RELATING VISUAL AND SEMANTIC IMAGE DESCRIPTORS

J. STAUDER AND J. SIROT

Thomson, Corporate Research, Rennes, France.
E-mail: {jurgen.stauder, joel.sirot}@thomson.net

H. LE BORGNE, E. COOKE AND N.E. O'CONNOR

Centre for Digital Video Processing, Dublin City University, Dublin 9, Ireland
E-mail: {hlborgne, ej.cooke, oconnorn }@eeng.dcu.ie

This paper addresses the automatic analysis of visual content and extraction of metadata beyond pure visual descriptors. Two approaches are described: Automatic Image Annotation (AIA) and Confidence Clustering (CC). AIA attempts to automatically classify images based on two binary classifiers and is designed for the consumer electronics domain. Contrastingly, the CC approach does not attempt to assign a unique label to images but rather to organise the database based on concepts.

1. Introduction

Semantic metadata for multimedia and semantic-level search and retrieval functions are crucial for efficient and easy to use multimedia services. However, there is often no default semantic metadata available to a system user. When visual content is considered, metadata can be associated to four semantic levels [1,2]: (level 1) visual *e.g.* round, (level 2a) generic objective *e.g.* landscape, (level 2b) instantiated objective *e.g.* the Rockies, (level 3): abstract, emotional, *e.g.* important. In first generation image retrieval tools, *e.g.* QBIC, Photobook, Blobworld, and professional multimedia asset management tools, *e.g.* iBase, iPhoto, PictureIt, the problem is bridging the semantic gap between visual features (level 1) and objective image content (level 2).

The aim of the European IST project aceMedia[‡] is to support users in intuitively accessing, managing, communicating, and enjoying collections of content. In this context, this paper addresses the automatic analysis of visual content and extraction of metadata superior to pure visual descriptors. Two approaches are described, both targeting level 2a type metadata. The first approach is Automatic Image Annotation (AIA) for applications of minimal user interaction in the domain of consumer electronics (CE). Problematic requirements of CE applications are the high degree of automatic processing,

[‡] This research was funded by the European IST FP6-001765 project AceMedia - http://www. acemedia.org

semantic meaningful results and low visibility and effect of processing errors. Image annotation by supervised classification is a technology that recently was shown to be capable of addressing semantic scene types [3, 4, 5]. Using well-chosen databases for learning, semantic-visual concepts such as indoor, outdoor, city, and landscape can be detected automatically. We present a complete system consisting of visual descriptor extraction, supervised learning and automatic detection of semantic visual concepts in images.

100% natural 70% natural Ambiguous 70% artificial 100% artificial

Figure 1*: Classification via the confidence-clustering KNN classifier*

The second approach focuses on the organisation of an image database in such a way that it reflects the semantics of the images. Useful for CE as well as for professional applications, this Confidence Clustering (CC) approach does not attempt to assign a unique label to images but rather to organise the database based on concepts. This idea was introduced by Oliva *et al* [12] who indicated its usefulness for the classification of "natural scenes". For instance, in a city/landscape paradigm images such as those illustrated in Figure 1 would be assigned confidence values as oppose to specific labels. For this, we consider standard and widely used image descriptors, similar to those proposed in the MPEG-7 standard (see Section 2.1) and investigate the influence of different distances to compute the similarities. The main contribution of this work is the use of a "confidence-clustering KNN classifier" that allows an efficient fusion of descriptors and demonstrates the potential of our approach. The paper is structured as follows. First, Section 2 reviews some basics and state of the art. In Section 3, the method for automatic image annotation is described. In Section 4, confidence-clustering approach for image organization is presented. Finally, Section 5 gives first preliminary experimental results and Section 6 a conclusion.

2. Review: Bridging the Semantic Gap

2.1. *Visual Descriptors*

The goal of the ISO/IEC MPEG-7 standard [9] is to allow interoperable searching, indexing, filtering, and browsing of audio-visual (AV) content. In order to describe this AV content the MPEG-7 standard specifies a set of descriptors. A descriptor defines the syntax and the semantics of an elementary

AV feature, which may be low-level, *e.g.* colour, or high-level, *e.g.* author. The aceToolbox developed within aceMedia is based on the architecture of the MPEG-7 experimentation Model [10] and uses a subset of these low-level visual descriptors of colour and texture in order to identify and categorise images. A brief overview of each descriptor is provided below. For more detail the reader is referred to [11].

Colour Structure Descriptor (CSD): is designed to express local colour features. This is achieved by creating a colour histogram based on the number of times each colour occurs within an 8x8 block as it scans across the image. Scalable Colour Descriptor (SCD): measures colour distribution over an entire image. It is defined in the HSV colour space and produces a colour histogram encoded by a Haar transform, allowing for a scalable representation. Colour Layout Descriptor (CLD): is designed to capture the spatial distribution of colour in an image. The feature extraction process consists of two parts; grid based representative colour selection and DCT transform with quantization. Edge Histogram Descriptor (EHD): is designed to capture the spatial distribution of edges by dividing the image into 16 non-overlapping blocks and then calculating 5 edge directions in each block. The output is a 5 bins histogram for each block. Homogenous Texture Descriptor (HDT): Describes directionality, coarseness, and regularity of patterns in images. It is computed by first filtering the image with a bank of orientation and scale sensitive filters, and then computing the mean and standard deviation of the filtered outputs in the frequency domain.

2.2. ***Semantic Image Classification***

Both the AIA and the CC approach for image organization are based on statistical classification. Classification approaches have been shown to be able to bridge the gap between available visual features and required semantic categories. Firgensohn *et al.* [6] propose a classification scheme for semantic categories in video images such as audience, close-ups, and graphics. Statistics from DCT transform coefficients are used to train uncorrelated Gaussians resulting in a recall rate of 84%. For photo collections, Huang et al. [7] classified images into classes such as sunset, flowers, clouds, and motorcars. These two approaches are characterized in that the variety of content and number of classes are limited.

To detect semantics in photo collections of larger variety, Mojsilovic et al. [8] and Vailaya *et al.* [5] conducted usage studies and found hierarchical categories of type indoor/outdoor, city/landscape, and mountain/water. These classes represent the semantic level 2a content. The authors propose further approaches for automatically classifying images into these categories. First, visual features such as colour, shape, texture, contours, and sub-band energies are calculated. Then, categories are learned from a learning set of images, *e.g.* Vailaya employed cluster analysis based on vector quantization. Semantic image classification has been applied successfully in the literature to the following four semantic visual concepts: “indoor”, “outdoor”, “city”, and “landscape”. A global colour histogram is employed by Yiu [3] and Szummer and Picard [4]. Being faster than local colour descriptors, the global colour histogram is also invariant against image orientation. The usefulness of texture features for indoor/outdoor classification has been demonstrated for dominating orientations [3], DCT [4], contours [5], and auto-regressive models [4,5].

3. Automatic Image Annotation

3.1. *Overview*

We developed two binary classifiers: indoor/outdoor and city/landscape. For each classification problem, an appropriate (small sized) observation has to be formed from the available visual descriptors. Under presence of errors and noise in visual descriptor extraction, descriptor parts with low information should be excluded. A further argument to limit the size of observation is the curse of dimensionality, which relates to the difficulties of density estimation (and all it implies for classification) at dimensions from 20 on. We choose a subset of coefficients by experiments and manual inspection.

3.2. *Indoor/outdoor detection*

For indoor/outdoor detection, local colour descriptors can easily detect clues like blue sky, green vegetation or red tinted indoor scenes [5]. Earlier experiments [2] showed that local colour descriptors are able to capture the localization of colour and separate indoor and outdoor clusters. For our implementation, we chose the global colour histogram. Contrary to MPEG SCD, it is not scalable and uses RGB color space. This global descriptor is much faster, rotation invariant, and works on partial images. For indoor/outdoor classification, we use additionally a global texture descriptor that can detect typical high frequencies in outdoor images and typical vertical structures in indoor scenes. The texture descriptor consists of global energies calculated in four sub bands using a 16-tap linear phase QMF filter. Finally the observation vector contains 24 coefficients.

3.3. *City/landscape detection*

For city/landscape detection we use only a contour descriptor similar to EHD. It is capable of distinguishing between dominating horizontal and vertical, more or less long contours in city images and more or less short contours of any direction in landscape images. The contour descriptor is a histogram of contour directions. First, edge pixels are detected by thresholding luminance gradient amplitudes calculated by a Deriche filter. Then, edges are skeletized and concatenated by local search for strongest gradients. Connected line segments are polygonized, and finally, their direction is calculated. To cope with short line segments and different image resolutions, a short 12-bin histogram is extracted. This gives an observation vector of 12 dimensions.

4. Confidence-Clustering for Image Organisation

4.1. *Overview*

The method presented here aims to organise a whole database according to semantic concepts. In a city/landscape paradigm for instance, our method aims to determine in relative terms whether an image can be considered a city/landscape as well as what degree of confidence is associated with this classification. The decision rule is based on a K-nearest-neighbours (KNN) classifier. A KNN classifier defines nonlinear boundaries by giving the same label to a query as the majority of its K nearest neighbours in the feature space. For a two-class problem the use of an odd K ensures a given query is attributed to one of the classes. If K has an even value there is the risk that neither of the classes can be attributed to the query. Since no decision can be taken for these images we introduce a third "undetermined" class to which they are assigned. This "non-decision" is a first step toward a full system that will be better able to organise the database, by using smooth decision rules.

4.2. *Fusion of several descriptors*

The fusion of several image descriptors is a crucial point for retrieval systems. A classic approach is to normalise the distances between images according to the different descriptors, then add these distances to obtain a unique distance for each pair (additive fusion). A KNN classifier can then be used to obtain the decision for each image. A drawback of this additive fusion is that it computes the average of the distances (by summing them) and therefore risks neglecting the good performances of a given descriptor because of the poor performances of another. We chose to take the decision about the class of an image directly in the feature space. For this, the nearest neighbours are independently identified for every descriptor, and are then summed. Then we use the same KNN classifier (with an undetermined class) as described in Section 4.1.

5. Results

AIA Approach: For each of the classification problems indoor/outdoor and city/landscape we set up four distinct databases, two for each category, labelled with "1" (learning) and "2" (verification). The following table describes the databases and their size.

Database	**Size**	**Database**	**Size**	**Database**	**Size**	**Database**	**Size**
Indoor 1	500	Outdoor 1	500	City 1	437	Landscape 1	626
Indoor 2	402	Outdoor 2	400	City 2	497	Landscape 2	643

The learning databases are defined in a strict sense using images that belong semantically and visually clearly to their respective class. High quality as well as low quality photos were taken into account. The criteria used for choosing images for the verification database is their semantic information as opposed to their visual content. Classification was optimised by manual inspection of images and iterative processing. Criteria for optimisation are the recall rate and precision of recall. Considering the application of image annotation (contrary to the CC approach) high precision was preferred with respect to recall rate. When learning with a Support Vector Machine from databases labelled "1" and predicting for databases labelled "2", we obtained a recall of 85% at 90% precision for "indoor" and a recall of 83% at 91% precision for "city". The methods of Yiu [3], Szummer/Picard [4] and Vailaya et al. [5] state a recall rate of 90%, 90%, and 93%, respectively. A comparison is approximate, since these authors used different databases and did not indicate the precision of their results. Figure 2 and Figure 3 show examples for correct and wrong classification. Notably images with red tint, snow, many people or building close-ups are sometimes classified as indoor. Colour descriptors seem to be more important for indoor/outdoor classification than texture descriptors.

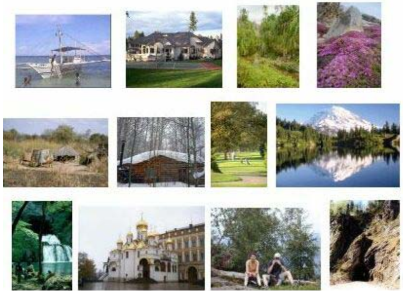


Figure 2: Outdoors images correctly classified as "outdoor"

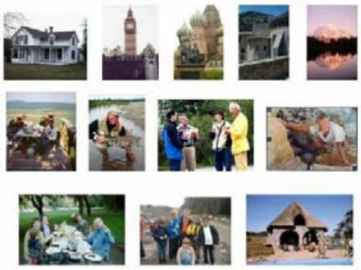
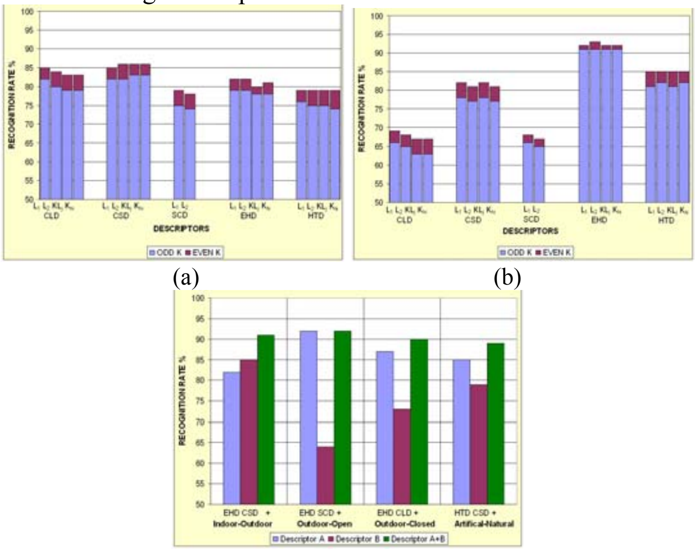


Figure 3: Outdoors images classified as "indoor"

CC Approach: Applying the protocol described in Section 4, we conducted two series of experiments using a database of 1900 images. This corpus is divided into 540 images of "outdoor cities", 445 images of "indoor rooms", and 431 landscapes with a large field-of-view (termed "open" landscapes). The first series of experiments deals with the individual performances of descriptors when we change the metrics used to compute the similarities between images. We considered four distances: L_1 and L_2 derived from the L_p Minkowsky distances, the Jeffrey divergence (symmetrical Kullback-Leibler divergence) and finally the $\chi 2$ statistics. The latter two distances are designed to be applied on distributions i.e. functions with null or positive values and hence, they cannot be used to compute the dissimilarities between images represented by the SCD. This initial set of results show that on one hand the performances are quite robust to the change of distance (Figure 4(a)(b)). On the other hand, it makes explicit the difference from one descriptor to another in the different experiments. In case (b) landscapes are classified against outdoor cities. Herein, the EHD outperforms any colour descriptor by more than 10%. While, case (a) shows a classification between "indoor scenes" and "outdoor cities" and the colour descriptors have quite better performances than the texture or edge descriptors.



(a)

(b)

(c)

Figure 4: Classification results: (a) Indoors – Outdoors, (b) Cities – Landscapes (open), (c) Improvements in classification via descriptor combination

In all cases we have also presented the effect of the "undetermined class" (in red in Figure 4(a)(b)) that systematically improves the results. It may seem obvious since this process allows us to get rid of all the "ambiguous" images and thus such an improvement may seem "artificial". In fact, it underlines the reason for defining an organisation paradigm rather than a strict classification when one wants to report semantic information about images. This is particularly true for images such as "holiday pictures" that can cover a large range of topics, and thus can belong to different "semantic axes". The second series of experiments, illustrated in Figure 4(c), addresses the combination of a colour and texture descriptor (edge histogram or homogeneous texture). In this case, represented by the third bar in each experiment, the results are significantly improved in comparison with a single-feature-based classification, first and second bar.

6. Conclusion

This paper presents two approaches for determining the so-called “generic objective level” in the semantic metadata annotation for digital libraries. The first approach is automatic annotation of images by detection of semantic concepts while the confidence-clustering method aims at organising image databases according to “semantic axes”. Both methods exhibit good performances in several classification tasks when compared to previous work in this field.

References

1. J. Eakins, M. Graham. "Content-based image retrieval", Information Security Forum, Newcastle, Oct. 1999.
2. J. Stauder, G. Gouzien, B. Chupeau, J.R. Vigouroux, E. Kijak: "Semantic image browsing using hidden categories and confidence values", Electronic Imaging 2003, Santa Clara, USA, January 20-24, 2003.
3. E.C. Yiu: "Image classification using color cues and texture orientation", Master thesis, MIT, 1996.
4. M. Szummer, R. Picard: "Indoor-outdoor image classification", Proc. IEEE Workshop on Content-based Access to Image and Video Databases, Bombay, India, January 1998.
5. A. Vailaya, M. Figueiredo, A. Jain, H.-J. Zhang: "Content-based hierarchical classification of vacation images", Proc. IEEE ICMCS, Italy, June 1999.
6. A. Girgensohn, J. Foote: "Video classification using transform coefficients", ICASSP-99, vol. 6, pp. 3045-3048.
7. J. Huang, S.R. Kumar, R. Zabih: "An automatic hierarchical image classification scheme", ACM Conference on Multimedia, Bristol, England, September 1998.
8. A. Mojsilovic, B. Rogowitz: "Capturing image semantics with low-level descriptors", Proc. IEEE ICIP, Vol. I, pp. 18-21, October 2001.
9. B.S. Manjunath, P. Salembier, and T. Sikora: "Introduction to MPEG-7: multimedia content description standard", New York, Wiley, 2001.
10. http://www.lis.ei.tum.de/research/bv/topics/mmdb/mpeg7.html
11. B.S. Manjunath, J.-R.Ohm, V.V. Vasudevan, and A. Yamada: "MPEG-7 color and texture descriptors", IEEE Trans. Circuits Syst. Video Technol., vol 11, pp. 703-715, June 2001.
12. A. Oliva, A. Torralba, A. Guérin-Dugué, J. Hérault: "Global semantic classification of scenes using power spectrum templates". Proc. Challenge of Image Retrieval, Elec. Work in Computing series, Springer-Verlag, Newcastle, 1999.