trees'; DM: 'beach, clouds, beach, beach, beach'

(c) TSMM: 'plane, clouds, cityscape, train, grass'; DM: 'trees, clouds, trees, flowers, trees'



(d) TSMM: 'cityscape, birds, sky, cathedral, trees'; DM: 'boats, wild, cat, grass, grass, trees'

Figure 7. Some classification examples by the two approaches

Figure 8 shows the relationship between classification accuracies and numbers of training examples of both approaches. By using different numbers of training examples, on average our proposed approach shows promise. The general direct mapping approach needs a large training set to give comparable performance with the two-stage mapping approach. In addition, the performance of the proposed approach when using small numbers of training examples is better than the direct mapping one. This further implies that the *error correction* mechanism of the second-level

[1]The Matlab Support Vector Machine Toolbox is downloaded from http://www.isis.ecs.soton.ac.uk/isystems/kernel/.

classifier is able to improve the performance of image keywording. As larger numbers of training examples may be difficult to obtain in practice and they are computationally demanding during training, the proposed approach is a suitable solution for classifying image databases in a scale of at least 60 conceptual categories.

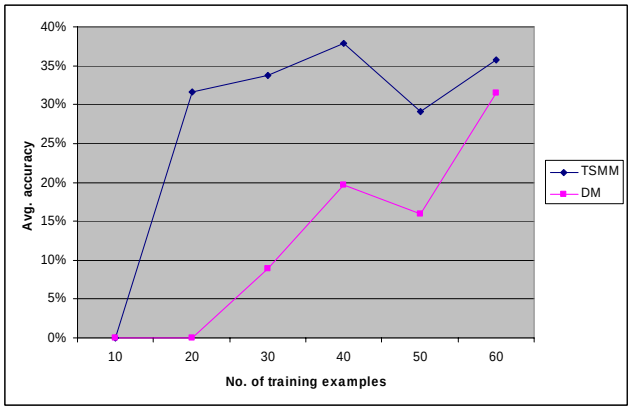


Figure 8. Classification accuracy vs. numbers of training examples

We found another interesting result about the number of unpredictable classes to which no examples from the test set have been assigned, i.e. those classes have zero rate classification accuracy. There are 19 and 28 classes out of 60 which have zero classification rates for TSMM and DM respectively. This may be caused by the imbalance of training examples for each class. However, this result can further indicate that under a certain scale problem, our proposed approach has more discriminative power or provides better mapping between low-level features and high-level concepts than the general single-step learning one. We believe that the above results can be improved by using more detailed colour and texture features and setting better parameters of SVMs, which will be considered in our future work.

5. EXPERIMENT II

This section presents the second experimental study including experimental setup and evaluation methodology by comparing the proposed approach using SVMs with a *k*-NN classifier. Both classification accuracy and numbers of classes with zero rate accuracy under the scale of 10, 30, 50, 70, 100, and 150 categories are examined. Section 4 has shown promising performances of our proposed approach by using SVMs, the aim of this study is to see the performance of SVMs under larger scale problems by comparing with a general classification approach, *k*-NN.

5.1 Experimental Setup

The dataset is based on Corel and the classification problem is scaled by 10, 30, 50, 70, 100, and 150 categories. In order to be more realistic, these categories include not only concrete classes but also *abstract* classes which mean abstraction, human activity, or an assemblage of multiple physical objects/entities defined by WordNet [38], e.g. *festival*, *parade*, and *studio* for the first, second and final cases respectively. They are based on the pre-classified categories of Corel. (Appendix II lists the chosen 100 concrete and 50 abstract classes.).

The proportion of training and testing examples per class is 30:20 and each example is composed of HSV colour and wavelet texture feature vectors which is the same as the previous study. For classifier design, the degree 2 Polynomial SVMs and a *k*-NN classifier (k = 1 to be the simplest classifier) are constructed.

5.2 Performance Evaluation

To evaluate the classification performances of both classifiers, we examine their classification accuracy and numbers of classes which have zero rate accuracy. Each test image is partitioned into five patches which are the same as the first study, and each image has five keywords assigned. For simplicity, we define accuracy as if at least one out of the five assigned keywords to an image is correct, then the image has correct keyword(s) assigned.

5.3 Results

The classification performance of both classifiers is shown in Figure 9 in which there are two numbers on each plot. The first one represents the rate of classification accuracy and the second one the number of classes which have zero rate accuracy.

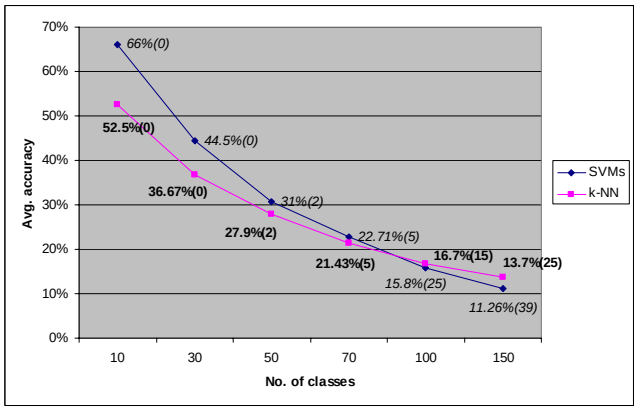


Figure 9. Performances of SVMs and *k*-NN

This result shows that SVMs outperform *k*-NN below the scale of 70 categories in terms of classification accuracy. However, both classifiers have the same performance for the difficult to identify classes, i.e. zero rate classification accuracy. When the scale problem increases to 100 and 150 categories, *k*-NN has higher classification accuracy and smaller numbers of difficult to identify classes than SVMs.

We believe that the poorer performance of SVMs under larger scale classification problems is caused by its learning structure since an SVM is trained by 'one-against-others' for binary classification. If the positive class closes to one or some of the negative classes in the feature space, i.e. feature overlapping, then this SVM would not perform as good as other classification approaches, such as *k*-NN. However, we could further draw a conclusion that SVMs have better classification performances than other learning models under smaller scale classification problems since the discrimination of smaller numbers of classes is high and the learning structure of SVMs makes them easier to be identified. On the contrary, *k*-NN has a great potential in dealing with larger scale classification problems.

6. EXPERIMENT III

This section presents a qualitative study for performance evaluation of the 150 SVMs. The aim of this study is to compare the performance of the learning machine with human annotators performance. Corel has just a single conceptual class assigned to each image, and it might be that our results were in some dependent on the this (or other particular) characteristics of the Corel collection rather than the actual process of annotating this particular set of images with the 150 conceptual keywords in the test.

6.1 Experimental Setup

We asked five judges (PhD research students) who are not experts in image indexing and retrieval to annotate the images. There were three male judges and two females who were all English first language speakers.

They were asked to annotate 60 images each (two images from each of 30 Corel categories) and assign between two and five keywords from a vocabulary of the 150 words drawn from the Corel classes used in previous experiment.

6.2 Performance Evaluation

The annotation results of the 150 SVMs for the 60 images are compared with each of the five human annotations. The definition of accuracy used is the same as Experiment II, that is if at least one of the five assigned keywords to an image is correct, then the image is regarded as having correct keyword(s) assigned.

6.3 Results

Table 3 shows the classification accuracy based on each of the five judges' annotations. On average, the system provides 22.27% classification accuracy. The correlation coefficient of the human annotations is moderately correlated, i.e. $r = 0.678$ at the significant level of 0.01 based on Pearson Product-Moment Correlation Coefficient [27].

Table 3. System performance based on the five judges

	Judge1	Judge2	Judge3	Judge4	Judge5
Accuracy	28%	20%	16.67%	21.67%	25%

It is interesting that the system performance is around twice as accurate measured in this way as compared to the accuracy results using Corel's classification as the ground truth. This qualitative study is encouraging that although the system has quite inaccurate annotation, i.e. 5.63% classification accuracy under the Corel data set, this is likely to be quite a stringent test of performance.

7. DISCUSSION

The advantages of the proposed approach over the general single-step learning one are that it has larger margin and smaller number of support vectors by using SVMs. In addition, by using the small test set the proposed approach shows better classification performances and needs smaller numbers of training examples. Furthermore, it has small numbers of unpredictable classes, i.e. zero classification rates.

Superficially, it may be thought of as a disadvantage that additional training for second level induction is required. On the contrary, the additional training is more than worthwhile since not only will it improve overall classification accuracy it can also actually overcome errors in first level training and classification. This accords with previous results using stacked generalisation [37].

The proposed approach could be considered for some early work which only employs a single low-level features for indexing and retrieval, such as colour [7, 19] and texture [1, 20]. That is, colour- and texture-based CBIR systems can be combined for image annotation.

For the choice of classification techniques, different classifiers have various performances under different scale classification problems. Therefore, it depends on the problem domain. In our study, we suggest that if the problem domain is under smaller scale classification problems, i.e. smaller numbers of classes, such as indoor/outdoor, natural/manmade classification, etc., SVMs have better performance. If the problem domain is to deal with larger numbers of classes, *k*-NN has a potential to perform better at least than SVMs although classification accuracy of current machine learning techniques is unlikely to be compatible with humans.

The experimental results of user-centred evaluation point out problems of current quantitative studies which are based on some chosen ground truth data sets. Although have not yet had the opportunity to fully analyse the reasons for the improved performance of the learning system against this test, it may be on occasions the Corel category is not an obvious one to the human annotator, and correspondingly this keyword/concept is difficult for the automatic system to assign. We therefore conclude that to make a full assessment and/or understanding of the performance of an image annotation system, user-centred studies need to be undertaken as well as more system centred ones.

8. CONCLUSIONS

In order to integrate multi-media data within the semantic web, it is necessary to find some mechanism, ideally automatic, to relate concepts in the ontology to items of multimedia data.

In this paper we have presented a series of experiments which show that an approach based on supervised learning in the context of a two stage learning model show promise for the purpose of automatically assigning keywords or concepts to still images.

It might be thought that a classification accuracy of as little as one in eight is not adequate for practical purposes. However, in practice the most likely use of such a system is in the context of a retrieval system supporting analytic metadata/keyword querying, browsing and query refinement techniques like relevance feedback. Unlike text, users can very rapidly assess the relevance or otherwise of large numbers of images presented to them in parallel. Presenting 24 thumbnail images on an initial query result screen appears not to be too many. On average, then, around three relevant images should be presented using the learning techniques we propose which provides an adequate starting point for relevance feedback.

The two prime issues of concern at the present time include the number of concepts for which no relevant images will be retrieved and the feasibility of obtaining training sets matched to concepts in ontologies.

Our future work will address the issue of "unreachable" concept classes by looking at working with more discriminating feature spaces: for example better regioning, using information gain and perhaps latent semantic analysis to use different feature spaces for different classes (either at the first or second level of the learning model). We also intend to investigate the use of concepts of

information gain and of generality/speciality in ontologies to organize the learning model to avoid this problem.

Tsai, McGarry and Tait [33] have investigated a mechanism to combine supervised and unsupervised learning to generate training sets of sufficient scale where only small samples of classified images are available. We propose also to investigate this approach to the problem of availability for suitable training sets.

Other avenues of future work include using higher *k*'s in the *k*-NN algorithm to provide a degree of generalization, user and task centred retrieval studies and dealing with larger number of concepts.

To conclude then, in this paper we have shown the feasibility of assigning conceptual classes from a vocabulary of 150 keyword terms to still images using a combination of fairly simple image processing and a two level learning classifier. The system is shown to be sufficiently effective to form the basis of a practical web image retrieval system in the future.

9. ACKNOWLEDGEMENT

The authors would like to thank our colleagues Chris Stokoe, James Malone, Sheila Garfield, Mark Elshaw, and Jean Davison for participating in the user-centred evaluation.

10. REFERENCES

[1] Aksoy, S. and Haralick, R.M. (2000) Using texture in image similarity and retrieval. In *Texture Analysis in Machine Vision*, Pietikainen, M. and Bunke, H. (Eds.), vol. 20, World Scientific, Singapore, pp. 129-149.

[2] Antani, S., Kasturi, R., and Jain, R. (2002) A survey on the use of pattern recognition methods for abstraction, indexing and retrieval of images and video. *Pattern Recognition*, vol. 35, pp. 945-965.

[3] Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D.M., Jordan, M.I. (2003) Matching Words and Pictures. *Journal of Machine Learning Research*, vol. 3, pp. 1107-1135.

[4] Berner-Lee, T., Hendler, J. and Lassila, O. (2001) The Semantic Web. *Scientific American*, May 17.

[5] Bishop, C.M. (1995) Neural networks for pattern recognition. Oxford University Press, Oxford.

[6] Campbell, N.W. and Thomas, B.T. (1997) Automatic segmentation and classification of outdoor images using neural networks. *International Journal of Neural Systems*, vol. 8, no. 1, pp. 137-144.

[7] Chan, P.K. and Stolfo, S.J. (1995) Comparative evaluation of voting and meta-learning on partitioned data. *Proceedings of the 12th International Conference on Machine Learning*, Tahoe City, California, July 9-12, pp. 90-98.

[8] Chang, E., Kingshy, G., Sychay, G., and Wu, G. (2003) CBSA: content-based soft annotation for multimodal image retrieval using Bayes point machines. *IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Conceptual and Dynamical Aspects of Multimedia Content Description*, vol. 13, no. 1, pp. 26-38.

[9] Chapelle, O., Haffner, P., and Vapnik, V.N. (1999) Support vector machines for histogram-based image classification. *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 1055-1064.

[10] Ciocca, G., Cusano, C., Schettini, R., and Brambilla, C. (2003) Semantic labeling of digital photos by classification. *Proceedings of Internet Imaging IV, SPIE 5018*, Santa Clara, CA, Jan. 20-24.

[11] Cortes, C. and Vapnik, V. (1995) Support vector networks. *Machine Learning*, vol. 20, pp. 273-297.

[12] Daubechies, I. (1992) *Ten lectures on wavelets*. Capital City Press, Vermont.

[13] Hong, P., Tian, Q., Huang, T.S. (2000) Incorporate support vector machines to content-based image retrieval with relevance feedback. *Proceedings of the IEEE International Conference on Image Processing*, vol. 3, Vancouver, Canada, Sep. 10-13, pp. 750-753.

[14] Horrocks, I and Patel-Schneider, P.F. (2003) Three thesis of representation of the semantic web. *Proceedings of the 12th ACM International World Wide Web Conference*, Budapest Hungary, May, 20-24, pp. 331-339.

[15] Horrocks, I. Patel-Schneider, P.F., van Harmelen, F. (2003) From SHIQ and RDF to OWL: the making of a web ontology language. *Journal of Web Semantics*, Volume 1, Issue 1, pp. 7-26.

[16] Huang, Y., Chan, K.L., and Zhang, Z. (2003) Texture classification by multi-model feature integration using Bayesian networks. *Pattern Recognition Letters*, vol. 24, pp. 393-401.

[17] Hsu, C.-W. and Lin, C.-J. (2002) A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks*, vol. 12, pp. 1288-1298.

[18] Iyatomi, H. and Hagiwara, M. (2002) Scenery image recognition and interpretation using fuzzy inference neural networks. *Pattern Recognition*, vol. 35, no. 8, pp. 1793-1806.

[19] Iyengar, G., Nock, H.J., and Neti, C. (2003) Discriminative model fusion for semantic concept detection and annotation in video. *Proceedings of the 11th ACM International Conference on Multimedia*, Berkeley, CA, Nov. 2-8, pp. 255-258.

[20] Jeon, J., Lavrenko, V., and Manmatha, R. (2003) Automatic image annotation and retrieval using cross-media relevance models. *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Toronto, Canada, July 28-Aug. 1, pp. 119-126.

[21] Kuroda, K. and Hagiwara, M. (2002) An image retrieval system by impression words and specific object names – IRIS. *Neurocomputing*, vol. 43, no. 1-4, pp. 259-276.

[22] Lai, T.-S. and Tait, J. (2000) CHROMA: a content-based image retrieval system. *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Berkeley, California, Aug. 15-19, pp. 324.

[23] Li, S., Kwok, J.T., Zhu, H., and Wang, Y. (2003) Texture classification using support vector machines. *Pattern Recognition*, vol. 36, no. 12, pp. 2883-2893.

[24] Lin, W.-H. and Hauptmann, A. (2002) News video classification using SVM-based multimodal classifiers and combination strategies. *Proceedings of the 10th ACM*

International Conference on Multimedia, Juan les Pins, France, Dec. 1-6, pp. 323-326.

[25] Monay, F. and Gatica-Perez, D. (2003) On image auto-annotation with latent space models. *Proceedings of the 11th ACM International Conference on Multimedia*, Berkeley, CA, Nov. 2-8, pp. 275-278.

[26] Osuna, E., Freund, R., Girosi, F. (1997) Training support vector machines: an application to face detection. *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, Puerto Rico,June 17-19, pp. 130-136.

[27] Pagano, R.R. (2001) *Understanding statistics in the behavioral sciences*, Sixth Edition. Wadsworth/Thomson Learning, California.

[28] Schapire, R.E., Freund, Y., Bartlett, P., and Lee, W.S. (1997) Boosting the margin: a new explanation for the effectiveness of voting methods. *Proceedings of the 14th International Conference on Machine Learning*, Nashville, Tennessee, USA, July 8-12, pp. 322-330.

[29] Serrano, N., Savakis, A., and Luo, J. (2002) A computationally efficient approach to indoor/outdoor scene classification. *Proceedings of the IEEE International Conference on Pattern Recognition*, Quebec, Canada, Aug. 11-15, pp. 146-149.

[30] Sheikholeslami, G., Chang, W., and Zhang, A. (1998) SemQuery: semantic clustering and querying on heterogeneous features for visual data. *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 5, pp. 988-1002.

[31] Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., and Jain, R. (2000) Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349-1380.

[32] Szummer, M. and Picard, R.W. (1998) Indoor-outdoor image classification. *Proceedings of the IEEE International Workshop on Content-based Access of Image and Video Databases*, Bombay, India, Jan. 3, pp. 42-51.

[33] Tsai, C.-F., McGarry, K., and Tait, J. (2003) Image classification using hybrid neural networks. *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Toronto, Canada, July 28-Aug. 1, pp. 431-432.

[34] Tong, S. and Chang, E. (2001) Support vector machine active learning for image retrieval. *Proceedings of the ACM International Conference on Multimedia*, Ottawa, Ontario, Canada, Sep. 30-Oct. 5, pp. 107-118.

[35] Vailaya, A., Figueiredo, M.A.T., Jain, A.K., and Zhang, H.-J. (2001) Image classification for content-based indexing. *IEEE Transactions on Image Processing*, vol. 10, no. 1, pp. 117-130.

[36] Vapnik, V. (1998) *Statistical learning theory*. John Wiley, New York.

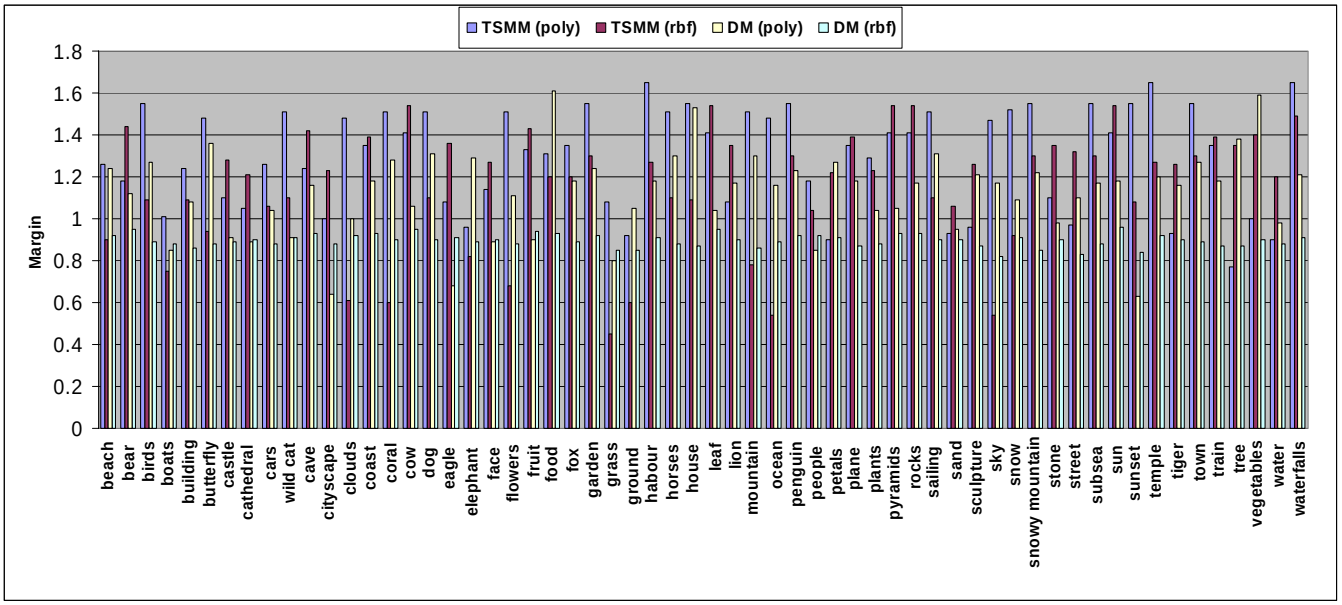
[37] Wolpert, D.H. (1992) Stacked generalization. *Neural Networks*, vol. 5, no. 2, pp. 241-259.

[38] WordNet [On-line] Available from: http://www.cogsci.princeton.edu/~wn/

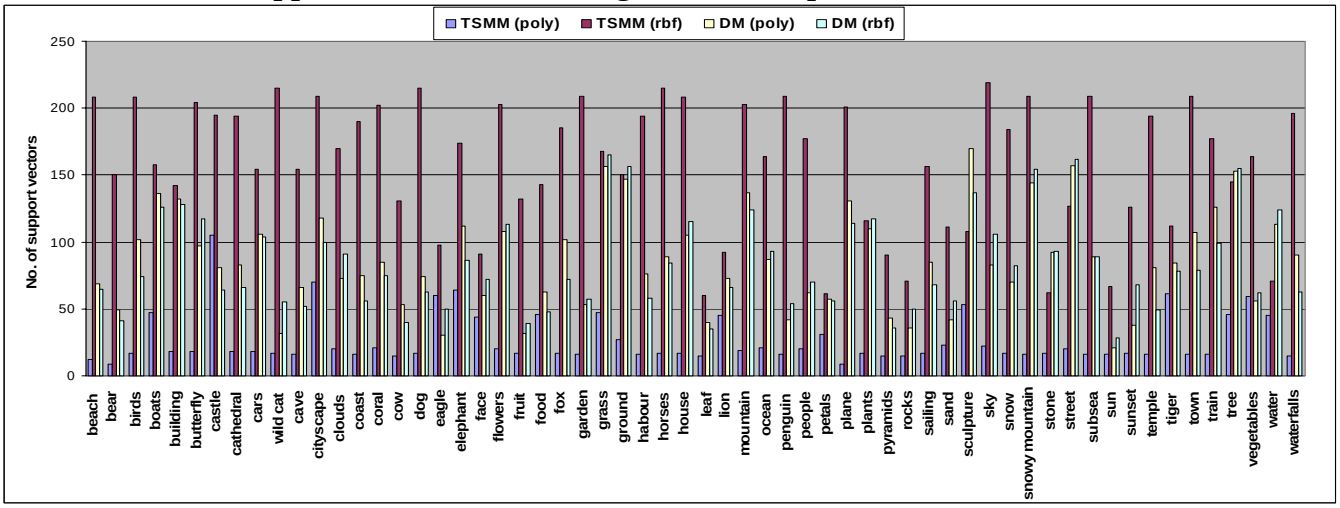
[39] Wyszecki, G. and Stiles, W.S. (2000) *Color Science: Concepts and Methods, Quantitative Data and Formulae*, 2nd Edition. John Wiley & Sons.

11. APPENDIX I

11.1 Margins of the 60 high-level concept SVMs



11.2 Numbers of support vectors of the 60 high-level concept SVMs



12. APPENDIX II

12.1 The Concrete Classes

Agates	Buses	Coast	Firework	Homes	Monuments	Perennial	Reptile	Subsea	War plane
Antelope	Butterfly	Cuisines	Flags	Horses	Mountain	Pills	Road	Tall ship	Waterfall
Antique	Cactus	Dessert	Flora	Jewelry	Mushroom	Plants	Rock form	Texture	Waves
Balloon	Car	Dogs	Flower	Lighthouse	Offices	Polo	Rodeo	Things	Wildcats
Beach	Cards	Dogsled	Flower bed	Machinery	Old dish	Predator	Roses	Tools	Wild bird
Bobsled	Castles	Doors	Foliage	Mammals	Old doll	Primates	Sail	Train	Wild fish
Bonsai	Cats	Drinks	Fractals	Man	Orchids	Pub signs	Sculpture	Tulips	Wild goat
Botany	Children	Everglade	Fruit	Marble	Owls	Puma	Shells	Valley	Whale
Beads	Churches	Fabric	Graffiti	Masks	Palaces	Pyramids	Stamps	Vegetable	Work ship
Building	Clothing	Firearms	Hawk	Minerals	Penguin	Race car	Steam engine	Volcano	Women

12.2 The Abstract Classes

Architecture	Barnyard	Cruise	Fashion	Game	Harbours	Nature	Pastoral	Space	Tropical
Autumn	Battles	Dawn	Festival	Gardens	Industry	Night	Rafting	Sports	Vineyard
Aviation	Compete	Desert	Fitness	Glamour	Interior	Old works	Ruins	Summer	Waterway
Ballet	Computer technology	Estate	Forests	Golf	Leisure	Parades	rural	Sunsets	Wet sports
Barbecue	Couples	Farm	Fountain	Hanover	Market	Park	Scene	Surfing	Winter