

Efficient image selection for concept learning

A. Dorado, D. Djordjevic, W. Pedrycz and E. Izquierdo

Abstract: In semantic-based image classification, learning concepts from features is an ongoing challenge for researchers and practitioners in different communities such as pattern recognition, machine learning and image analysis, among others. Concepts are used to add knowledge to the image descriptions linking high- and low-level numerical interpretation of the image content. Augmented descriptions are useful to perform more ‘intelligent’ processing on large-scale image databases. The semantic component casts the classification into the supervised or learning-from-examples paradigm, in which the classifier obtains knowledge by generalising specific facts presented in a number of design samples (or training patterns). Consequently, selection of suitable samples becomes a critical design step. The introduced framework exploits the capability of support vector classifiers to learn from relatively small number of patterns. Classifiers make decisions based on low-level descriptions containing only some image content information (e.g. colour, texture, shape). Therefore there is a clear drawback in collecting image samples by just using random visual observation and ignoring any low-level feature similarity. Moreover, this sort of approach set-up could lead to sub-optimal training data sets. The presented framework uses unsupervised learning to organise images based on low-level similarity, in effort to assist a professional annotator in picking positive and negative samples for a given concept. Active learning to refine the classifier model follows this initial design step. The framework shows promising results as an efficient approach in selecting design samples for semantic image description and classification.

1 Introduction

The rapid growth in consumer-oriented electronic technologies, for example, digital cameras, camcorders and mobile phones, along with the expansion in networking is facilitating production and consumption of striking amounts of digital information. It is also bringing a change in the way people process such information.

The challenge is in incorporating mechanisms in multimedia systems to resemble the way humans make decisions on the basis of how they interpret what they perceive. Those interpretations are subjective because of the different physiological and psychological responses of each beholder to visual stimuli. It has captured the attention of researchers in computer vision, pattern recognition and other related fields in the last decades. These efforts are focused on the task of adding knowledge to the image content in order to enable more ‘intelligent’ processing.

Although the semantic component casts the systems into the supervised or learning-from-examples paradigm,

methods applied on low-level primitives could allow a reduction of systems’ dependency on the designer.

Traditionally, proposed methods in machine learning and pattern recognition (e.g. clustering analysis) are used to designate a passage from visual features to human understanding of the image content in order to provide a way that a computer can execute the recognition process [1].

Designers use patterns in the form of labelled content to train the system. In such an approach, the learning process is based on basic visual interpretation of the image content indicating observed elements in the scene, for example, landscape, cityscape [2–4]. In this way, visual features can be linked to linguistic concepts at the highest level of abstraction.

This bottom-up approach from low-level to semantic meaning, used in most image content retrieval systems, relies completely on matching procedures at the lowest level of content interpretation. It is well known that two objects can be similar in their visual primitives, but semantically different to a human observer. Therefore substantial noise could be introduced in propagating interpretations using only low-level similarity.

From the other perspective, propagation based only on high-level similarity (top-down approach) puts a heavy burden on the designer and has undesirable effects in system performance, being stalled at a certain point for lack of information in decision making without the necessary human factor.

Combined approaches that go from the bottom to the top and in the opposite way are the foundation of the critical paradigm of ‘bridging the semantic gap’ [5].

With this in mind and thinking on the feasibility of learning from human subjectivity, a framework to assist

concept learning from examples in semantic-based image classification is presented.

The framework exploits the capability of support vector classifiers (SVCs) to learn from a relatively small number of samples [6]. These samples can be chosen using random selection of images, which does not guarantee quality or good representation of the concept. However, manual searching has some drawbacks. One of them is the definition itself of 'a good' sample, which involves subjectivity and varies from one designer to another. This sort of manual search could imply the need to traverse the entire database in an effort to obtain higher efficiency. Consequently, selection of suitable examples becomes a critical design step.

This framework uses unsupervised learning as first step in designing the classifier. By applying clustering, it organises images based on low-level similarity in order to assist a designer in picking positive and negative samples for a given concept. Basically, clustering outcomes are used to identify sensitive points that can define the hyperplane between groups of images associated with certain concept.

The component of low-level feature similarity, although effective, presents shortcomings in terms of efficiency due to unavoidable introduction of misleading information relying only on machine's interpretation of the content. Here is where active learning starts to play an important role in allowing a posteriori training mode and facilitating system's adaptation. Therefore in order to refine the classifier model, the initial design step is followed by active learning (AL). It is to say, after clustering the space of image descriptors positive samples are selected from feature vectors situated in the well-populated regions surrounding cluster prototypes relevant to the chosen concept. Negative samples are selected from feature vectors placed in regions surrounding concept contradictory cluster prototypes and in regions where two or more clusters overlap. Afterwards, it captures hints from the professional annotator (designer) regarding to classification outcomes of image representations observed from the clustering results capturing the underlying low-level similarity of image patterns.

Experimental results show that within the proposed framework, the SVC exploiting both low-level and semantic information achieves higher accuracy than using either random selection of samples or only AL.

2 Problem of learning concepts

Although the problem of learning concepts has been studied for decades, it is still an open issue. Focusing on the problem, Saitta and Bergadano [7] presented an interesting comparative analysis of results from pattern recognition and theoretical machine learning. In their continuing work, Bhanu and Dong [8] proposed a framework for learning concepts based on retrieval experience, which combines partially supervised clustering and probabilistic relevance feedback. The challenge of finding suitable samples is also observed in training strategies as the one presented by Boutell *et al.* [9].

There are also several interactive approaches that have been proposed to enable long-term learning and system's adaptation [8, 10] as well as methods achieving some improvement in performance by introducing group-oriented search of sample images [11, 12]. Tong and Chang [13] proposed the use of a support vector machine (SVM) AL algorithm for conducting effective relevance feedback for image retrieval. It produces a learner that is particularly well suited to the query refinement task in image retrieval, which outperforms a number of traditional query refinement

schemes. Smith and coworkers [14] used SVMs and AL in a very similar way to Zhang and Chen [15]. Essentially, it is an extension of the method proposed by Tong and Chang [13]. The important difference to our approach is that all of these solutions are focused on content-based image retrieval.

In our approach, the problem of learning concepts is addressed in the context of semantic-based image classification. The concepts are used to add knowledge to the image descriptions linking human and low-level numerical interpretation of the image content.

Semantic-based classifiers perform the task of using content-based descriptions (feature vectors) to assign certain objects to a given concept (semantic class or category). The inductive training process in learning-from-examples is carried out by presenting declarative knowledge through a number of labelled objects. Specifically, the two types of information supplied here are images that positively exemplify the concept and hints that tell the classifier (learner) whether or not the concept can be attached to the images [16]. In this way, these learning protocols are applied in semantic image classification to introduce concept-wise human subjectivity.

In order to refine the classifier model the initial design step is followed by AL. Nguyen and Smeulders [17] proposed a similar strategy, but limited to two-class AL.

An image is considered to be either a positive (ascribed to) or negative (not related to) sample of a given concept, if it satisfies a criterion defined by a professional annotator. For instance, a picture is a positive sample of a 'city view image' if it depicts a scene containing buildings within a city skyline.

Normally, design samples are taken from a large-scale database. Time periods required, relaxations in the selection criteria, subjectivity of the beholder, poor quality of the picture because of occlusion, shadows, rotation and amount of available examples are some of the identified drawbacks in collecting training patterns.

Choosing samples just on the basis of human perception misses out on the fact that in the end, the classifier will be using descriptions with limited domain knowledge, and not the overall cognitive perception of the world assumed by humans. Then, the problem can be stated as follows: 'How to assist designers in selecting samples to train semantic-based image classifiers?' Accordingly, the following framework is proposed.

3 Framework to assist concept learning from examples

3.1 General overview

By making use of feature vectors provided by the feature extractor, a classifier is aiming to assign certain objects to a category [18]. Low-level features are organised by combining unsupervised (automatic) and partially supervised (semi-automatic) pattern recognition training modes. The objective is to find a classifier model that roughly resembles the semantic categorisation of images.

Fig. 1 illustrates the proposed framework for training the classifier. In a first step, clustering mechanisms are used to assist the professional annotator in collecting image samples to train the classifier.

An initial data set is built with the best-ranked images (highest membership degree) in the clusters. These images are associated with sensitive points of two types: positive samples, high membership degree in a cluster associated with the expected concept and negative samples, high

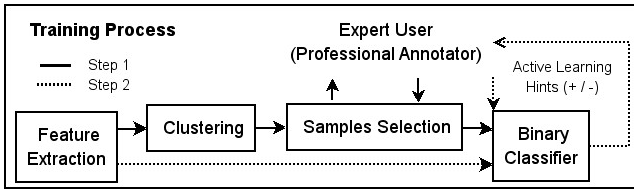


Fig. 1 Framework for training a SVM classifier

First step uses clustering to assist the professional annotator in selecting image samples
Latter step applies AL to refine the classifier model

membership degree in a cluster associated with a different concept than the one expected for the corresponding image. Focusing on these sensitive points facilitates the definition of hyperplanes between clusters containing images belonging to certain class (or concept). These positive and negative samples constitute the candidates of training patterns. Then, the annotator follows a sample selection procedure to decide if the candidates are suitable samples to obtain a basic classifier model. This procedure is shown in Fig. 2.

The second step in the training process applies AL (likewise relevance feedback) to refine the classifier model. The classifier predicts positive examples for the category from the unlabelled images, using the clusters as search space. The professional annotator provides hints indicating positive and negative images found among the classification results. These annotator's hints are collected to update the training data set. Furthermore, both positive and negative images are used to refine the classifier design. The introduced knowledge accumulated during the training interactions is used to increase the problem domain knowledge and enable long-term learning. The framework's components are detailed subsequently.

3.2 Unsupervised clustering

Clustering methods help to organise low-level features into groups, the interpretation of which may relate to some description task pertaining to the image content. Thus, features are clustered according to similarities among them [19]. Such a similarity between patterns is quantified or measured using a proximity metric. If the clusters are

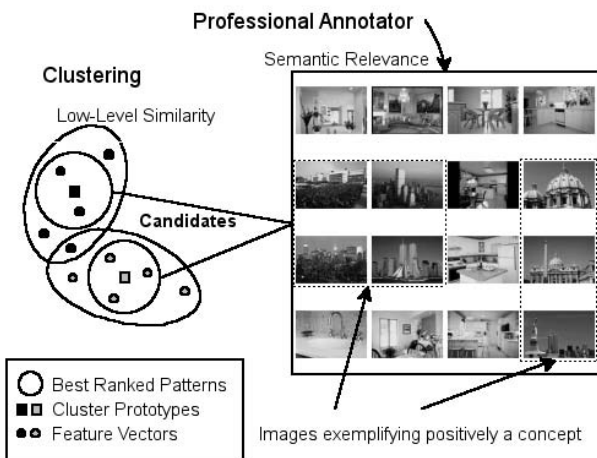


Fig. 2 Finding design samples for a first training round

Similarity at low-level is provided by clustering algorithm
Professional annotator indicates relevant images to the concept

described by an equivalence class

$$[v_j]_E \doteq \{x_i : x_i \in X, E(v_j, x_i) = 1\} \quad (1)$$

where v_j ($1 \leq j \leq c$) is a cluster prototype and x_i ($1 \leq i \leq N$) is a feature vector associated with the i th image in the data set X . Then the set of equivalence classes

$$\frac{X}{E} \doteq \{[v_j]_E\} \quad (2)$$

called a quotient set forms a partition of the feature space. Consequently, the clustering outcome can be used as a pre-processing classification procedure based on the map from X onto X/E , which is defined by

$$\phi: X \mapsto \frac{X}{E} \quad (3)$$

The cluster assignment is inherently unsupervised as no prior information about the data structure is utilised in the algorithm. However, objective function-based clustering methods can be used to determine the underlying structure of the training data set. Thus, the clustering results provide valuable information that can be exploited to assist a professional annotator in establishing links between the feature vectors and the concepts.

The nature of the problem demands an extension to deal with the underneath subjectivity and fuzziness of the human interpretation [20]. In the proposed framework, the clustering task is carried out using the standard Fuzzy C-Means (FCM) presented by Bezdek [21]. FCM is an optimisation technique based on minimisation of the objective function that measures the desirability of partitions of the data space. The objective (or criterion) function is a scalar index that indicates the quality of the partition and has the form

$$J(X, V, U) = \sum_{i=1}^N \sum_{j=1}^c u_{ij}^m d^2(x_i, v_j) \quad (4)$$

where X is a data space consisting of N p -dimension feature vectors to cluster, V is a set of c ($2 \leq c \leq N$) cluster prototypes and U is a matrix belonging to the set of all possible fuzzy partitions defined by

$$\mathfrak{S} = \left\{ U \in \mathfrak{R}_{Nc} \left| \begin{array}{l} \forall_{\substack{1 \leq i \leq N \\ 1 \leq j \leq c}} u_{ij} \in [0, 1], \\ \sum_{j=1}^c u_{ij} = 1, \\ 0 < \sum_{i=1}^N u_{ij} < N \end{array} \right. \right\} \quad (5)$$

where u_{ij} is the degree of membership of vector x_i in the cluster j , v_j is the p -dimension prototype of the cluster and m ($1 < m < \infty$) is a fuzzy exponent that determines the degree of overlap of fuzzy clusters. Setting $c = 2$ produces the minimum partition, whereas $c = N$ is the data space itself.

$d^2(\cdot)$ is any distance norm expressing the similarity between any feature vector and the prototype, formally defined as

$$d_{ij}^2 \doteq \|x_i - v_j\|_A^2 = (x_i - v_j)^T A (x_i - v_j), \quad 1 \leq i \leq N; \quad 1 \leq j \leq c \quad (6)$$

where A is the identity matrix for Euclidean distance and inverse of variance-covariance matrix of X for Mahalanobis distance.

The minimisation of the fuzzy objective function is a non-linear optimisation problem that can be solved using

the Picard iteration with the $\|U_{(t+1)} - U_{(t)}\| < \delta$ criterion. A deficiency is presented when after a number of iterations the solution converges to local minima, which is not necessarily the optimal one. The solution is unique or optimal if the prototypes of the clusters are always the same regardless of the initial partition matrix $U_{(0)}$.

An illustrative example of using clustering as preprocessing mechanism to find suitable samples is presented in Fig. 3. FCM provides the cluster prototypes as well as the feature space partition. Membership degrees of patterns to each cluster are used to collect candidate images of design samples. The best-ranked images, it is to say the nearest patterns to the prototypes, are organised into sets following the cluster partitions. These sets are presented to the annotator who selects images that positively and negatively exemplify the concept as design samples to train the classifier in a first round. A basic classifier model is obtained using these samples.

As the number of classes in semantic-based image classification can be predetermined, the optimal number of clusters to partition the data space can be equal to the class set cardinality. Subsequently, validity functions [22, 23] such as the fuzziness performance index [24] or the compactness and separation [25, 26] can be used.

3.3 Binary classifier

With a choice of many options for binary classification that could later lead to multi-class approaches, we have made use of good performances of SVC. Although this group of classifiers shows good performance for the generalisation task over various pattern recognition and information retrieval problems [27] it can also achieve good results with small training data sets [28], which makes it extremely appealing for our framework. In this section, will introduce basic concepts of SVCs and then we will explain the role they have in our system.

The idea of supervised learning approach is incorporated within this classifier, it tries to empirically model a system that would classify or predict an accurate response of unseen data set based on limited training patterns. The

idea of the SVC is based on structural risk minimisation (SRM) principle, minimising not only empirical risk but also the upper bound of the expected risk. Suppose we have a training data set generated by an unknown probability distribution and assuming only two classes represented as

$$(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \dots, (\mathbf{x}_N, \mathbf{y}_N) \in \mathbf{R}^p \times \{-1, +1\} \quad (7)$$

where $(\mathbf{x}_i, \mathbf{y}_i)$ is a data sample and $\mathbf{y}_i = \Omega(\mathbf{x}_i) = 1$ represents the label if \mathbf{x}_i satisfies the designer-defined criterion regarding to a given concept and $\mathbf{y}_i = \Omega(\mathbf{x}_i) = -1$ when it does not. N is the overall number of available training samples. Here $\Omega(\cdot)$ denotes the classifier that maps input patterns into one of the classes

$$\Omega: \mathbf{R}^p \rightarrow \{-1, 1\} \quad (8)$$

SRM principle is based on minimising the training error and choosing a function class such that the upper bound of the test error is minimised.

If a class of hyperplanes is considered in the following form

$$(\mathbf{w} \cdot \mathbf{x}) + b = 0, \quad \mathbf{w} \in \mathbf{R}^p, \quad b \in \mathbf{R} \quad (9)$$

The corresponding decision function for the SVC classifier is denoted as

$$f(\mathbf{x}) = \text{sgn}((\mathbf{w} \cdot \mathbf{x}) + b) \quad (10)$$

And \mathbf{w} , b represent the normal weight vector and a bias of the hyperplane, respectively, that separates data samples based on their position from the hyperplane in the p -dimensional feature space.

The separating hyperplane is optimal if it separates a set of patterns without error and maximises the margin, the distance between feature vectors from each class that are closest to the hyperplane.

The optimal solution for separating hyperplane can be obtained by maximising the margin and having in mind that the problem is often noisy. This corresponds to maximising the minimal distance between convex hulls of both

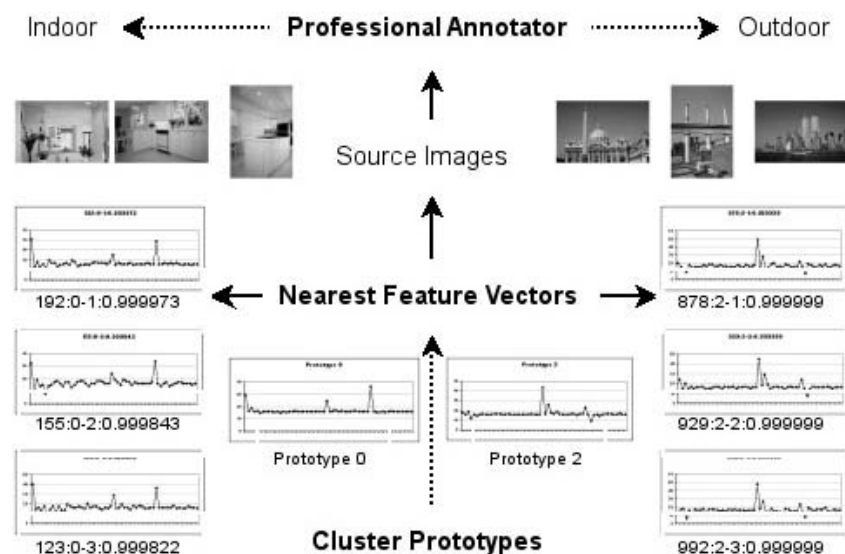


Fig. 3 Selection of sample images

Examples for training the classifier are chosen from nearest feature vectors to the clusters. The selection is based on similarity between vectors and cluster prototypes. In this case, feature vectors correspond to colour layout descriptions (58 bin histograms)

classes equal to $2/\|\mathbf{w}\|$ and introducing a slack variable to relax the hard margin constraint

$$\min \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_i^N \xi_i \quad (11)$$

C is the balancing factor between minimisation of the empirical risk and maximisation of the margin between classes. The above expression is valid under the following constraints

$$y_i((\mathbf{w} \cdot \mathbf{x}) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \dots, N \quad (12)$$

A solution to the minimisation problem is introduced through optimisation of (11) and (12), representing a quadratic problem often solved by conversion to Wolfe dual [29]. The later is easier to solve by minimising the Lagrangian with respect to primal variables \mathbf{w} , b and maximising with respect to dual variables α_i , leading to the following optimisation problem

$$\max_{\alpha} \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) \quad (13)$$

$$\sum_{i=1}^N \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C, \quad i = 1, \dots, N \quad (14)$$

From the Kuhn–Tucker complementary conditions, we have the following condition

$$\alpha_i \cdot [y_i((\mathbf{x}_i \cdot \mathbf{w}) + b) - 1] = 0, \quad i = 1, \dots, N \quad (15)$$

On the basis of (15) only points with non-zero α_i values can be support vectors (SVs). They lie on the margin, define it and are the only relevant samples from the training set.

In cases when the linear bound does not facilitate class separation, a non-linear mapping of input space into a higher dimensional vector feature space may enable a linear separation boundary. The non-linear mapping is denoted as

$$\Phi: \mathcal{R}^p \longrightarrow \mathfrak{S} \quad \mathbf{x} \longrightarrow \Phi \mathbf{x} \quad (16)$$

The optimisation task of (13) and (14) is solved by using the fact that only the inner product of training patterns is needed to define the hyperplane. Therefore SVC uses a kernel function $k(\mathbf{x}, \mathbf{x}')$ instead of directly calculating the inner products. In the feature space \mathfrak{S} , the inner product is represented as a kernel, with the similarity measure between two input vectors being

$$k(\mathbf{x}, \mathbf{x}') = (\Phi(\mathbf{x}) \cdot \Phi(\mathbf{x}')) \quad (17)$$

If k is a continuous kernel of a positive integral over a Hilbert space with each kernel function satisfying Mercer's condition [27], then the kernel k is a valid inner product in the feature space.

Introducing some of the kernels that could be used, we mention Gaussian radial basis function (RBF) kernel. This is a universal kernel [30], meaning that a linear combination of RBF kernel functions can approximate any continuous function. The appropriate feature space is then of infinite dimension and given any labelled data set, a linear hyperplane can be found, which separates classes in the Gaussian feature space [30].

As for most machine learning processes for SVM based approaches, there are also a number of parameters and decisions that need to be made in order to generate a classification model that would perform well on unseen

data (e.g. the upper bound for Lagrange multipliers C , standard deviation in RBF kernel).

One of the frequently used kernel functions is RBF kernel given by

$$K_{\text{Gaussian}}(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{2\sigma^2}\right) \quad (18)$$

The goal is to find optimal parameters and weighting for the kernel function used with respect to the scenarios.

We have carried out a number of tests using a classical RBF kernel given by (18). In this case, an on-line parameter adaptation is very time demanding, as we do not know the testing data set for AL in advance. In this effort, we have also tested a modification of RBF kernel for SVM hoping to adopt it to the non-linear behaviour of low-level feature vectors used. The proposed modification uses SVM and employs kernel-learning approaches to optimise the non-linear mapping introduced with kernels for a better correspondence to the chosen features.

In our tests, several image descriptors are combined in order to improve the effectiveness of the classifier. This raises the need of using appropriate distances for each descriptor as norms within the RBF kernel. In our approach, the kernel within SVM has the following form

$$K_{\text{Gaussian}}(\mathbf{x}, \mathbf{y}) = \exp\left(\frac{-d(\mathbf{x}, \mathbf{y})}{2\sigma^2}\right) \quad (19)$$

the distance $d(\mathbf{x}, \mathbf{y})$ is a linear combination of dynamically weighted and normalised distances for each descriptors used, on the basis of the MPEG-7 standard [31].

$$d(\mathbf{x}, \mathbf{y}) = \sum_i \mathbf{w}_i \bar{d}_i(\mathbf{x}, \mathbf{y}) \quad (20)$$

Weights calculation relies on the assumption that a particular descriptor somehow resembling the user preferences obtains higher weight. Thus, the weights for each descriptor are determined as inverse of variance over all positive examples given by the designer. There is no assurance that the new kernel satisfies the Mercer's condition, guaranteeing kernels to be real-inner products. Although it is possible to still apply the SVM to such kernels, there is no longer assurance that the optimal hyperplane maximises the margin [32], but we have empirically observed higher consistency and improvement in performance.

3.4 Active learning

As depicted in Fig. 1, the system captures hints of domain knowledge, which relate to the classification problem. During the second step of the training process, a professional annotator provides hints indicating to the classifier

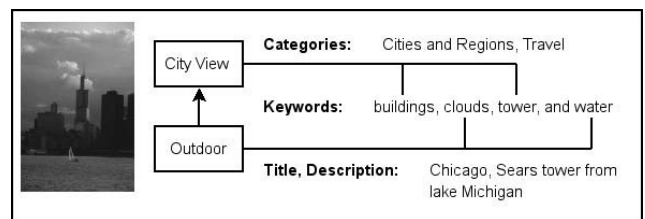


Fig. 4 The keyword-oriented classification is useful to index more labels

Category names used in the experiments (e.g. outdoor and city view) can be linked to more elaborated annotations (e.g. cities and regions, buildings, clouds)



Fig. 5 Top-ten of ranked images by highest memberships in the clusters

Categories: indoor and outdoor
 Low-level similarity based on colour features
 Each row corresponds to a representative set of the cluster

whether or not its decisions were correct or not (positive or negative hint).

The classifier uses those hints to adjust the boundaries between patterns containing (or not) the concept. These boundaries are defined by the hyperplane based on the SVs.

The idea of this supervised learning step is not to estimate distributions of the known/unknown patterns, but to learn the SVs. These vectors define the optimal non-linear decision hyperplane and are determined from the known training set.

4 Experimental studies

4.1 Test conditions

Experiments were conveyed with images selected from Corel stock gallery. Two groups consisting of 1035 and 1200 photographs, respectively, were organised into a number of semantic categories. The first group was used to classify indoor (kitchens and bathrooms, office interiors, museums, etc.) and outdoor (contemporary buildings, Rome, Chicago, architecture 1 and 2, etc.) images. The second group was used to classify animals (dogs, tropical sea life, etc.), city views (New York city, Ottawa, etc.), landscapes (autumn, Yosemite, etc.) and vegetation (perennial plants, American gardens, etc.) images.

As illustrated in Fig. 4, the category names were simplified according to the objectives of the case studies. Professional annotations of Corel images involve more information: title, categories and keywords (fotosearch.com/corel). It is

worth to stress that keyword-oriented classification is useful to describe images with a controlled vocabulary. The keywords can also be used to search related annotations in semantic ontology models.

The indoor/outdoor feature space was built with vectors containing colour layout descriptions (58 bin histograms), whereas the animal/city view/landscape/vegetation feature space combines colour structure, edge histogram and homogeneous texture descriptions (398 bin histograms). Each of these MPEG-7 descriptors has a particular syntax and semantics [31]. The matching procedures in the experiments use the basic L2 norm.

Training data sets were randomly generated with 60% of the images. The remaining images (40%) were used for testing the classifier model.

4.2 Clustering analysis

Clustering results for the indoor/outdoor classification problem indicate that colour is an appropriate descriptor to create a separable feature space in this domain. The similarity of best-ranked images in the five clusters, on the basis of their membership degrees, resembles partially the expected semantic grouping.

As depicted in Fig. 5, the first set (row 1) contains samples of indoor images, except the fifth image that corresponds to a building façade close-up. In the sequel, most of the displayed images are good candidates of outdoor (sets 2 and 3) and indoor (sets 4 and 5). The sixth and tenth images

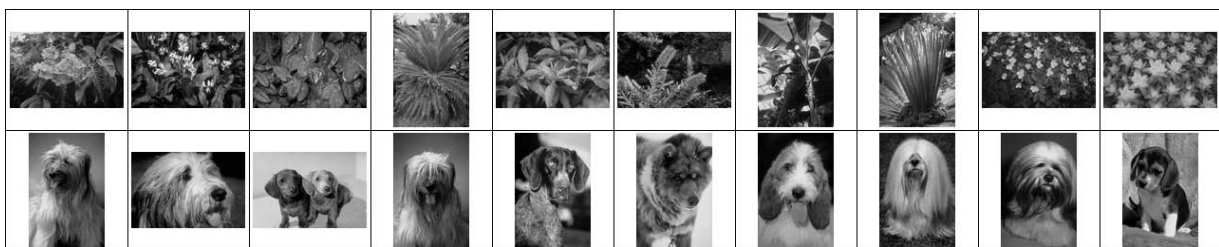


Fig. 6 Sets satisfying criteria for the semantic categorisation: vegetation (row 1) and animal (row 2)



Fig. 7 *Category overlapping*

Row 1: landscape-city view; row 2: vegetation-city view; row 3: animal-vegetation; and row 4: city view-landscape

in the fourth set (row 4) are negative examples of indoor category, although their colour distribution is closer to the prototype of this group. Using low-level similar images as negative examples, helps the classifier in defining the optimal non-linear decision hyperplane.

Following figures contain the best-ranked images in the clusters for the classification problem of categories animal, city view, landscape and vegetation. Low-level similarity is based on colour and texture features. Feature space was partitioned into ten clusters.



Fig. 8 *Sample of sets whose content mixed objects from different categories*

Sets 1 (row 1), 5 (row 2) and 6 (row 3) are exemplars of how low-level similarity can be derived in semantically meaningless grouping



Fig. 9 *Samples selection can be affected by images containing objects from another category*



Fig. 10 *Samples of images do not match clearly the semantic criteria*

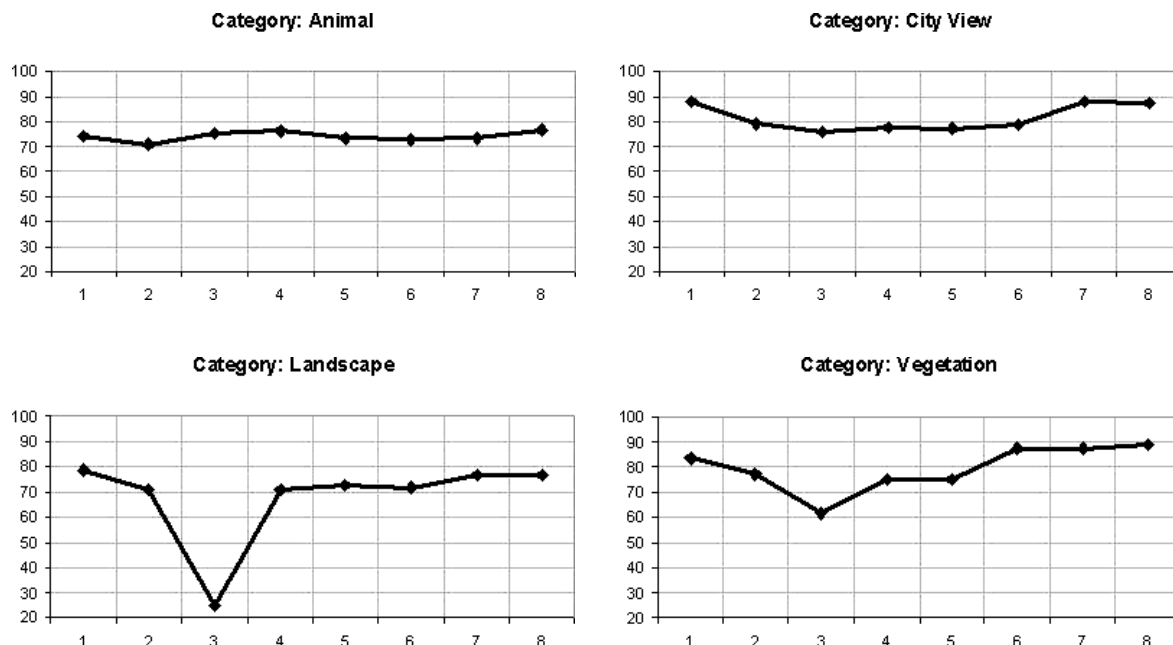


Fig. 11 Classification results using random selection of images

x-axis indicates the number of iteration in which the annotator provides new samples to the classifier
y-axis shows the resulting accuracy

Fig. 6 shows a sample of image sets satisfying criteria for the semantic categorisation. It means that images found in each set can be directly attached to a category as follows: set from cluster 4 (row 1) to vegetation and set from cluster 7 (row 2) to animal.

Fig. 7 gives a sample of image sets overlapping criteria for the semantic categorization. The first two rows corresponding to sets taken from clusters 8 and 9, respectively, present a minimum overlap. The first set can be ascribed as landscape except by the last image (tenth column), which is a sample of city view; the second one satisfies criteria of category vegetation except by the image in the second column containing a city view scene. The third and fourth rows show overlapping between categories animal–vegetation and city view–landscape with strong commonalities in their distributions of colour and texture descriptions. This fact is reflected in the sets derived from the clustering results.

As expected, some sets in the ranked images contain objects from more than two categories. It shows why the clusters cannot be attached to a single category. Consequently, relying on low-level similarity derives in semantically meaningless grouping (Fig. 8).

Fig. 9 displays some samples of images taken from cluster 10. Most of the images are landscapes, although there are some manmade objects or animals that could mislead their categorisation.

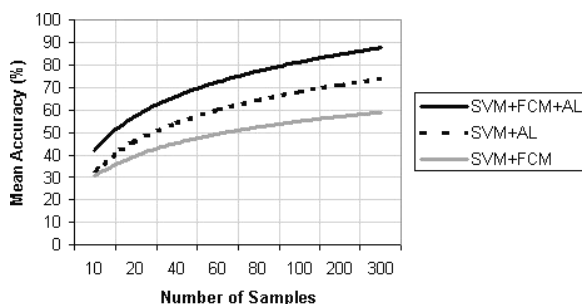


Fig. 12 Mean accuracies achieved in the indoor/outdoor classification problem using the training approaches detailed in Table 1

One of the problems in selecting training samples is the quality of the images. It can be regarded as definition (pixel resolution, colour, etc.) as well as semantic content. Fig. 10 shows images that do not match either completely or clearly the high-level categorisation. These images introduce noise in the learning process and subsequently affect the classifier performance.

4.3 Framework assessment

In order to evaluate stability of the classifier model, a set of experiments were carried out using random selection of samples. Conversely, this approach skips the clustering procedure. As can be observed in Fig. 11, the classification results lack of stability. It is because samples collection is

Table 1: Training approaches used to assess the classifier performance

Training approach	Description
SVM + FCM	SVM classifier assisted with hints provided by a professional annotator governed by clustering (FCM) results during the training phase. Samples are selected from the nearest patterns (see Figs. 5–9) to the cluster prototypes
SVM + AL	SVM classifier using only AL. The classifier is trained with hints provided by a professional annotator
SVM + FCM + AL	SVM classifier is trained combining both clustering results and AL

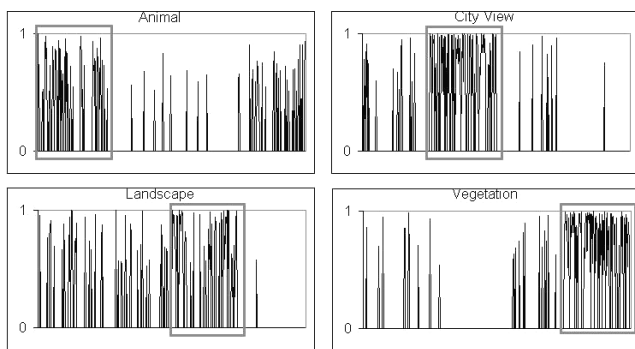


Fig. 13 Two-class classification outcomes

Input patterns are organised along the x -axis y -axis indicates the probability of membership to the corresponding category

Table 2: Performance of two-class classifiers (%)

Animal	City view	Landscape	Vegetation
74.38	87.29	74.18	87.29

based upon visual inspection along with subjective criteria of the annotator without taking into account any low-level similarity. In contrast, clustering mechanisms not only assist in the sample selection, but also contribute to the system's stability (Fig. 12).

The three training approaches summarised in Table 1 are used to assess the performance of the classifier within the proposed framework.

Mean accuracies obtained in the experimental studies are presented in Fig. 12. The lowest accuracy is obtained when the SVC learns only from clustering outcomes; the classifier

behaves better when using AL; the accuracy is improved selecting samples from clusters.

Accuracy in the first approach (SVM + FCM) decreases rapidly, although it is expected because of the sensible reduction on the required supervision. The professional annotator needs only to indicate the class label of each cluster. This lightens the burden of annotation while introducing some noise at the same time.

The second approach (SVM + AL) depends entirely on the images shown to the user. An inconvenience here is the overall subjectivity because of the fact that selection of sample relies completely on the images ignoring any relationship (low-level similarity) between the image descriptions.

The third approach (SVM + FCM + AL), corresponding to the proposed method, shows a higher performance. It also has the advantage of taking into account the underlying low-level structures (revealed by the clusters). It minimises the required supervision and partially exploits the semantic information provided from the professional annotator.

When it comes to the multi-class problem, several interesting classification scenarios may arise, which in the sequel lead to a certain quantification of the results obtained in this manner. The two-class classifiers may produce the following outcome: (1) only one classifier identifies the class, (2) none of the classifiers identify a class; this is described as lack of decision, (3) a few classifiers identified several classes; this is described as lack of specificity of classification. Under these circumstances, two situations may occur. First, the correct class is within the set of these classes. The result is correct but not specific. Second, the correct class is not in these classes being identified by the classifiers. In this case, the classification result is neither correct nor specific. In the latter case, the two-class classifier with higher probability defines the class to be assigned. The probability of membership for


Animal	 0.981087	 0.976889	 0.964562	 0.954610	 0.940054
City View	 0.999999	 0.999999	 0.999997	 0.998680	 0.998563
Landscape	 0.998122	 0.997713	 0.997342	 0.996973	 0.995615
Vegetation	 0.997746	 0.996170	 0.99457	 0.992258	 0.991586

Fig. 14 Samples of images correctly classified

Probability is indicated below each image

Animal	 0.932426 Vegetation	 0.927963 Vegetation	 0.904276 Vegetation	 0.902754 Vegetation	 0.843108 Vegetation
City View	 0.980284 Landscape	 0.961845 Animal	 0.945897 Animal	 0.904977 Landscape	 0.896745 Landscape
Landscape	 0.99898 Animal	 0.994494 City View	 0.99264 Animal	 0.960764 City View	 0.958875 City View
Vegetation	 0.94527 Animal	 0.93526 Animal	 0.898635 Landscape	 0.863435 Animal	 0.855575 Animal

Fig. 15 Samples of misclassified images as the assigned categories on far left column
Probability and expected class are indicated below each image

each image to a class is represented through training SVM and fitting parameters of additional sigmoid function to posterior probability of the classes [33]. It is illustrated in Fig. 13. Input patterns are organised along the x -axis. The y -axis corresponds to the obtained probability in the corresponding two-class classifier. The boxes indicate the expected category.

Table 2 presents accuracies achieved by the two-class classifiers. Some samples of correctly classified and misclassified images are given in Fig. 14 and Fig. 15.

5 Conclusions

A framework to assist efficiently a professional annotator in choosing image samples to train a semantic classifier was presented. The approach uses clustering mechanisms to reveal the underlying structure in training data in order to shift low-level features toward high-level information.

The training process applies AL to capture hints from the annotator. Problem domain knowledge is accumulated in order to enable long-term learning. This learning mode reduces the burden of collecting samples randomly as well as improves the quality of the chosen ones taking into account low-level similarity. The AL is also a practical way to introduce system's adaptation and can be extended onto the generalisation stage in the form of relevance feedback.

The applied keyword-oriented classification is useful to describe images with a controlled vocabulary. These keywords can also be used to search related annotations in semantic ontology models.

The real challenge in using classifiers with kernel methods is in the scarce training data sets available and

necessity for real-time optimisation, which makes off-line adaptation method very impractical. These training methods generally rely on substantial data training sets and sensitive parameter tuning. Relatively high precision and accuracy is possible to achieve, but dependant on the data samples used and therefore very random. Hence, learning kernel matrix from data samples without the need to set any parameters is the initial idea for the presented kernel manipulations and for future research.

6 Acknowledgments

Support from the Natural Sciences and Engineering Research Council (NSERC) and Canada Research Chair (W.P.) is gratefully acknowledged. Part of the research leading to this paper was also done within the framework of the project aceMedia – Integrating knowledge, semantics and content for user-centred intelligent media services (www.acemedia.org).

7 References

- 1 Simon, J.C.: 'Recent progress to formal approach of pattern recognition and scene analysis', *Pattern Recognit.*, 1975, 7, pp. 117–124
- 2 Gorkani, M.M., and Picard, R.W.: 'Texture orientation for sorting photos "At a glance"'. Proc. IEEE IAPR, 1994, vol. 1, pp. 459–464
- 3 Vailaya, A., Figueiredo, M.A.T., Jain, A.K., and Zhang, H.-J.: 'Image classification for content-based indexing', *IEEE Trans. Image Process.*, 2001, 10, pp. 117–130
- 4 Vailaya, A., Jain, A., and Zhang, H.-J.: 'On image classification: city vs. landscape'. Proc. IEEE Workshop on Content-Based Access of Image and Video Libraries, 1998, pp. 3–8
- 5 Dorai, C., and Venkatesh, S.: 'Bridging the semantic gap with computational media aesthetics', *IEEE Multimedia*, 2003, 10, p. 15

- 6 Jain, A.K., Duin, P.W., and Mao, J.: 'Statistical pattern recognition: a review', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2000, **22**, (1), pp. 4–37
- 7 Saitta, L., and Bergadano, F.: 'Pattern recognition and Valiant's learning framework', *IEEE Trans. Pattern Anal. Mach. Intell.*, 1993, **15**, (2), pp. 145–155
- 8 Bhanu, B., and Dong, A.: 'Concepts learning with fuzzy clustering and relevance feedback', *Eng. Appl. Artif. Intell.*, 2002, **15**, pp. 123–138
- 9 Boutell, M.R., Luo, J., Shen, X., and Brown, C.: 'Learning multi-label scene classification', *Pattern Recognit.*, 2004, **37**, pp. 1757–1771
- 10 Rui, Y., Huang, T.S., Ortega, M., and Mehrota, S.: 'Relevance feedback: a power tool for interactive content-based image retrieval', *IEEE Trans. Circuits Syst. Video Technol.*, 1998, **8**, (5), pp. 644–655
- 11 Nakazato, M., and Huang, T.S.: 'Extending image retrieval with group-oriented interface'. IEEE Conf. on Multimedia and Expo, 2002, pp. 201–204
- 12 Yoshizawa, T., and Schweitzer, H.: 'Long-term learning of semantic grouping from relevance-feedback'. Proc. 6th ACM SIGMM Int. Workshop on Multimedia Information, 2004, pp. 165–172
- 13 Tong, S., and Chang, E.: 'Support vector machine active learning for image retrieval'. Proc. 9th ACM Int. Conf. on Multimedia, 2001, pp. 107–118
- 14 Smith, J.R., Tseng, B., Naphade, M.R., Lin, C.Y., and Basu, S.: 'Learning to annotate video databases'. Proc. SPIE Conf. on Storage and Retrieval on Media Databases, 2002
- 15 Zhang, C., and Chen, T.: 'An active learning framework for content based information retrieval', *IEEE Trans. Multimedia*, 2002, **4**, (2), pp. 260–268
- 16 Valiant, L.G.: 'A theory of the learnable', *Commun. ACM*, 1984, **27**, (11), pp. 1134–1142
- 17 Nguyen, H.T., and Smeulders, A.: 'Active learning using pre-clustering'. Proc. 21st Int. Conf. on Machine Learning, 2004
- 18 Duda, R.O., Hart, P.E., and Stork, D.G.: 'Pattern classification' (Wiley-Interscience, 2001, 2nd edn.)
- 19 Jain, A.K., Murty, M.N., and Flynn, P.J.: 'Data clustering: a review', *ACM Comput. Surv.*, 1999, **31**, (3), pp. 264–323
- 20 Pedrycz, W.: 'Fuzzy sets pattern recognition: methodology and methods', *Pattern Recognit.*, 1990, **23**, (1/2), pp. 121–146
- 21 Bezdek, J.: 'A convergence theorem for the ISODATA clustering algorithms', *IEEE Trans. Pattern Anal. Mach. Intell.*, 1980, **2**, (1), pp. 1–8
- 22 Pedrycz, W.: 'Knowledge-based clustering: from data to information granules' (J. Wiley, New York, NY, USA, 2005)
- 23 Halkidi, M., Batistakis, Y., and Vazirgiannis, M.: 'On clustering validation techniques', *J. Intell. Inf. Syst.*, 2001, **17**, (2–3), pp. 107–145
- 24 Roubens, M.: 'Fuzzy clustering algorithms and their cluster validity', *Eur. J. Oper. Res.*, 1982, **10**, pp. 294–301
- 25 Xie, X.L., and Beni, G.: 'A validity measure for fuzzy clustering', *IEEE Trans. Pattern Anal. Mach. Learn.*, 1991, **13**, pp. 841–847
- 26 Xie, Y., Raghavan, V.V., and Zhao, X.: '3M algorithm: finding an optimal fuzzy cluster scheme for proximity data'. Proc. IEEE Int. Conf. Fuzzy Systems, 2002, vol. 1, pp. 627–632
- 27 Müller, K.-R., Mika, S., Rätsch, G., Tsuda, K., and Schölkopf, B.: 'An introduction to kernel-based learning algorithms', *IEEE Neural Netw.*, 2001, **12**, (2), pp. 181–201
- 28 Duin, R.: 'Classifiers in almost empty spaces'. Proc. 15th Int. Conf. on Pattern Recognition, 2000, vol. 2, pp. 1–7
- 29 Keerthi, S.S., Shevade, S.K., Bhattacharyya, C., Murthy, K.R.K.: 'Improvements to Platt's SMO algorithm for SVM classifier design'. Technical Report, Department of CSA, Bangalore, India, IISc, 1999
- 30 Perez-Cruz, F., and Bousquet, O.: 'Kernel methods and their potential use in signal processing', *IEEE Signal Process. Mag.*, 2004, **21**, (3), pp. 57–65
- 31 Manjunath, B.S., Ohm, J.-R., Vasudevan, V.V., and Yamada, A.: 'Color and texture descriptors', *IEEE Trans. Circuits Syst. Video Technol.*, 2001, **11**, (6), pp. 703–715
- 32 Chapelle, O., Haffner, P., and Vapnik, V.N.: 'Support vector machines for histogram-based image classification', *IEEE Trans. Neural Netw.*, 1999, **10**, (5), pp. 1055–1064
- 33 Platt, J.: 'Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods', in Smola, A., Bartlett, P., Schölkopf, B., and Schuurmans, D. (Eds.): 'Advances in large margin classifiers' (MIT Press, 1999), pp. 61–74

Copyright of IEE Proceedings -- Vision, Image & Signal Processing is the property of Institution of Engineering & Technology and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.