Multifeature analysis and semantic context learning for image classification

ZHANG, Q; Izquierdo, E

Multi-Feature Analysis and Semantic Context Learning for Image Classification

QIANNI ZHANG and EBROUL IZQUIERDO, Queen Mary, University of London

This paper introduces an image classification approach in which the semantic context of images and multiple low-level visual features are jointly exploited. The context consists of a set of semantic terms defining the classes to be associated to unclassified images. Initially, a multi-objective optimisation technique is used to define a multi-feature fusion model for each semantic class. Then, a Bayesian learning procedure is applied to derive a context model representing relationships among semantic classes. Finally, this context model is used to infer object-classes within images. Selected results from a comprehensive experimental evaluation are reported to show the effectiveness of the proposed approaches.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval models*; I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis—*Object recognition*

General Terms: Algorithm

Additional Key Words and Phrases: Image classification, Object detection, Multi-feature fusion, semantic context modelling

ACM Reference Format:

Zhang, Q. and Izquierdo, E. 2012. Multi-Feature Analysis and Semantic Context Learning for Image Classification. ACM Trans. Appl. Percept. 0, 0, Article 0 (2012), 20 pages.
DOI = 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

1. INTRODUCTION

This paper addresses the challenging problem of automatic image classification and tagging. The aim is to derive high-level concepts or tags for images by jointly exploiting the effects of interrelated semantic classes and low-level visual features. The novel contributions of this work lays in two main aspects: First, an approach to derive multi-feature fusion models using a suitable multi-objective optimisation technique. This approach is referred to as Multi-Feature Learning (MFL). Second, a novel Semantic Context Learning (SCL) method is developed to effectively promote image classification performance by exploring contextual relationships in images. The proposed approach uses a block-based representation scheme for analysing visual elements in images.

In previously reported works on image classification and retrieval, accuracy is often limited due to the unreliable outputs from the classifiers. This is because classifiers often rely on single low-level features, which are not always capable of interpreting visual objects into varying and complicated se-

Author's address: Q. Zhang and E. Izquierdo, School of Electronic Engineering and Computer Science, Queen Mary, University of London, London, U.K; email: {qianni.zhang and ebroul.izquierdo}@eecs.qmul.ac.uk

DOI 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

mantic meanings. To tackle this problem, the combination of low-level features for image classification has been widely considered in the literature. However, different visual features and their similarity measures cannot be combined straightforwardly [Deselaers et al. 2008]. Thus, the question related to the definition of a models joining several similarity functions requires careful consideration. A number of approaches have been proposed to address this question. One general approach was to append different feature vectors into a large global feature vector or combine distances obtained in each feature space [Chun et al. 2008; Li and Wang 2003; Jain and Vailaya 1996; Berman and Shapiro 1999; Rui et al. 1998]. Weights obtained using different methods can be applied on each feature space according to their importance. Some research used a generic programming framework to derive a nonlinear combination of image similarities [Torres et al. 2009]. Another general approach related to merging multiple classifiers that are trained on individual features, for example, using multiple kernel learning (MKL) for an optimal combination of kernels which captured different feature channels [Zhang and Ye 2009; Gehler and Nowozin 2009; Vedaldi et al. 2009]. Alternatively, several other computer vision approaches were based on local interest point detectors and descriptors invariant to geometric and illumination variations [Mikolajczyk and Schmid 2005; Csurka et al. 2004].

In this paper a different MFL approach is used to enable multi-feature based image classification. It relies on a multi-objective learning algorithm for deriving a unique semantic based multi-feature fusion model [Zhang and Izquierdo 2007]. Such a model aims to represent the most descriptive visual patterns of an object class considering different visual aspects. Here, the MFL approach is used to learn a suitable model for combining features and metrics. The core of MFL is the widely used Multi-Objective Optimisation (MOO) strategy, described in [Steuer 1986; Knowles and Corne 2000]. The main advantage in this strategy is that it is able to find a general optimum across potentially conflicting visual signatures for a concept. Thus, the obtained multi-feature model simultaneously encapsulates different aspects of the most representative visual patterns for each unique concept, without however assigning fixed relevance factors to each one of the used visual features. To the best of our knowledge, the MOO strategy has never been applied in feature fusion and content based image classification scenarios before the authors' earlier work [Zhang and Izquierdo 2007], in which the initial version of MFL approach was proposed. In the initial version, the focus of MFL was on finding suitable feature weighting factors that can later be used to make simple rankings of images. Based on the previous work, the new contribution in this paper is that the MFL approach is integrated into a classifier so that it can be directly used for an image classification and annotation purposes. More importantly, from implementation perspective, the new MFL approach has been carefully optimised to ensure its capability of handling large datasets. It is worth noting that the MFL approach focuses on deriving suitable fusion models for multiple features. Instead of relying on any of the single features, the MFL approach can use any set of features as proposed in the literature. The used features could be MPEG-7 visual features [Chang et al. 2001], well developed colour and texture features [Swain and Ballard 1991; Tuceryan and Jain 1993], local distinctive image descriptors [Lowe 2004], codebooks obtained using bag-of-features approach [Zhang et al. 2011], etc. The proposed multi-feature fusion model is not restricted to the features being employed. However, comparison of different visual features is out of the scope of this paper.

In realistic image classification scenarios, images are complex and they usually consist of many semantically meaningful objects interrelated to each other. The SCL method is designed for jointly analysing visual-semantic context information hidden in the content based on Bayesian models. Using the SCL method, the visual-semantic context can be learned and exploited to further improve the classification power of the MFL approach. In the literature, classification of visual content according to their semantic meanings can incorporate domain knowledge to influence the decisions by context, grammar, semantics and other high-level information. In this aspect, schemes for conceptual

synthesising and reasoning at semantic level have been developed, including statistical association, conditional probability distributions, different kinds of monotonic and non-monotonic and fuzzy relationships [De Jong et al. 2007; Fan et al. 2008; Koskela et al. 2007; Lavrenko et al. 2004; Zhu et al. 2005]. Many object-based image classification frameworks relied on probabilistic models for semantic context modelling and inference [Kherfi and Ziou 2007; Aksoy et al. 2005; Vailaya et al. 1998; Zhang and Zhang 2007; Li and Wang 2003]. In contrast to object-based classification solutions, some works use context modelling in scene classification e.g., *indoor* vs. *outdoor* or *cityview* vs. *landscape* [Vailaya et al. 1998; Rasiwasia and Vasconcelos 2009]. Context modelling and exploitation has also been widely applied in the field of video annotation and retrieval, based on techniques such as boosted conditional random field (CRF) [Jiang et al. 2007] or domain adaptive semantic diffusion technique [Jiang et al. 2009]. The SCL method proposed in this paper is different from the existing research works targeting the task of object based image classification, in the sense that it can automatically learn a *visual-semantic context model* from a small training set without relying on dedicated model designs that are restricted to particular scenarios. Thus the definition of the model structure is a fully automatic learning process. There are two main advantages in such a method. First, the method is not restricted to a certain scenario or application domain but can be applied to any database with its own semantic context. Second, the user is not required to be an expert in the application scenario in order to be able to define the context model according to the domain knowledge. Rather, the context model is automatically learned using a small sized training set and the multi-feature similarities. The automatic learning process is conducted using a search-based algorithm – *K2*. The learned visual-semantic context model is then used to calculate the probabilities conveying the effects that the existing concepts have on each other.

Some works in the literature shared similar ideas in some aspects to the SCL approach. Despite of these similar aspects, SCL has its unique and novel contribution to the research topic on semantic context modelling. For example, the work in [Rasiwasia and Vasconcelos 2009] took a similar idea of context modelling using the output of some visual-feature based concept detection methods. However, this system was designed for a scene classification task; while in this paper, by using the block based representation, the proposed system tries to focus on local regions at object level in order to detect object concepts in natural complex images, which are often ignored at scene level. Moreover, their main approach used mixtures of Dirichlet distributions for learning the concept models, while the SCL uses *K2* algorithm context modelling. The work in [Tang et al. 2009b] focused on detecting concepts from user generated content including images and their associated tags using a sparse graph-based semisupervised learning approach. However, it relied on textual data for learning and classification while the proposed approach in this paper is purely content based and the intermediate semantic features are directly inferred from the visual features. [Qi et al. 2007] used a Correlative Multi-Label paradigm based on Gibbs Random Fields which simultaneously classifies concepts and models correlations between them. [Rabinovich et al. 2007] used a CRF framework to maximise object label agreement according to contextual relevance. A commonality in these three works [Qi et al. 2007; Tang et al. 2009a; Rabinovich et al. 2007], was that they constructed correlative models from external resources, such as LSCOM and Google sets, in off-line settings. Comparing to these works, SCL is a purely data-driven approach, which means it does not require any external guidance about the statistics of high-level contextual relationships, i.e. Google sets, LSCOM, etc. The inputs required by SCL is multi-feature similarities directly obtained from the MFL module and the ground-truth annotation of the class in concern. Moreover, a training set for SCL module consists of only five hundred training samples, while training sets used in MFL contain only 20 image blocks for each class. The other systems in the literature usually need significantly bigger training sets.

In addition to the two new methods, MFL and SCL, an important aspect of this research is that it uses regular image blocks as the basic units for visual representations of objects associated to semantic concepts. Feature extraction for visual representation, as well as, the generation of training sets are based on such image blocks. In the literature, most region-based approaches rely on image segmentation [Athanasiadis et al. 2007; Datta et al. 2008; Jing et al. 2004; Li et al. 2008; Natsev et al. 2004; Wang et al. 2001]. However, there are several intrinsic difficulties in using image segmentation for semantic classification. Firstly, due to the complexity of natural images, segmentation algorithms based on visual features usually segment regions, not semantically meaningful objects. Indeed, precise extraction of objects from images using automated segmentation is an open problem in computer vision. Secondly, segmentation algorithms often add heavy computation loads to the system. Moreover, some human assisted segmentation algorithms may impose burdens on users and thus are unfeasible when large databases are processed. Considering these facts, alternative schemes for image decomposition have been exploited recently [Dagli and Huang 2004]. In this paper, it will be shown experimentally that by combining this approach with other appropriately designed methods, good performance can be achieved, while maintaining a low computational load. Block regions are used as basic units without assuming precise segmentation or structures for representing objects. In general, the goal of block-region based object representation schemes is to reduce the influence of noise coming from the background and surrounding objects, and to identify suitable visual patterns for a given object without introducing errors or heavy computational cost associated with image segmentation. The work in [Vogel and Schiele 2007] demonstrated that using block regions and semantic modelling for natural scene detection, good performance can be achieved. The concepts used in this work were relatively easy concepts for block regions, such as *sky*, *cloud*, etc. In our paper, we intend to test the block-based representation scheme on some more challenging object concepts, which have clear borders of their shapes, such as *lion* and *car*. An interesting observation on our experiments, is that this 'simplistic' approach to object extraction does provide a reasonable performance when they are applied in the proposed image classification framework. However, without losing generality, the proposed framework has the full flexibility in supporting other representation schemes such as segmented regions, if effective and efficient segmentation algorithms are available [Levinshtein et al. 2009].

The rest of the paper is organised as follows: Section 2 first gives an overview of the image classification framework, in which proposed ideas of MFL, SCL and block based representation scheme, are instantiated. This section also describes the pre-processing steps in the block-based analysis scheme. Section 3 introduces the MFL approach to multi-feature fusion. Section 4 presents the SCL method. In Section 5 the experimental setup and selected results from comprehensive evaluations are presented. The paper closes with conclusions in Section 6.

2. IMAGE CLASSIFICATION FRAMEWORK AND BLOCK-BASED VISUAL ANALYSIS

The research described in this paper consists of two main aspects – the MFL approach for multi-feature fusion and the SCL method for semantic context inference, both of which rely on the block-based representation scheme. These ideas and approaches are instantiated in an integrated image classification framework for experimental purposes. The general structure of the experimental framework used in this paper is illustrated in Fig. 1. All experiments carried out and presented in this research are conducted based on this framework.

In the following of this section, we present the details on the visual analysis steps based on the block-based representation scheme for semantic objects. The reasons for using the block-based representation scheme are explained and the advantages and deficiencies are discussed. The block based analysis module of the framework comprises five processing steps: (1) Image decomposition; (2) Low-

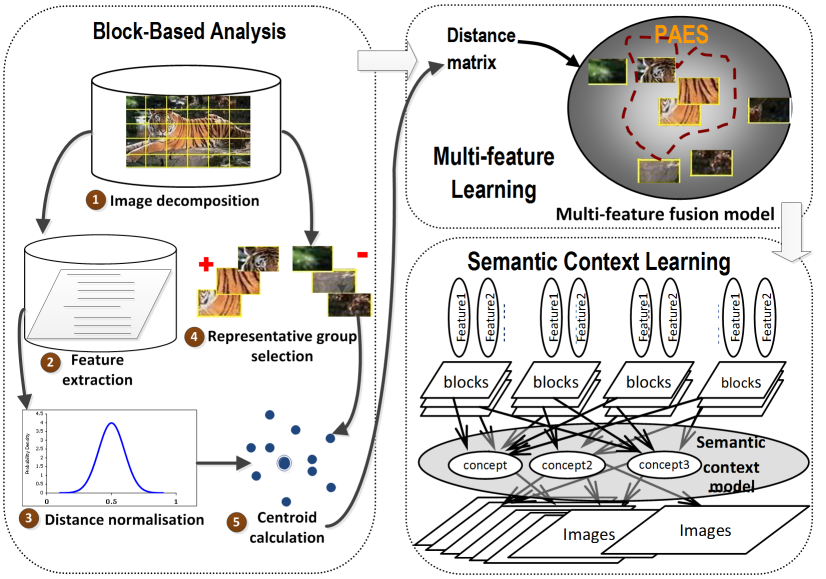

Fig. 1. Diagrammatic outline of the proposed system for multi-feature analysis and semantic context learning.

level feature extraction; (3) Distance calculation and normalisation; (4) Representative group selection; (5) Centroid calculation.

2.1 Visual analysis in block regions

This sub-section describes the first two pre-processing steps: image decomposition and low-level feature extraction. In the first step, each image in the database is partitioned into a grid of blocks. In Fig.2 an example is presented illustrating the blocks representing existing objects in an image, including *cloud*, *vegetation*, *elephant* and *rock*.

Next, a set of n low-level features, $\{F_j : j = 1, 2, ..., n\}$, are extracted from each block in the whole image database. The proposed technique is independent of the choices of low-level features. The features selected for experimental purposes in this paper are described in Section 5.

2.2 Distance calculation and normalisation

In the third step, suitable feature spaces endowed with appropriate normalised distance functions are built. Here it is assumed that each feature F_j has an inherent distance function $\tilde{d}$. Thus, it is assumed that n feature spaces $(F_j, \tilde{d})$ are available. Clearly, the approach to estimate an optimised multi-feature distance relies on comparing features and distances from different feature spaces. In most cases, the underlying distances have very different value ranges, scales and densities. To ensure comparability, a critical normalisation step is required.

Consider two blocks in the dataset Γ, and their feature vectors v_i and v_k, which are extracted from feature space F_j. The distance between the two vectors are estimated as $d' = \tilde{d}(v_i, v_k)$. Once the original distance is estimated using the dedicated metric of the feature space, a normalisation step is carried out to transform all distance values into the unitary range $[0, 1]$:

$$d = \frac{(d' - \mu)}{\sigma}, \tag{1}$$

Fig. 2. An example of block regions displaying concepts.

Table I. A bank of suitable PDFs.

Distribution	PDF	Mean	Variance
Normal $\Theta = (\mu, \sigma)$	$\frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	μ	σ^2
Gamma $\Theta = (y, \sigma)$	$x^{k-1}\frac{exp(-x/\theta)}{\Gamma(y)\theta^k}$	$y\theta$	$y\theta^2$
Laplace $\Theta = (\mu, \beta)$	$\frac{1}{2b} exp\left(-\frac{\lvert x-\mu\rvert}{\beta}\right)$	μ	$2\beta^2$
Log-norm $\Theta = (\mu, \sigma)$	$\frac{1}{x\sqrt{2\pi\sigma^2}} exp\left(-\frac{(ln\ x-\mu)^2}{2\sigma^2}\right)$	e^{μ}	$(e^{\sigma^2}-1)e^{2\mu+\sigma^2}$
Rayleigh $\Theta = (\mu)$	$\frac{x}{\sigma^2} exp\left(-\frac{x^2}{2\sigma^2}\right)$	$\sigma\sqrt{\frac{\pi}{2}}$	$\frac{4-\pi}{2}\sigma^2$
Exponential $\Theta = (\mu)$	$\lambda exp(-\lambda x)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

where d is the normalised distance, μ and σ are the mean and standard deviation of the underlying distance distribution for that feature space. To estimate the values of corresponding distribution parameters for a feature space, such as μ and σ, we approximate the true distance distributions using a set of Probability Distribution Functions (PDFs). As presented in Table I, six different commonly used PDFs are considered, including Normal, Gamma, Laplace, Log-norm, Rayleigh and Exponential. These PDFs are denoted as p_k, $k = 1, ..., 6$. Based on the distances calculated from the dataset, the parameters Θ_k of p_k are determined and the best distribution approximation is selected by evaluating the Kullback-Leibler (KL) divergence between the actual distribution and the PDFs.

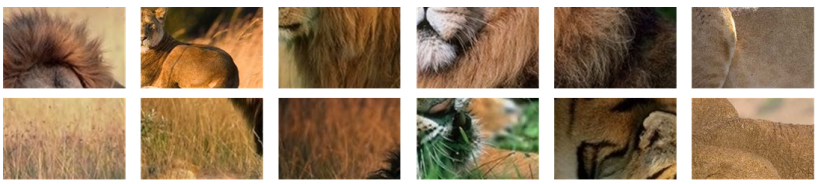

Fig. 3. Positive (first row) and negative (second row) samples of concept lion.

In the sequel, normalised distances are denoted as d and the normalised feature space as (F_j, d). Furthermore, for the sake of simplicity, a normalised feature space (F_j, d) is called a "feature space" and a normalised distance d is called a "distance".

2.3 Representative group selection and centroid calculation

In the next two processing steps, the aim is to build a suitable representation for a given object class. This is done by first selecting a list of sample blocks to visually represent the class, and then generalising a 'centroid' to indicate the center of the class in the targeted multi-feature space model.

A semantic class or object can often be represented by a single key-word or tag, e.g., *tiger*. Due to the complexity of natural images, it is hard to find any single image or image region giving an optimal visual representation of a single class. Therefore in this paper, we use a small group of block samples to approximate an optimal representation for an object class. This group of samples is referred to as the *representative group*. For the given class, let us denote such a representative group as $R = \{b_i, i = 1, 2, ..., m\}, R \in \Gamma$. Here, m is the total number of representative samples in R and b_i is a sample block.

To improve the discriminative power of the proposed classifier, two types of representative samples are considered in the proposed MFL approach. R^+ contains relevant examples for the semantic class as defined by ground-truth. They are referred to as *positive samples*. R^- contains samples which do not represent the class of concern but are easy to be classified as relevant to it. These samples are referred to as *negative samples*. $R = R^+ \cup R^-$. In the example given in Fig. 3, blocks in the first row are positive samples for the concept *lion*. They contain a yellowish hair texture of a lion. The blocks in the second row are negative samples which represent other objects with "lion-like" colour and texture patterns.

If new classes are added to the set of semantic classes or the database is populated with new images containing new concepts, new representative groups for the incoming concept classes need to be generated.

The final processing step of the block-based analysis module calculates the centroid of a class in each of the considered feature spaces using the positive representative group R^+. A centroid $\bar{v}$ in feature space (F_j, d) can be calculated by finding the sample with the minimal sum of distances to all other positive samples in R^+. Let v_i and v_k be the feature vectors extracted from representative samples b_i and b_k , $i, k \in [1, |R^+|]$, in feature space (F_j, d). $d(v_i, v_k)$ is the distance between these two blocks. Then the centroid of the representative group in the feature space (F_j, d) can be defined as:

$$\bar{v} = \underset{i \in [1,|R^+|]}{\operatorname{argmin}} \left\{ \sum_{k \in [1,|R^+|], k \neq i} d(v_i, v_k) \right\}. \tag{2}$$

Taking $\bar{v}$ as an anchor, for a given block b_k in Γ, where $k \in [1, |\Gamma|]$, the distance in feature space F_j from b_k to the centroid $\bar{v}$ can be calculated using feature vector v_k of b_k:

$$\bar{d}_k = d(\bar{v}, v_k). \tag{3}$$

All the centroids of different feature spaces form a particular set of vectors $\bar{V} = \{\bar{v}_1, \bar{v}_2, ..., \bar{v}_n\}$. $\bar{V}$ is referred to as the *generalised centroid* of R, since $\bar{V}$ may be represented by multiple positive samples. The generalised centroid $\bar{V}$ is always calculated considering only positive samples. This is because the negative samples can locate anywhere in the feature space and calculating the generalised centroid of both positive and negative representative samples would become meaningless and adversely affect the subsequent classification process.

For the representative group R containing a total number of m representative samples, a generalised distance matrix $\bar{M}$ of size $m \times n$ can thus be constructed:

$$\bar{M} = \begin{matrix} \bar{d}_{1,1} & \bar{d}_{1,2} & ... & \bar{d}_{1,n} \\ \bar{d}_{2,1} & \bar{d}_{2,2} & ... & \bar{d}_{2,n} \\ \vdots & & \ddots & \vdots \\ \bar{d}_{m,1} & \bar{d}_{m,2} & ... & \bar{d}_{m,n} \end{matrix}, \tag{4}$$

in which the element $\bar{d}_{i,j}$ in row i and column j is the normalised distance from representative block b_i, $i = 1, 2, ..., m$, to the centroid vector $\bar{v}_j$ in feature space (F_j, d), $j = 1, 2, ..., n$.

Now, let us assume that a total number of r semantic classes are considered in an experiment. Based on the methods described in this section, for each class, a representative group R can first be selected, a generalised centroid $\bar{V}$ can then be calculated and a generalised distance matrix $\bar{M}$ can be built as in (4).

3. MULTI-FEATURE SPACE LEARNING BY MOO

The main goal of the MFL approach is to define a suitable multi-feature model for each semantic class. The core in this approach is a multi-objective optimisation strategy that is used for learning an optimal combination model through optimising multiple objective functions. Since several representative samples are required in order to display a representative visual pattern for a semantic class, the contribution of each single sample may conflict with others. Thus, we construct an objective function for each of the representative samples, and use the MOO strategy to find a solution that can achieve a common optimum for all these functions.

Compared to multi-kernel learning methods [Gehler and Nowozin 2009; Vedaldi et al. 2009] or late fusion methods of multiple classifiers [Zhang and Ye 2009], the proposed MFL does not require using kernels for classification or ranking of data. As described in Section 2, it uses only normalised similarities in each feature space which cost very little in computation compared to the use of kernels. In this section, we describe the novel approach to fusion model optimisation in the MFL module.

3.1 Objective functions for optimisation-based learning

As a result of the steps described in Section 2, for a semantic class, a generalised distance matrix $\bar{M}$ is derived. The next stage would be to build the objective functions from the generalised distance matrix. The objective functions will be used as the basis for performing optimisation in the subsequent stage.

In order to take into account the contribution from each single representative sample, we use each row in $\bar{M}$ to construct one objective function. Each objective function is formed as weighted linear combinations of feature-specific distances. The weighting factors α are used as the optimisation variables in the next stage. Considering $\bar{M}$ in (4), a number of m objective functions can be constructed:

$$D(A) = \left\{ \begin{array}{c} D_1 = \sum_{j=1}^{n} \alpha_j \bar{d}_{1,j}, \\ D_2 = \sum_{j=1}^{n} \alpha_j \bar{d}_{2,j}, \\ \vdots, \\ D_m = \sum_{j=1}^{n} \alpha_j \bar{d}_{n,j}. \end{array} \right\} \tag{5}$$

The MFL approach seeks to learn from the representative group, a suitable set of coefficients $A = \{\alpha_j : j = 1, 2, ..., n\}$ subject to the constraint: $\sum_{j=1}^{n} \alpha_j = 1$. Objective functions for the positive representative samples are denoted by D_i^+ and those for negative representative samples are denoted by D_i^-.

Learning a multi-feature fusion model is now transformed into a problem of finding a solution that optimises each of these objective functions in (5). The weighted linear combination model D, together with the targeted optimal weighting factors $\hat{A}$, are referred to as the multi-feature fusion model in this paper.

3.2 Optimising a multi-feature fusion model using MFL

Generally speaking, an optimum is usually defined as the maximum or minimum of some objectives. The optimal solution $\hat{A}$ should lead to the maximal or minimal value of the objective function(s) in $D(A)$ among all possible scalar combination of A that satisfy the constraint(s). In the proposed MFL approach, the optimum is regarded as the minimum of all positive objective functions D_i^+ and maximum of all negative objective functions D_i^-. The problem of learning a multi-feature model is now posed as to find a solution that optimises each of these, in some cases contradicting, objective functions. Observe that different representative blocks may display different visual characters but these differences need to be harmonised. In other words, it means that simultaneous optimisation of multiple objectives is required. This optimisation process can be achieved using the MOO strategy. While various algorithms have been developed based on the MOO strategy, in this work *Pareto Archived Evolution Strategy* (PAES) [Knowles and Corne 2000] is adopted as the MOO algorithm to optimise the fusion model for visual descriptors. In PAES, the optimisation process first generates a set of potential *Pareto optimal solutions* according to the algorithm 1.

Using one sentence to describe the rule in this algorithm for defining the Pareto optimum $\hat{A}$, it can be stated as:

> *There does not exist another* $A' \in \Phi$ *such that* $D_i(A') \leq D_i(\hat{A})$*, for all* $i \in \{1, 2, ..., m\}$*, where* Φ *is the whole set of possible solutions.*

This algorithm usually generates a set of potential *Pareto optimal solutions* $\Phi = \{A_1, A_2, ...\}$. Thus, a second step is required to decide which one of these solutions is the most suitable or most feasible one. The meaning of optimum in our specific task can be described as the fusion model in which all the vectors for positive representative blocks are closely gathered around the *generalised centroid* while the vectors for negative representative blocks are randomly scattered around the *generalised centroid*. Thus, a sensible selection criterion can be defined as to find the $\hat{A}$ that satisfies:

$$\hat{A} = \underset{\hat{A} \in \Phi}{\operatorname{argmin}} \frac{\sum_{i \in [1, |R^+|]} D_i(\hat{A})}{\sum_{k \in [1, |R^-|]} D_k(\hat{A})}. \tag{6}$$

At this point the MFL approach has achieved its goal in finding a most suitable multi-feature fusion model for a semantic class. Then in the next stage, the obtained fusion model $D(\hat{A})$ will be employed to build a multi-feature based image classifier.

ALGORITHM 1: PAES Algorithm

ran = generateInitialRandonSolution()*;
$archive$.add(ran);
main: mut = mutate(ran);
if *(ran.dominates(mut))* **then**
    mut = null; // discard mut
else if *(mut.dominates(ran))* **then**
    ran = mut; $archive$.add(mut);
else if *(mut.dominatedby($archive$))* **then**
    mut = null;
else
    $isNew$ = test(ran, mut, $archive$); // check whether mut is a new solution
    if *($isNew$)* **then**
        $archive$.add(mut);
end
if *(checkTerminationCriterion() == ture)* **then**
    return;
else
    goto main; // goto line 3
end
*Here, ran is a random data sample, mut is a mutant data sample, $archive$ is the archive of Pareto optimum.

3.3 Image classification in optimised multi-feature space

For a particular semantic class w, an optimal multi-feature fusion model $D(\hat{A})$ is obtained from the previous optimisation step. Using $D(\hat{A})$, the multi-feature distance of any block $b \in \Gamma$ considering class w, can be calculated by:

$$\hat{D}_{b,w} = D_{b,w}(\hat{A}) = \sum_{j=1}^{n} \hat{\alpha}_j \bar{d}_{j,b,w}, \tag{7}$$

where $\hat{\alpha}_j \in \hat{A}$. Among all blocks of an image g, the block b that has the smallest multi-feature distance to the generalised centroid of class w is defined as the *salient block* for class w in this image. In the classification stage, a similarity value $S_{b,w} = 1 - \hat{D}_{b,w}$, is used instead of the distance value that demonstrate the dissimilarities. For instance, salient blocks for the classes *tiger*, *vegetation* and *rock* are highlighted in Fig. 4 with solid borders. Here, the similarities of this image to the semantic classes *tiger*, *vegetation* and *rock* are 0.9438, 0.7675, 0.6927, respectively.

Images that have larger similarities than an empirically defined threshold are considered as relevant to the target semantic class. In this way, an initial semantic image classification can be achieved by the first two modules in the framework. In Section 5, some experiment results using this classifier are presented, comparing to several single feature based classification results. It shows that the proposed multi-feature based classification approach performs better than the best single feature-based classification.

4. SEMANTIC CONTEXT LEARNING AND INFERENCE

The performance of MFL module is considerably better compared to classifications based on single features. However, there may be more potential for improvements beyond the intrinsic limitation of relying on low-level visual features to describe high-level semantics. Indeed, low-level visual features are generally considered as more ambiguous and semantically impoverished compared with sound patterns, phonemes, words or dialogues. In the example given in Fig. 4, enlarged blocks with red solid

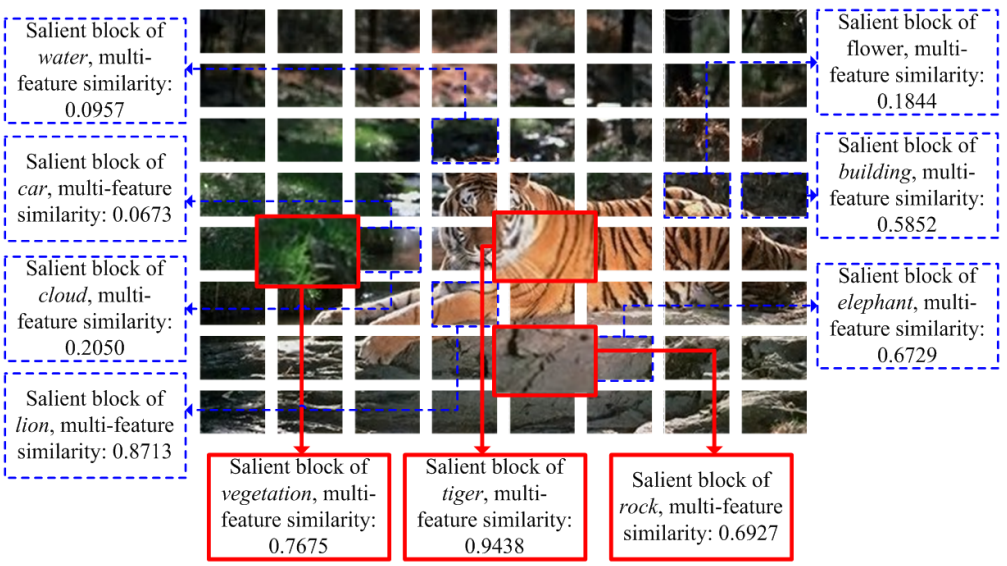

Fig. 4. An example of an image and its salient blocks for classes.

borders contain concepts that actually exist in this image. The similarities between these salient blocks and the classes are higher than their corresponding thresholds, so this image can be correctly labelled with keywords *tiger*, *vegetation* and *rock*. Blocks with blue dashed borders are salient blocks of some classes that do not match with this image. The obtained similarities of the salient blocks to *water*, *car*, *cloud*, *flower*, and *building* classes are lower than their thresholds. Thus, this image is correctly not-labelled with the corresponding keywords. The similarities of the salient blocks to *lion* and *elephant* classes are higher than their thresholds, so these two corresponding keywords are wrongly labelled on this image. These mistakes are unavoidably due to the limited discrimination power of visual feature based classification.

This observation motivated us to exploit the semantic context existing in images and thus improve the system performance. To illustrate the meaning of context, let us consider Fig. 4 again. Here, we use the similarity values to represent the probabilities of a block belonging to a particular class. Thus, the probability of an image belongs to class *tiger* should be increased when the same image also belongs to *vegetation* or *rock* classes, and should be decreased if the image belongs to *building* or *car* classes. We note that the semantic interrelations in images form a meaningful context, which can be potentially used to improve the image classification performance.

Semantic context learning is achieved by jointly considering the multi-feature similarities extracted from each block using the MFL module, and exploring the relationships between them. The proposed SCL module uses a Bayesian network to model these relationships. As shown in Fig. 5 (left), two steps are considered in the SCL module. In the *learning step*, a small set of images are used as a training set to obtain the *semantic context inference model*. In each one of these images, both visual and semantic evidences are taken into account as training data. Then, in the *inference step*, the learned context model is used to infer on the interrelationships between multiple classes within an input image, synthesis visual-semantic evidences and generate improved semantic labels.

A key feature of the Bayesian Network is its capability to handle incomplete information gracefully. Thus, despite of the fact that the multi-feature visual similarities extracted from the MFL module are not 100 percent correct, they can be used as a good evidence in building a meaningful Bayesian

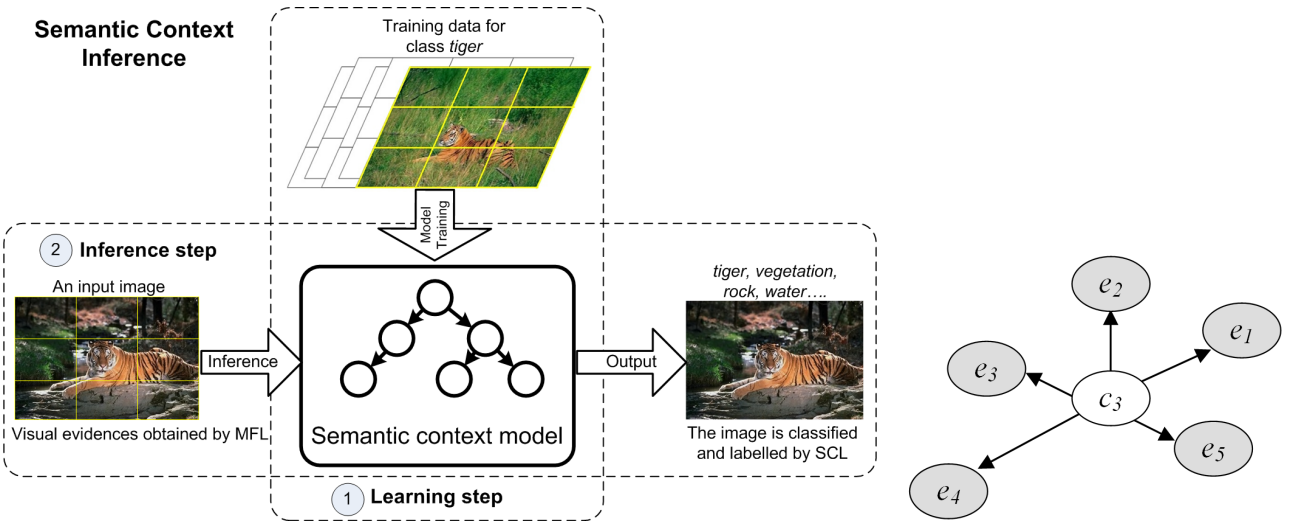

Fig. 5. Left: Illustration of the learning, inference and classification process based on the visual-semantic context model; Right: An example of a Bayesian network structure modelling a simple scenario of five classes.

network model. Improved classification is then achieved by solving a joint inference problem relying on the integration of existing clues to disambiguate the visual-semantic evidences in a suitable context defined through an experimental scenario.

Assume a total number of r semantic classes are considered in the experiment, denoted as $W = \{w_1, w_2, ..., w_r\}$. Multi-feature based classifications are performed on all the images for all these classes using the MFL module. The corresponding input for obtaining a context model for class w_i is a set of training images that come with two types of information. The first type of information is the ground-truth annotation on the class w_i. This data is referred to as the *semantic evidence*, denoted as c_i. The other type of the input information is the multi-feature similarities of the training images to all classes in W. This data is referred to as the *contextual visual evidences*, denoted as $E = \{e_i, i = 1, 2, ..., r\}$ where e_i is the contextual visual evidence for class w_i and E is the collection of this evidence for all classes.

Taking the simple example as in Fig. 4 again, if $W = \{cloud, rock, tiger, vegetation, water | r = 5\}$ and the class in concern is w_3 *tiger*, a typical naïve Bayesian network can be constructed as depicted in Fig. 5 (right). Since w_3 is the current concerned class, the node for semantic evidence c_3 is the parent node and the contextual visual evidence nodes $e_1, ..., e_5$ are the child nodes. For this example structure, the joint probability of c_3 and E is denoted as $P(E, c_3)$. According to the Bayes' rule, $P(E, c_3)$ can be re-written as:

$$P(E, c_3) = P(c_3) \cdot P(E|c_3) = P(c_3) \cdot P(e_1, e_2, ..., e_5|c_3). \tag{8}$$

Then if we apply the chain rule of probability theory on (8):

$$\begin{aligned} P(e_1, e_2, ..., e_5|c_3) &= P(e_1|c_3) \cdot P(e_2, e_3, e_4, e_5|e_1, c_3) \\ &= P(e_1|c_3) \cdot P(e_2|e_1, c_3) \cdot P(e_3, e_4, e_5|e_1, e_2, c_3) \\ &= P(e_1|c_3) \cdot P(e_2|e_1, c_3) \cdot P(e_3|e_1, e_2, c_3) \cdot P(e_4, e_5|e_1, e_2, e_3, c_3) \\ &= P(e_1|c_3) \cdot P(e_2|e_1, c_3) \cdot P(e_3|e_1, e_2, c_3) \cdot P(e_4|e_1, e_2, e_3, c_3) \\ & P(e_5|e_1, e_2, e_3, e_4, c_3). \end{aligned} \tag{9}$$

Intuitively, it is reasonable to assume that the multi-feature similarity of one image to class w_1 is an independent value to its multi-feature similarity to class w_2. Therefore, any two evidences in $\{e_1, e_2, ..., e_5\}$ should be independent to each other. Based on this assumption, the following terms can be re-written:

$$\begin{aligned} &P(e_2|e_1, c_3) = P(e_2|c_3) \\ &P(e_3|e_1, e_2, c_3) = P(e_3|c_3) \\ &P(e_4|e_1, e_2, e_3, c_3) = P(e_4|c_3) \\ &P(e_5|e_1, e_2, e_3, e_4, c_3) = P(e_5|c_3). \end{aligned} \tag{10}$$

Thus, the joint probability $P(E, c_3)$ can be re-written as:

$$\begin{aligned} P(E, c_3) &= P(c_3) \cdot P(E|c_3) = P(c_3) \cdot P(e_1, e_2, ..., e_5|c_3) \\ &= P(c_3) \cdot P(e_1|c_3) \cdot P(e_2|c_3) \cdots P(e_5|c_3) \\ &= P(c_3) \cdot \prod_{i=1}^{5} P(e_i|c_3), \end{aligned} \tag{11}$$

In a general case, assuming r semantic classes are available and the class of concern is c_k, $k = 1, 2, ..., r$, instead of (11) we have:

$$P(E, c_k) = P(c_k) \cdot \prod_{i=1}^{r} P(e_i|c_k). \tag{12}$$

There are many methods for learning both the structure and parameters of Bayesian networks from the given training data. Learning the network structure and parameters is in fact a search through the space of all possible links and parameters of the set of nodes n_i, $i = 1, 2, ..., t$, where t is the total number of nodes. In the proposed SCL method, given a set of pre-defined classes and the training data, a visual-semantic context model is constructed by applying the *K2* algorithm [Cooper and Herskovits 1992]. *K2* is a greedy search technique. It starts from an empty network with random initial settings and create a Bayesian network by iteratively adding a directed arc to a given node n_i from the parent node whose addition most increases the *K2* score of the resulting graph structure. The iterations terminates when no more possible additions could increase the *K2* score. The evaluation metric for calculating the *K2 score* of network structure is described as follows.

Given a database Δ, the *K2* algorithm searches for the Bayesian network structure G^* with maximal $Pr(G^*|\Delta)$, where $Pr(G|\Delta)$ is the probability of network structure G given the database Δ. For two Bayesian network structures G_1 and G_2, we have

$$\frac{Pr(G_1|\Delta)}{Pr(G_2|\Delta)} = \frac{Pr(G_1, \Delta)}{Pr(G_2, \Delta)} \tag{13}$$

Thus, the problem of calculating $Pr(G|\Delta)$ boils down to estimate $Pr(G, \Delta)$. Let $N(G) = \{n_i, i = 1, 2, ..., t\}$ be the set of nodes in Δ, where each node n_i has p_i possible values $\{\upsilon_{ik} : k = 1, 2, ..., p_i\}$. Besides, each node n_i has a set of parent nodes $\pi(n_i)$, with a total number of q_i instantiations. Define s_{ijk} to be the number of cases in Δ in which node n_i has the value υ_{ik} and ω_{ij} to be a unique instantiation of $\pi(n_i)$, $j \in [1, q_i]$ and $s_{ij} = \sum_{k=1}^{p_i} s_{ijk}$. The set of conditional probability distributions associated to a directed acyclic graph G is further denoted as GPr. Assuming that the cases occur independently and

the conditional probability density function $pdf(GPr|G)$ is uniform, then the ***K2 score*** can be computed as:

$$Pr(G,I) = Pr(G)\prod_{i=1}^{t}\prod_{j=1}^{q_i}\frac{(p_i-1)!}{(s_{ij}+p_i-1)!}\prod_{k=1}^{p_i}s_{ijk}!, \tag{14}$$

where $Pr(G)$ is the prior on the network structure that equals to a constant, thus it can be ignored. Therefore, the evaluation metric for computing the ***K2 score*** of a network structure G is given by:

$$K2score = \prod_{i=1}^{t}\prod_{j=1}^{q_i}\frac{(p_i-1)!}{(s_{ij}+p_i-1)!}\prod_{k=1}^{p_i}s_{ijk}!. \tag{15}$$

Assume the nodes are in a given order and n_i cannot be a parent of n_j if $i > j$. The algorithm starts a iterative process for each node n_i, including the following steps: 1) calculate the score for the case where n_i has no parents; 2) calculate the score for the case where n_i has a parent among all nodes that have smaller indices than i. If any of these are greater the case with no parents, select the node n_j which gives the maximum and add an arrow from n_i to n_j. This process is iterated by adding more parents and continue until no further nodes increasing the score can be found.

After the structure G^* of the Bayesian context model is learned, an inference process is carried out to measure joint probability distributions between the *contextual visual evidences* and the semantic class in each input image, as in (12). The multi-visual-feature similarities of images are replaced by the *posterior probabilities* inferred using the visual-semantic context model learned for a class. By applying the Bayes classifier using the context model of a class, images in the database are classified and labelled accordingly.

5. EXPERIMENTS

In this section, we present our experimental results and analyse the performance of the multi-feature fusion model and semantic context model obtained using MFL and SCL approaches. In Section 5.1, the results are evaluated based on a dataset called DB-COMB10000. In Section 5.2, we compare the performance of the proposed approaches with some other state-of-the-art approaches using a benchmarking dataset in the MIRFlickr database.

Without losing generality, seven visual primitives have been used to assess the performance of our experiments: Colour Layout Descriptor (CLD), Colour Structure Descriptor (CSD), Dominant Colour Descriptor (DCD), Edge Histogram Descriptor (EHD), Texture based on Gabor Filter (TGF), Grey-Level Co-Occurrence (GLC) and Colour Histograms in HSV Space (HSV). Observe that the first four primitives are MPEG-7 descriptors [Chang et al. 2001], while the other three are well established descriptors from the literature [Swain and Ballard 1991; Tuceryan and Jain 1993]. It is important to stress that the proposed approaches are not tailored to a given number of low-level descriptors; instead, any descriptor bank can be used.

Training sets of SCL module were randomly selected subset from the experimental image collections. In our experiments each training set contained five hundred images. The representative groups for the MFL module were also selected from the these training sets. Positive representative samples were collected directly based on existing ground-truth of the training sets. In case when negative samples were needed, they were selected by using the first a few false positive samples in the result of R^+ based MFL classification.

5.1 Experiments in DB-COMB10000 dataset

The proposed image classification approaches were first evaluated using an image collection named *DB-COMB10000*. DB-COMB10000 contains 10000 images selected from the *Corel* database and the *LabelMe* database [Russell et al. 2008]. The aim is to produce a dataset that is close to realistic image repositories in which images are complex and contains multiple semantic objects in both background and foreground. The context formed by multiple concept classes is an essential ground for inference and classification in the SCL method. For that reason this paper did not use popular object classification databases such as PASCAL VOC [Everingham et al. 2010], in which each image only contains a single object class. The DB-COMB10000 dataset has pictures depicting a variety of topics, taken in different time, places and conditions. As a consequence, only a small portion of images belong to some of the considered classes used in this paper, especially for the classes representing un-common objects such as *elephant* and *tiger*. Thus, the classification task in this dataset was very difficult and represents a realistic image classification scenario. Notice that the DB-COMB10000 dataset is fully annotated and of reasonable size comparing to some of the state-of-the-art image databases in the literature that also have full annotations.

The proposed approaches in this paper mainly target object based image classification. Thus, a set of object classes were selected here for experiments, covering most object-like semantic elements of image content. For DB-COMB10000, ten classes have been selected for testing based on subjective observation throughout the experimental dataset. These classes are: *building*, *car*, *cloud*, *elephant*, *flower*, *lion*, *rock*, *tiger*, *vegetation* and *water*.

The first experiment was performed based on the multi-feature fusion models generated by the MFL approach. To assess the performance of the multi-feature fusion models, the obtained results were compared to similar classifications but employing only single features. Thus for each class, eight experiments were performed, including one Multi-Feature Based Classification (MFBC) using the obtained class-specific multi-feature fusion models, and seven single feature based classifications (SFBCs) based on each one of the seven visual features. Fig.6 (left) shows the mean precision-recall curves across ten classes for the MFBC result and seven single feature based classification results. It can be observed in this figure, that the proposed multi-feature fusion models significantly improved the classification performance against any of the used single visual features.

A second experiment was performed to assess the performance of the SCL approach. Here, the multi-feature similarities were used as the *contextual visual evidences* for SCL based inference and classification. A Bayesian network was constructed to model this semantic context by carrying out the learning process as depicted in Fig. 5. In the inference process, a posterior probability can be calculated through this Bayesian model, indicating how likely each testing image belongs to a target class given both its multi-feature similarity and the visual-semantic context model. This round of experiments is referred to as the Context Inference Based Classification (CIBC). To show the classification performance of CIBC, the same mean precision-recall curve is depicted in Fig.6 (right). For comparison, the curve for MFBC is also depicted in the same figure. A clear observation from this picture is that the CIBC results have shown clear advantage over the MFBC results.

The results of MFBC and CIBC experiments show that the classes of an image can be effectively determined using the two proposed approaches: MFL and SCL. MFBC performed much better than any single feature based classification, which means that the MFL approach was able to build a suitable multi-feature fusion model for a particular semantic class. The obtained multi-feature fusion model synthesised the useful aspects from different features, and thus provided better discrimination power compared to the single features. Furthermore, the CIBC produced even better results compared to MFBC. This means that the SCL approach was able to build a useful visual-semantic context model

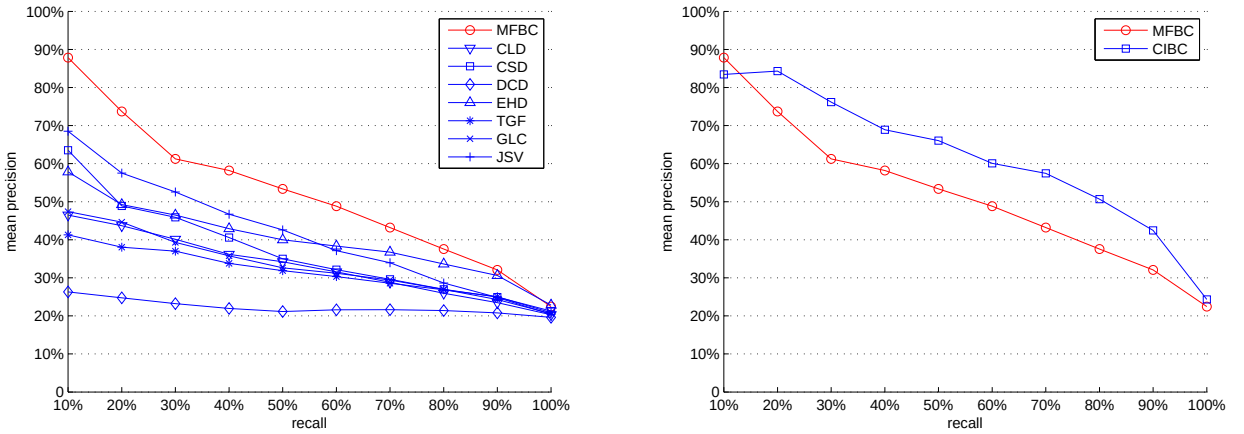

Fig. 6. Left: Mean precision-recall curves of ten classes using MFBC and SFBCs. The curve marked with o presents results of MFBC, while the other curves present results of SFBCs; Right: Mean precision-recall curves of ten classes using the two proposed approaches.

using the imperfect visual evidences, which in our case were the multi-feature similarities generated by the MFL approach. The inference process based on this model can boost the classification performance significantly.

5.2 Image classification in MIRFlickr25000 dataset

MIRFlickr25000 dataset [Huiskes and Lew 2008] is an image collection consists of 25000 images that were downloaded from the social photography site Flickr.com through its public API. These images are representative of a generic domain. MIRFlickr mainly targets concrete visual concepts in high-quality colour images, which suits the purpose of the object-based image classification system. This image collection has been used in image annotation tasks of the ImageCLEF benchmarking forum. In this task, usually both visual features and semantic features, i.e., user tags associated to the images, are used. However, in this paper, the proposed image classification system is purely visual based.

MIRFlickr provides full annotation of 24 potential labels and 14 regular (subjective) annotations. These concepts classes were considered in our experiments except two types of concepts. First, concepts about scenes were not considered in this paper. The task for the proposed framework is object-centered image classification. Scene classification is out of the scope of this paper. Second, concepts about human, i.e., female or baby, were not used in our experiments. Human detection and recognition has been a famous and well-studied research topic. Many specialised systems are available in the literature dedicated to this task. Our proposed system aims at general image classification, and is not suitable in detecting of human-related concepts. Therefore, the object classes being considered in our experiments include 14 labels: *sky*, *plant life*, *structure*, *clouds*, *sea*, *river*, *tree*, *flower*, *dog*, *bird*, *car*, *water*, *animals*, *lake*; and 8 regular annotations: *clouds(r)*, *sea(r)*, *river(r)*, *tree(r)*, *flower(r)*, *dog(r)*, *bird(r)*, *car(r)*.

The same seven visual features were used in the MFBC and CIBC experiments. Classifiers of MFBC and CIBC were trained for each of the classes and the mean precision-recall curves across the 22 classes are presented in Figure 7.

For benchmarking purpose, the resulting average precisions (APs) of each of the 22 classes were compared with the output of two visual feature based classification methods: a linear discriminant classifier (LDA) and a support vector machine classifier (SVM) with RBF kernel, provided in [Huiskes

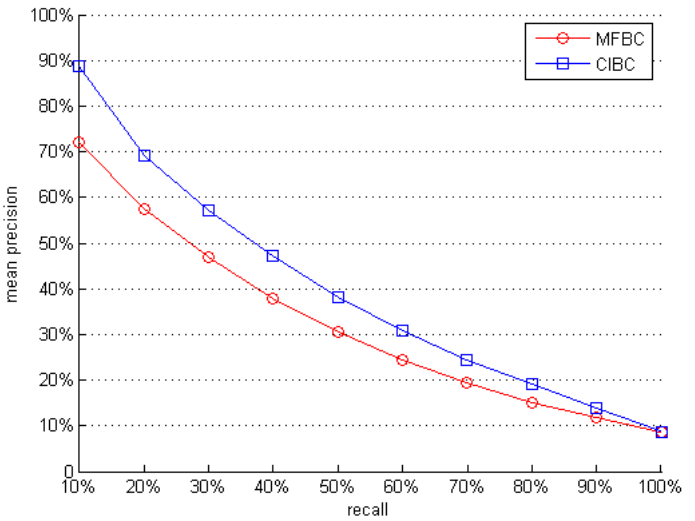

Fig. 7. Mean precision-recall curves of 14 classes using the two proposed approaches.

Table II. Comparison in terms of AP of MFBC, CIBC, LDA and SVMs.

%	*animals*	*bird*	*bird(r)*	*car*	*car(r)*	*clouds*	*clouds(r)*	*dog*
MFBC	27.8	12.1	11.0	26.4	23.3	45.0	25.9	17.3
CIBC	29.6	11.6	10.6	27.0	22.9	50.9	29.6	17.8
LDA	26.8	9.7	9.6	14.2	12.2	57.7	44.5	10.8
SVM	27.8	12.8	12.9	17.9	22.7	65.1	51.1	15.5
relevant rate	12.9	3.0	2.0	5.0	1.7	14.5	5.4	2.7

%	*dog(r)*	*flower*	*flower(r)*	*lake*	*plant life*	*river*	*river(r)*	*sea*
MFBC	17.6	28.8	23.0	13.9	62.9	14.1	11.0	21.5
CIBC	22.9	30.2	17.4	15.8	64.7	18.1	13.0	35.9
LDA	11.2	30.1	31.8	13.9	64.2	13.0	6.9	25.5
SVM	15.6	46.9	51.9	18.8	68.7	17.9	10.2	36.6
relevant rate	2.3	7.4	4.4	3.0	34.8	3.7	0.6	5.3

%	*sea(r)*	*sky*	*structure*	*tree*	*tree(r)*	*water*	Mean
MFBC	11.4	56.7	61.0	42.6	16.3	30.6	27.3
CIBC	14.4	63.9	62.5	43.5	13.3	36.6	29.6
LDA	9.1	74.9	61.5	43.4	14.4	35.7	28.2
SVM	12.6	77.5	62.6	51.4	20.5	44.8	34.6
relevant rate	0.8	31.0	40.4	18.3	2.7	13.1	9.8

et al. 2010]. In Table II, we present AP scores per concept class, for MFBC, CIBC, LDA and SVM. For reference, a rate of relevant images per concept in the ground-truth database is also included.

For 17 classes out of 22, CIBC approach yielded higher AP scores compared to MFBC approach. In average, CIBC had a higher mean AP score compared to LDA, while MFBC had slightly lower AP than LDA (less than 1%). SVM performed better than the other three approaches in most classes. It should

be noted that the better performance of SVM approach came at a much higher cost in calculation, compared with the two proposed approaches.

In terms of performance, in our experiment, there were mainly two real-time steps to be considered: 1) distance calculation using the multi-feature fusion model in MFL and 2) semantic context inference in SCL. A computer with Windows XP operating system and Intel Core-Duo E5200 CPU was used in all experiments. For one classification process in about 20 thousand images on concept *car*, step 1) used 13.37 seconds and step 2) used less than 1 second. As these on-line steps are based on straightforward arithmetic, the overall system has potential to be extended for application in large-scale image databases. Some of the off-line processes, such as feature extraction, may be time consuming but as they take place off-line, they would not affect the performance of the system.

6. CONCLUSIONS

In this paper, we have described an innovative framework for content based image classification supported by three closely collaborative approaches: multi-feature learning, semantic context learning and block-based representation scheme for semantic classes. The block-based representation scheme is an intuitive but effective way to extract object classes from complex images. The multi-feature learning process involves a multi-objective learning algorithm that is able to build optimised multi-feature fusion models for each semantic class. The aim is to harmonise the different interests of multiple representative block samples and generate a commonly agreed solution that satisfies the overall optimum. By using this approach, different aspects of the characteristic visual patterns in an object can be synthesised to generate a unified multi-feature space. The semantic context inference method has then been employed to enhance the classification performance by jointly exploring the context among object classes represented by blocks of an image. The visual-semantic context, which can be represented by the interrelationships among object classes, is modelled using a Bayesian network. This Bayesian model has been generated through learning process based on a search based algorithm *K2* in a fully automatic manner.

We have shown that, with extensive experiments, the proposed framework for image classification was able to effectively classify visual content that belongs to a class. We have also shown that the proposed approaches were able to handle a variety of problems in image classification without requiring extensive human expertise and efforts for constructing the Bayesian model. We first demonstrated the adequacy of the multi feature learning approach in semantic classification based on block regions. The multi-feature based classification employing the learned fusion model outperformed single visual feature based classifications, and other similar multi-feature based classification methods described in the literature. On top of that, we have shown that the proposed semantic context learning method can further boost the classification performance. The advantage of the semantic context learning method is that it exploits the very important, but usually ignored, visual and semantic context information to assist the classification. Thus, it can often improve the classification accuracy compared to the classifiers relying on purely visual features. The improvement introduced by the semantic context inference method has demonstrated the effectiveness of exploring the resourceful context information in addition to the isolated visual content and vocabularies.

ACKNOWLEDGMENTS

The research that lead to this paper was supported by the European Commission under contract FP7-287704 CUBRIK.

REFERENCES

AKSOY, S., KOPERSKI, K., TUSK, C., MARCHISIO, G., AND TILTON, J. 2005. Learning bayesian classifiers for scene classification

with a visual grammar. *Geoscience and Remote Sensing, IEEE Transactions on 43,* 3, 581–589.

ATHANASIADIS, T., MYLONAS, P., AVRITHIS, Y., AND KOLLIAS, S. 2007. Semantic image segmentation and object labeling. *Circuits and Systems for Video Technology, IEEE Transactions on 17,* 3, 298–312.

BERMAN, A. AND SHAPIRO, L. 1999. A flexible image database system for content-based retrieval. *Computer Vision and Image Understanding 75,* 1-2, 175–195.

CHANG, S., SIKORA, T., AND PURL, A. 2001. Overview of the MPEG-7 standard. *Circuits and Systems for Video Technology, IEEE Transactions on 11,* 6, 688–695.

CHUN, Y., KIM, N., AND JANG, I. 2008. Content-based image retrieval using multiresolution color and texture features. *Multimedia, IEEE Transactions on 10,* 6, 1073–1084.

COOPER, G. AND HERSKOVITS, E. 1992. A Bayesian method for the induction of probabilistic networks from data. *Machine learning 9,* 4, 309–347.

CSURKA, G., DANCE, C., FAN, L., WILLAMOWSKI, J., AND BRAY, C. 2004. Visual categorization with bags of keypoints. In *statistical learning in computer vision, Proceedings of the workshop on*. Vol. 1. 22.

DAGLI, C. AND HUANG, T. 2004. A framework for grid-based image retrieval. In *Pattern Recognition, Proceedings of the 17th International Conference on*. Vol. 2. IEEE, 1021–1024.

DATTA, R., JOSHI, D., LI, J., AND WANG, J. 2008. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys 40,* 2, 5.

DE JONG, F., WESTERVELD, T., AND DE VRIES, A. 2007. Multimedia search without visual analysis: the value of linguistic and contextual information. *Circuits and Systems for Video Technology, IEEE Transactions on 17,* 3, 365–371.

DESELAERS, T., KEYSERS, D., AND NEY, H. 2008. Features for image retrieval: an experimental comparison. *Information Retrieval 11,* 2, 77–107.

EVERINGHAM, M., VAN GOOL, L., WILLIAMS, C. K. I., WINN, J., AND ZISSERMAN, A. 2010. The pascal visual object classes challenge. *International Journal of Computer Vision 88,* 2, 303–338.

FAN, J., GAO, Y., LUO, H., AND JAIN, R. 2008. Mining multilevel image semantics via hierarchical classification. *Multimedia, IEEE Transactions on 10,* 2, 167–187.

GEHLER, P. AND NOWOZIN, S. 2009. On feature combination for multiclass object classification. In *Computer Vision, Proceedings of the IEEE 12th International Conference on*. 221 –228.

HUISKES, M. AND LEW, M. 2008. The MIR flickr retrieval evaluation. In *Multimedia information retrieval, Proceedings of the 1st ACM international conference on*. 39–43.

HUISKES, M. J., THOMEE, B., AND LEW, M. S. 2010. New trends and ideas in visual concept detection: the MIR flickr retrieval evaluation initiative. In *Multimedia information retrieval, Proceedings of the international conference on*. 527–536.

JAIN, A. AND VAILAYA, A. 1996. Image retrieval using color and shape. *Pattern recognition 29,* 8, 1233–1244.

JIANG, W., CHANG, S.-F., AND LOUI, A. 2007. Context-based concept fusion with boosted conditional random fields. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*. Vol. 1. I–949 –I–952.

JIANG, Y.-G., WANG, J., CHANG, S.-F., AND NGO, C.-W. 2009. Domain adaptive semantic diffusion for large scale context-based video annotation. In *Computer Vision, 2009 IEEE 12th International Conference on*. 1420 –1427.

JING, F., LI, M., ZHANG, H., AND ZHANG, B. 2004. Relevance feedback in region-based image retrieval. *Circuits and Systems for Video Technology, IEEE Transactions on 14,* 5, 672–681.

KHERFI, M. AND ZIOU, D. 2007. Image collection organization and its application to indexing, browsing, summarization, and semantic retrieval. *Multimedia, IEEE Transactions on 9,* 4, 893–900.

KNOWLES, J. AND CORNE, D. 2000. Approximating the nondominated front using the Pareto archived evolution strategy. *Evolutionary computation 8,* 2, 149–172.

KOSKELA, M., SMEATON, A., AND LAAKSONEN, J. 2007. Measuring concept similarities in multimedia ontologies: Analysis and evaluations. *Multimedia, IEEE Transactions on 9,* 5, 912–922.

LAVRENKO, V., FENG, S., AND MANMATHA, R. 2004. Statistical models for automatic video annotation and retrieval. In *Acoustics, Speech, and Signal Processing, Proceedings of the IEEE International Conference on*. Vol. 3. IEEE, 1044–1047.

LEVINSHTEIN, A., STERE, A., KUTULAKOS, K., FLEET, D., DICKINSON, S., AND SIDDIQI, K. 2009. Turbopixels: Fast superpixels using geometric flows. *IEEE Transactions on Pattern Analysis and Machine Intelligence 31,* 12, 2290–2297.

LI, F., DAI, Q., XU, W., AND ER, G. 2008. Multilabel neighborhood propagation for region-based image retrieval. *Multimedia, IEEE Transactions on 10,* 8, 1592–1604.

LI, J. AND WANG, J. 2003. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence 25,* 9, 1075–1088.

LOWE, D. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision 60,* 2, 91–110.

MIKOLAJCZYK, K. AND SCHMID, C. 2005. A performance evaluation of local descriptors. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 27,* 10, 1615 –1630.

NATSEV, A., RASTOGI, R., AND SHIM, K. 2004. Walrus: A similarity retrieval algorithm for image databases. *Knowledge and Data Engineering, IEEE Transactions on 16,* 3, 301–316.

QI, G.-J., HUA, X.-S., RUI, Y., TANG, J., MEI, T., AND ZHANG, H.-J. 2007. Correlative multi-label video annotation. In *Multimedia, Proceedings of the 15th international conference on*. 17–26.

RABINOVICH, A., VEDALDI, A., GALLEGUILLOS, C., WIEWIORA, E., AND BELONGIE, S. 2007. Objects in context. In *Computer Vision, Proceedings of the IEEE 11th International Conference on*. 1–8.

RASIWASIA, N. AND VASCONCELOS, N. 2009. Holistic context modeling using semantic co-occurrences. In *Computer Vision and Pattern Recognition, Proceedings of the IEEE Conference on*. 1889 –1895.

RUI, Y., HUANG, T., ORTEGA, M., AND MEHROTRA, S. 1998. Relevance feedback: A power tool for interactive content-based image retrieval. *Circuits and Systems for Video Technology, IEEE Transactions on 8,* 5, 644–655.

RUSSELL, B., TORRALBA, A., MURPHY, K., AND FREEMAN, W. 2008. Labelme: a database and web-based tool for image annotation. *International Journal of Computer Vision 77,* 1, 157–173.

STEUER, R. 1986. *Multiple criteria optimization. Theory, computation, and application*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics.

SWAIN, M. AND BALLARD, D. 1991. Color indexing. *International journal of computer vision 7,* 1, 11–32.

TANG, J., HUA, X.-S., WANG, M., GU, Z., QI, G.-J., AND WU, X. 2009a. Correlative linear neighborhood propagation for video annotation. *Systems, Man, and Cybernetics, IEEE Transactions on 39,* 2, 409–416.

TANG, J., YAN, S., HONG, R., QI, G.-J., AND CHUA, T.-S. 2009b. Inferring semantic concepts from community-contributed images and noisy tags. In *Multimedia, Proceedings of the 17th ACM international conference on*. MM '09. 223–232.

TORRES, R. D. S., FALCÃO, A. X., GONÇALVES, M. A., PAPA, J. A. P., ZHANG, B., FAN, W., AND FOX, E. A. 2009. A genetic programming framework for content-based image retrieval. *Pattern Recognition 42,* 2, 283 – 292.

TUCERYAN, M. AND JAIN, A. 1993. Texture analysis. *Handbook of pattern recognition and computer vision 276*.

VAILAYA, A., JAIN, A., AND ZHANG, H. 1998. On image classification: City images vs. landscapes. *Pattern Recognition 31,* 12, 1921–1935.

VEDALDI, A., GULSHAN, V., VARMA, M., AND ZISSERMAN, A. 2009. Multiple kernels for object detection. In *Computer Vision, 2009 IEEE 12th International Conference on*. 606 –613.

VOGEL, J. AND SCHIELE, B. 2007. Semantic modeling of natural scenes for content-based image retrieval. *International Journal of Computer Vision 72,* 2, 133–157.

WANG, J., LI, J., AND WIEDERHOLD, G. 2001. Simplicity: Semantics-sensitive integrated matching for picture libraries. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 23,* 9, 947–963.

ZHANG, J. AND YE, L. 2009. Content based image retrieval using unclean positive examples. *Image Processing, IEEE Transactions on 18,* 10, 2370–2375.

ZHANG, L., ZHANG, K., AND DONG, X. 2011. Online sparse learning utilizing multi-feature combination for image classification. In *Image Processing, Proceedings of the 18th IEEE International Conference on*. 197 –200.

ZHANG, Q. AND IZQUIERDO, E. 2007. Combining low-level features for semantic inference in image retrieval. *Journal on Advances in Signal Processing*, 12.

ZHANG, R. AND ZHANG, Z. 2007. Effective image retrieval based on hidden concept discovery in image database. *Image Processing, IEEE Transactions on 16,* 2, 562–572.

ZHU, X., WU, X., ELMAGARMID, A., FENG, Z., AND WU, L. 2005. Video data mining: Semantic indexing and event detection from the association perspective. *Knowledge and Data engineering, IEEE Transactions on 17,* 5, 665–677.