1. INTRODUCTION

In modern and continuously growing media databases, the inability of accessing accurate and desired content can be as limiting as the lack of content itself. Unlike textual information, which is human defined and precise in meaning, a picture, or audio-video content has a hidden component of creative reasoning of the human brain. This gives the content an overall shape and meaning far beyond capabilities of any language-based representation. In the last decades, the research in information retrieval has moved from strictly text based retrieval systems to multimedia content databases fusing information together. In case the content has been manually labelled then the problem of retrieval may appear easier. However, the adequacy of such a solution depends on human interaction, which is expensive, time consuming and therefore infeasible for many applications. Furthermore, such semantic based annotation is completely subjective and depends on semantic accuracy in describing the media. Though this last issue can be alleviated by constraining the annotation to words extracted from a predefined dictionary, taxonomy or advanced ontology, the challenge of automatic annotation and retrieval using semantic structures natural to humans remains critical.

If the aim is to retrieve audiovisual content using semantic structures, e.g., words or sentences, which are natural to humans, two profound challenges become evident: how to deal with the subjective interpretation of images by different users under different conditions; and how to link a semantic-based query with low-level metadata. The first problem originates in the fact that perceptual similarity is user and context dependent. The second challenge is a synonym of the semantic gap [1]. To provide a personalised, user tailored result, the machine needs to learn user preferences and to discover low-level pattern representation of these preferences. To bridge the semantic gap the machine needs to learn associations between low-level patterns and semantic concepts. In either case the need for learning and inference technology involving user’s input becomes evident [2]. The most natural way of getting user’s subjective information and preferences into the system is by using relevance feedback (RF) models. The idea underpinning this model is to integrate a “relevance feedback” loop into the system with the user at the centre and the machine learning from user’s feedback. By analysing feedback these models predict and learn user’s preferences in order to iteratively improve retrieval results. Semi-automatic adaptive learning strategies based on relevance feedback are aimed at learning relations between high-level semantic concepts used by humans to identify objects in an image and low-level descriptions of the same visual information.

The approach presented in this paper is based on several challenges as merging different low-level content features into meaningful descriptors with high semantic discrimination power; dealing with the subjective interpretation of images by different users under different conditions; and modelling semantic objects as mosaics made of irregular structures obtained through low-level and spatial clustering of building blocks. The mentioned aspects are connected through multi-feature space modelled in a structured way exploiting both low-level features and spatial relations. They are integrated into the learning approach through kernel based optimisation techniques.

The objective of this paper is twofold. Firstly, to describe relevant developments in RF-driven image annotation and retrieval, as given in section 2. Secondly, to introduce a model to infer semantic concepts for the annotation of objects in images as a user centred system; this is elaborated in section 3. Section 4 reports some results and the paper closes with conclusions in section 5.

2. STATE-OF-THE-ART IN USER DRIVEN SYSTEMS

Research in information retrieval whether based on textual description or low-level content is aiming at overcoming the drawbacks of machine limited behaviour and incorporating human understanding into the complex equation of machine responses. The learning process can generally rely on only positive examples [3-4], both positive and negative...
examples [5-6], multiple levels of relevance [7], or fuzzy membership functions [8].

A wide class of approaches is based on learning object structure via training on segmented regions. For instance, Ratan et al. in [9] identified an image as a bag of regions with every region represented by its own low-level features. Entire bags are labelled as relevant or irrelevant. In [10], Forsyth and Fleck, defined “body plans” for objects (e.g. horses) based on low-level features and geometrical relations among parts of a scene. Similarly, in [11] Bayesian inference is used on block based local image features for relevance feedback learning. Alternatively, in [12] authors use attribute relational graphs for contexts of objects and scenes. In [13] Fauqueur and Boujemaa do not consider regions individually but according to the class they belong to (positive or negative query category) and further on use query composition techniques.

On the other side, based on the learning method, user-driven relevance feedback models can be classified into: a) descriptive learning models, and b) discriminative learning models. A more detailed description of each class of model is presented in following sub-sections.  

2.1 Descriptive Learning Models

In early retrieval systems RF modules were a part of query by example (QBE) paradigm. One of the first retrieval systems using RF is MARS [7]. It introduced re-weighting techniques and query point movement (QPM), for exploitation of user’s feedback. The re-weighting scheme assumed independency of all features, and positive samples were modelled with single Gaussian distribution. The alternative approach, based on QPM, shifted the new query feature vector towards positive examples and away from negative examples. In [14], the PicHunter system used relevance feedback in a form of relative judgment and an extension of k-d trees to stochastic settings led to the proposed “stochastic-comparison search”. The first non-heuristic technique was introduced by Ishikawa et al. in [3]. The approach was solving an optimisation problem and considering correlation among features. A new optimal query was generated by minimising the total distances of positive examples from the new query, based on Mahalanobis distance.

While early probabilistic RF systems assumed that positive images follow a single Gaussian distribution later systems were based on more complex distributions such as Gaussian mixture models (GMMs). Qian et al. [15] used GMMs to represent the distribution of relevant images using both relevant and irrelevant examples as well as unlabelled data. Estimation of parameters for the mixture of Gaussians was based on positive relevance feedback and the assumption that positive examples may be grouped in the feature space and mutually separated by negative examples. In an effort to avoid assumptions for shape of a density model, Meilhac and Nastar in [16] used a non-parametric distribution for targeted samples based on Parzen window density. They defined the problem as a difference of densities for relevant and irrelevant samples. Every labelled image was declared a centre of a Gaussian with a high probability that the neighbouring elements are similar. The approach was incremental and user’s feedback was used to more precisely estimate the model.

2.2 Discriminative Learning Models

Discriminative learning models do not primarily concentrate on estimating the correct distribution of relevant and irrelevant data but rather on estimating the boundaries between classes. A straightforward way for dealing with non-linearity of visual feature space was introduced through discriminative approaches and kernel based algorithms [17]. Neural networks implicitly enable classification and ranking and thus can also effectively be used for relevance feedback, [8] and [18]. Statistical methods based on kernel SVMs and their good generalisation capabilities as well as active learning were considered for visual RF by Hong et al. [19], Tian et al. [20], and Tong and Chang [21].

One of the directions modern relevance feedback techniques are heading considers solving a two class (relevant/irrelevant) problem by use of discriminative analysis (DA). However conventional linear DA approach tries to cluster all irrelevant images into one class, whereas it is known that negative examples belong to many different classes. Wu et al. in [22], considered image retrieval to be a transductive learning problem, using both labelled and unlabelled data samples. A linear transformation based on the labelled data set that could be generalised onto the unlabelled dataset was found through multiple DA. The transformation maximises the ratio between inter-class and intra-class scattering, with all negative examples scattered among each other in the feature space. The main disadvantage is the assumptions that both positive and negative examples cluster into distinctive classes, that is, each negative example is treated as from a different class. However several negative examples can come from the same class and splitting them into different classes can mislead the resulting discriminating subspace. Hence, Zhou and Huang in [17] suggested a variation of the above approach in a form of Biased DA. They assumed that all the positive examples are clustered in one class while negative examples can come from an uncertain number of classes. In this case, a biased classification problem is defined with multiple classes but with the user being interested only in one class. The goal is to determine a mapping that would enable maximum clustering of positive examples moving them away from the negative examples. As a modification of the method above, an extension to non-linear space based on kernel methods, was also proposed. Kernel Biased DA is formulated through expressing the problem using inner product in the transformed space. However, kernel BDA is sensitive to imbalance between positive and negative samples and needs a considerable amount of training to perform well.

In the last couple of years there has been a huge impact of SVMs on classification, recognition and retrieval problems. The input data is mapped into higher-dimensional feature space using a non-linear transform. In the newly defined space linear discrimination boundary between classes can be easily found. In [19] and [20] SVMs are used for estimating the weight of relevant images in the covariance matrix of Mahalanobis distance. Similarly, Chen et al. in [23] used in
their RF approach a statistical learning machine, called one-class SVM. A kernel based one-class SVM acts as a density estimator for positive examples. Though the approach shows promising results negative examples when they exist, are a source of information completely ignored. Similarly, Guo et al. in [24] developed a constrained similarity measure for image retrieval, based on SVMs. Relevant images inside the boundary were ranked based on the Euclidean distances from it and images outside of the boundary were treated with less relevance.

Hence the choice for using SVMs for the learner in our user- driven approach is due to the following characteristics of SVMs: a) Good generalisation and learning properties even based on limited number of training examples. b) No prior assumption of the distribution for different classes is needed. c) There is a possibility of adapting kernels to the data by integration of prior knowledge, unlabelled samples and active learning.

3. USER- DRIVEN SYSTEM WITH ADAPTIVE KERNEL

We approach the problem from two sides. On the learning side we used SVMs within the relevance feedback module. The SVM optimisation approach is iteratively trained by the user in multi-feature space. This is a low-level feature space consisting of a number of colour, texture and edge features. The non-linear mapping that models the similarities of input patterns and defines the overall decision boundary different classes is defined through a kernel. It is expected that the complex nature of the patterns in images can be learned over time when using a combined distance function in multi-feature space. This combined distance function, as a linear combination of various distance functions is embedded into a Gaussian kernel leading to the adaptive convolution kernel (ACK):

\[ K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left( -\sum_{l=1}^{L} \omega_l \cdot d^{(l)}(\mathbf{x}^{(l)}_i, \mathbf{x}^{(l)}_j) \right) \]

(1)

where \( K(\mathbf{x}_i, \mathbf{x}_j) \) is a Gaussian convolution kernel over \( L \) joint feature vector space, with vectors \( \mathbf{x}_i \in \mathbb{X} \) where each individual feature is denoted as \( \mathbf{x}^{(l)}_i \). Here, \( d^{(l)}(\cdot, \cdot) \) denotes a distance function in a particular individual feature space and \( \omega_l \) are appropriate weights defined as ratio of variance over irrelevant and relevant samples in individual feature spaces, updated in every iteration. The convexity of the optimisation problem underpinning SVMs based on the ACK kernel and a set of MPEG-7 colour, texture and edge features, with different distance functions, has been proven in [25].

On the other, object modelling, side a method to organise individual image blocks into meaningful structured data conveying spatial information is presented. The aim is to keep the descriptor resolution at block level, local to objects rather than for whole images, while at the same time capturing higher-level semantic information. Firstly, elementary non-overlapping structures covering the whole image are generated, and then key-representative structures within each image are obtained using a modification of k-medoids clustering. This method combines distance function in multi-feature space with spatial proximity of initially generated clusters of blocks based on low-level similarity as described in [25]. The result is a set of representative structures based on low-level similarity and conveying spatial proximity, hence a rough object representation is obtained per each image, as in Fig 1 a) Each image is now represented as a set (a bag) of structures.

![Figure 1](image367x513.png to 477x633.png)

Figure 1 – a) Mosaics of structured elements after clustering using low-level similarity and spatial proximity of structured elements. b) Generating across- image clusters of structures (connected with blue lines). The next structure \( i \) or \( j \), from image labelled as 3, to be a member of the cluster is the one with smallest sum of distance \( \text{Sum} \) to all already chosen members of that cluster.

Additionally, in a relevance feedback scenario, user provided feedback through iterations in a form of relevant and irrelevant images is further exploited. Most similar structures across relevant and irrelevant images are captured, in a form of across- image clusters. In Fig 1 b) it is depicted how one of the two structures \( i \) or \( j \) is chosen to be a part of a particular across image cluster. The similarity is obtained as low-level similarity to all elements that are already members of that cluster. Most similar structure across images are clustered together, and further on ordered by decreasing levels of coherency. Where, coherency denotes the average distance among all cluster elements. Hence, in this way sets of most coherent clusters are obtained. And structures identified as common to irrelevant images are excluded from both the relevant and irrelevant set of structure representatives. The reasoning being, that if irrelevant images have very coherent clusters which are similar to clusters across-relevant images, then all the structures that are members of these clusters need to be removed from representative sets (e.g. coherent back-
ground across these sets). The similarity between two across-image clusters is defined as low-level similarity between their medoids. Hence, the obtained set of representative structures for relevant and irrelevant images is further refined, removing background noise by capturing similarity across images.

The overall complexity of obtaining the refined sets of structures for each image is very small due to scars feedback information provided by the user in every iteration. Therefore the overall procedure is included in the SVM learning approach as an integral part of its kernel. Given two images and their decomposition into the above mentioned sets of representative structures with variable cardinality, the kernel to measure similarity is defined as the sum of local kernels, the ACK kernels, in multi feature space. It is estimated using all possible combinations of structure pairs between the two images.

4. EXPERIMENTAL RESULTS

The proposed set kernel with local ACK kernels operates on representative image blocks and can be denoted as SET(ACK). This kernel was experimentally evaluated with three other kernels operating in multi-feature space over whole images. These are: Gaussian radial basis function kernel (RBFK), Laplace kernel (LK), and the ACK kernel. These kernels were incorporated into the relevance feedback module with precision and recall as performance measures. In order to show improvement of performance for various kernels in the relevance feedback scenario, the ratio between retrieved images and the size of the relevant class, was set to one. In each iteration ten additional images are labelled with maximum 40% of the class being used as a training set. Two different databases were used in the experiments reported in this paper. The first database is a subset from the Corel stock and it consists of seven concepts. These are: buildings, clouds, cars, elephants, grass, lions and tigers. The second dataset is composed of 25 distinct classes from the COIL database [26].

In Figs. 2 and 3 results from the LK, RBFK, ACK, and SET(ACK) kernel are presented. For most categories there is a huge variance in low-level features when performing retrieval on whole images. Even though the results were averaged over a number of independent runs, the mentioned variance leads to inconstancy that inspired development of set kernels on structured feature spaces. The SET(ACK) kernel incorporating spatial information outperforms classical RBFK and LK as well as the ACK kernel on whole images. As it can be noticed, the ACK kernel from (1) when applied over whole images has higher precision then classical kernels. It can also be noticed that in case of SET(ACK) kernels, contrasting to performance of kernels on whole images, there is no considerable decrease in performance as the number of training samples increase, e.g., after the second iteration.

This is due to the fact that noisy elements from the background are significantly reduced, thus, they have a negligible influence on the performance. These results justify the idea of using structured descriptors with low-level similarity and also incorporating spatial information about building blocks.

5. CONCLUSIONS

In this article relevant developments in user-driven image annotation and retrieval are presented. A new set kernel, coupled with clustering approaches, defined in spatially structured space has also been proposed. The proposed scheme was motivated by the observation that labelling complete images as relevant to a given concept introduces a lot of noise due to the variety of non-relevant objects in complex scenes. The kernel encloses both multi-feature space and spatial information about localised image structures, enabling a higher transparency between low-level image features and semantic concepts. The approach organises individual image blocks into meaningful structured data conveying spatial information. In each iteration of the RF session, the proposed kernel captures most similar structures across images labelled...
as relevant and across images labelled as irrelevant. If clusters across irrelevant images are similar to some cluster for relevant images, then appropriate structures are excluded from representative sets for each image. The proposed scheme has been tested with approaches operating on whole images and an evident advantage of the designed method has been confirmed.

REFERENCES


©2007 EURASIP