Detecting Semantic Concepts In Digital Photographs: Low-level Features Vs. Non-Homogeneous Data Fusion

by

Jovanka Malobabić, B.Eng

Supervisor: Dr. Noel Murphy

Thesis submitted for the degree of

Master of Engineering

Centre for Digital Video Processing
and
School of Electronic Engineering
Dublin City University

July 2007

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Master of Engineering is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed : _____

Jovanka Malobabić

ID No.: 50955213

Date: _____

Abstract

Semantic concepts, such as faces, buildings, and other real world objects, are the most preferred instrument that humans use to navigate through and retrieve visual content from large multimedia databases. Semantic annotation of visual content in large collections is therefore essential if ease of access and use is to be ensured. Classification of images into broad categories such as indoor/outdoor, building/non-building, urban/landscape, people/no-people, etc., allows us to obtain the semantic labels without the full knowledge of all objects in the scene.

Inferring the presence of high-level semantic concepts from low-level visual features is a research topic that has been attracting a significant amount of interest lately. However, the power of low-level visual features alone has been shown to be limited when faced with the task of semantic scene classification in heterogeneous, unconstrained, broad-topic image collections. Multi-modal fusion or combination of information from different modalities has been identified as one possible way of overcoming the limitations of single-mode approaches. In the field of digital photography, the incorporation of readily available camera metadata, i.e. information about the image capture conditions stored in the EXIF header of each image, along with the GPS information, offers a way to move towards a better understanding of the imaged scene.

In this thesis we focus on detection of semantic concepts such as artificial text in video and large buildings in digital photographs, and examine how fusion of low-level visual features with selected camera metadata, using a Support Vector Machine as an integration device, affects the performance of the building detector in a genuine personal photo collection. We implemented two approaches to detection of buildings that combine content-based and the context-based information, and an approach to *indoor/outdoor* classification based exclusively on camera metadata. An outdoor detection rate of 85.6% was obtained using camera metadata only. The first approach to building detection, based on simple edge orientation-based features extracted at three different scales, has been tested on a dataset of 1720 outdoor images, with a classification accuracy of 88.22%. The second approach integrates the edge orientation-based features with the camera metadata-based features, both at the feature and at the decision level. The fusion approaches have been evaluated using an unconstrained dataset of 8000 genuine consumer photographs. The experiments demonstrate that the fusion approaches outperform the visual features-only approach by of 2-3% on average regardless of the operating point chosen, while all the performance measures are approximately 4% below the upper limit of performance. The early fusion approach consistently improves all performance measures.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter outlines the research objective addressed and provides a context to the thesis. In the introductory part, it gives an overview of the relevant concepts in the wider research area, which includes content-based image indexing and retrieval, as well as image and scene understanding. Further on, we present a short comparison between consumer and corporate photo collections, and highlight the differences between the two in terms of potentially exploitable characteristics of images in each, as well as challenges they each present to research in the field of visual content analysis in general, and for the task of semantic classification in particular. Next, we briefly discuss the scope for improved semantic annotation of digital photographs. Finally, an outline of the thesis and the list of associated publications are presented.

1.1 General objective

The last decade has seen an enormous surge in the number of digital images being generated. The arrival of digital cameras and, more recently, camera phones in such a ubiquitous manner has brought yet another challenge – how to organise and manage large collections of digital photographs - how to facilitate a fast, yet user-friendly and easy retrieval of and access to a desired photo in a collection of several thousands photographs?

Is this an indoor or outdoor scene? What objects are present in the scene? Is there a building or some other human-made structure in the scene? Is there any artificial text included? Which

part of the world was the photo captured in? These are the sorts of questions that must be addressed when aspiring to annotate images in a fashion that facilitates user-friendly retrieval from a large collection of images. Users prefer expressing their needs in natural language, and rather that looking for "images with large objects that exhibit strong edges in horizontal and vertical directions, in muted colours", they prefer to look for "images depicting large building(s) in Copenhagen". To facilitate this type of query, which also emphasises the close bond between the location information and the image semantics [85], the images need to be annotated with semantic concepts rather than low-level descriptors. As manual annotation of images with linguistic terms (visual concepts) is time-consuming and expensive, and inevitably prone to subjectivity in perception, the need arises for automated annotation with semantic concepts as a viable alternative. Apart from semantic annotation, automatic understanding of image content is a key requirement to a plethora of applications in content-specific/content-sensitive image enhancement, analysis, organisation, etc.

In summary, the general objectives of this work are (i) to extract knowledge from digital visual content for the purpose of automated annotation of large personal photo collections with semantic concepts, and (ii) to examine the impact of multimodal data fusion of content-based and the context-based features on semantic concept detection performance. In this work, we focus on the task of detecting large building objects in still images, and the impact of integration of content-based evidence with the contextual evidence on the performance of the building detector. Here, the task of large building detection is viewed as an image classification problem.

1.2 Content-based indexing and retrieval of visual content

Content-based indexing and retrieval (CBIR) of visual content is a research area dealing with indexing and retrieval of still images and video, which are indexed by their own visual content, using low-level features such as shape, colour, texture, etc. Unlike text-based retrieval which relies on manual annotations (or the surrounding text, captions, etc.), the indexing in CBIR is performed automatically. Different visual features and their combinations may be used for content-based image representations. However, due to perceptual subjectivity, it is not possible to identify a single best representation for a given feature. For instance, in the case of a texture

feature, a number of different texture descriptors are typically used, such as the Tamura texture representation [83], co-occurrence matrix-based [25], Wavelet transform-based [2], Fourier power spectrum-based [56], etc. Besides, a suitable representation is usually task dependent as well [61].

The idea of CBIR emerged as a complementary approach to earlier text-based image indexing and retrieval. This happened at a time when large-scale image collections started to appear, thus rendering manual image annotation infeasible and impractical, especially in near-real time applications. More importantly, the automated approach to indexing offers the possibility of dealing with the issue of perception subjectivity and annotation imprecision that originated in a combination of the subjectivity of human perception and the richness of the image content [25].

1.2.1 Content-based image retrieval vs. image classification

Content-based image retrieval (CBIR) and image classification are closely related: the ultimate aim of image classification is to generate linguistic terms that can be used for semantic image indexing and, therefore, organisation of and retrieval from image databases. The aim of CBIR, on the other hand, is to provide methods for searching image databases based on visual features. However, while there are many similarities between image retrieval and classification tasks, image classification is considered an easier task as explained in the following.

The aim of image classification is to categorise an image into one of predefined, usually broad mutually exclusive classes: indoor or outdoor, cityscape or landscape, building or non-building, no-people or people-present, etc. For a classification, the available training set can be made arbitrarily large (of course, bearing in mind the associated costs). Image classification is oftentimes approached in a hierarchical manner. Decomposing the problem into a set of two-class classifications, conducted in a number of stages, whilst using a simple feature representation tailored to the task at the hand, has been shown to be one of the most effective approaches [82,88].

However, in the case of image retrieval (by example), the number of potential image classes that a query image may belong to is large and remains unknown until the time a query is made.

This lack of knowledge on the query image's class prevents us from selecting the most suitable or most discriminative subset of features. Furthermore, the input space is usually high-dimensional, while the training set on which to learn is much smaller than the one available in classification [90].

1.2.2 The "Semantic Gap"

Notwithstanding the apparent richness that a content-based representation may provide in terms of high-dimensionality, numbers and the different types of visual descriptors used, there still remains a gap between such a low-level representation of an image, which is semantically poor, and the high-level semantics that are inherent in a human query [75]. In the literature, this gap between the user's semantic query and the low-level information extracted from images is referred to as *the semantic gap*. In [78], Smeulders *et al* define the semantic gap as follows:

> "....The lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation."

The main challenge in closing this gap is how to best translate a user's request into the "language" of low-level features. How can we, for instance, adequately capture the semantics of an image depicting a piece of baroque architecture, or the Roman amphitheatre in the centre of Pula, a close-up of humans engaged in lively conversation, a joyous atmosphere of a Christening or sombre mood of a funeral, in terms of texture, colour and shape?

In [64], Mojsilović *et al* describe a set of psychophysical experiments conducted in an effort to "gain insight into the semantic categories that guide the human perception of image similarity". Having established the most important semantic categories (such as portraits, people indoors, people outdoors, crowds, cityscapes, architecture, waterscapes, landscapes with human influence, sky/clouds, animals, textures, etc.) in the first set of experiments, they conducted further experiments in order to find correlations between these semantic categories and low-level descriptors. The objective of the work was the identification of the most appropriate low-level descriptors that could best capture the semantics of each image category and thus help to bridge the semantic gap. Using a set of 40 features, for each of the 20

semantic categories they determine a feature combination that discriminates that category from all other images. Their results suggest that, even though the visual features cannot fully capture the entire semantics of an image, there exists a significant correlation between them.

1.3 Image and scene understanding

The term *image understanding* refers to the process of generating a linguistic or natural language description of a given image, or attaching a textual description to an image automatically. It provides a description of an image in terms of objects, places, people and events. Image and scene understanding is a part of high-level vision - "...the highest processing level in computer vision" according to Sonka *et al* [81]. Semantic interpretation of an image provides answers to questions such as: "What objects are present in the scene? What location does the image depict? What is happening, what event does it depict?"

There exist different levels of understanding in the hierarchy of image understanding: the lowest level is the level of objects. Further up, in order of complexity, understanding entails understanding of the relationships between the objects in the scene (spatial and otherwise). Understanding and interpreting the mood and atmosphere the imaged scene conveys is the most complex task and comes at the very top of the image understanding hierarchy [52].

Thus far, image understanding has been successful in constrained domains such as medical and military applications, industrial inspection, etc. Unconstrained consumer photos, however, pose a far greater challenge due to the complete freedom associated with the process of image creation as well as the versatility in terms of content and scene composition. As a result, both the image content (e.g. no well-defined subject, no sense of conformance with the rules of good composition) and capture conditions (e.g. poor lighting, poor focus) vary a lot [49]. At present, image understanding mainly serves as a means to two distinct purposes: as an instrument to facilitate automated semantic indexing and content-sensitive adaptive image enhancement [20].

1.4 Digital photo collections: consumer vs. corporate photos

Digital photo collections constitute an important subset of general image collections. They can be divided into *consumer* collections and *corporate* collections:

- *Consumer photo collections* are heterogeneous by nature and generally unconstrained in scope. Consumer photos are captured without necessarily obeying photographic conventions (i.e. rules governing the capture of a good quality photograph) with regard to good composition, lighting, focus, exposure, etc. Such photos may be cluttered and contain a high number of objects in the scene, frequent occlusions, etc. In general, "they do not fall nicely into well-defined categories" [6]. While the number of distinct individuals occurring in an average consumer's photo collection is usually limited (e.g. family members, relatives, friends, etc.), such a collection is normally unconstrained with respect to its topic, content and composition. The number of photographers contributing to the collection may be relatively small, nevertheless the quality of their contributions may vary significantly. On a positive note, since fewer camera models are likely to be used, there is a greater consistency over types and quality of metadata associated with the images.

  As regards the motivation driving the capture of the image in a personal photo collection, it is largely emotional, it is merely to save the moment, to record an interesting object or an event worth saving from oblivion, to create a memento that can be shared now and kept for posterity. In contrast to photos in corporate photo collections, the temporal context of a photo in a consumer photo collection is valuable and informative, as people tend to take photos in bursts. Thus, the photos taken in temporal proximity are usually related, belong to the same event, share a common theme, etc. The temporal context can be exploited to reinforce the classification results and propagate labels to temporally close images.

- *Corporate photo collections* (digital archives and stock photography), on the other hand, can be broad in topic, but can also be domain specific. In the case of *digital archives*, owned and maintained by broadcasting corporations, the collections may cover a large time span. The number of distinct individuals appearing in such collections is virtually unlimited and their identity can often be quite specific and important. Collections of this type are normally the result of the work of a large number of photographers with their individual

styles. However, good rules of photography, such as good lighting, good composition, proper focus and exposure, conspicuous lack of clutter, etc., usually apply. In the case of *stock photography*, it is important that an image conveys a clear message. Therefore such images are unlikely to be cluttered or badly lit, the number of objects in the scene is expected to be low, etc. Essentially, those images are captured with a specific purpose in mind and thus focus on what is important. However, the identity of people present in such photos is rarely relevant. The time attribute may not be very relevant either.

In either case, the motivation driving the capture of images in corporate photo collections in general is by no means uncertain: it may be the intention to document an event for the wider public, or to create an image that could evoke certain emotions in people, provide some incentive, pass a certain message, and so on. They are characterised by their clarity of purpose.

1.5 "Is semantic image annotation feasible?"

The research thus far in the area of semantic image classification in broad topic image collections has shown that low-level features on their own do not have sufficient power to bridge the semantic gap between the high-level semantic concepts that humans communicate in, and content-based image description. The potential for filling that void may lie in using other contextual information that may be available. As capture devices become more powerful, more and more information is recorded at capture time [20]. For instance, the GPS information accompanying a digital photo easily answers the location-question. Dates and times, along with the location information can facilitate an automatic annotation of a photo with semantic labels with respect to the season (winter, spring, summer, autumn) and time of the day (dawn, morning, midday, dusk, night). Likewise, the EXIF's scene brightness tag could help determine whether the photo was taken indoor or outdoor. All this, in turn, could possibly assist other classification and annotation tasks by way of refining their results. In conclusion, the integration of visual features extracted from image content *with* information originating in other modalities is likely to offer an improved solution to the task of semantic image classification. Some of the challenges rest in identifying supplementary sources of information as well as finding smart and efficient ways of combining such diverse information. The work

described in this thesis explores some of these challenges.

1.6 Summary and thesis structure

This chapter provides context to the thesis by providing a brief introduction to the wider research area of content-based image retrieval, semantic classification and annotation of digital media. The target application of the methods presented in this thesis is an automated semantic annotation of large personal photo collections, with associated GPS location information. There are many possible semantic aspects of a large photo collection that might be explored such as indoor/outdoor, cityscape/landscape, waterscapes, sky/clouds, presence of people, crowds, presence of animals, architecture and so on. However, a semantic distinction that is consistently relevant is that between natural scenery and the human-made environment. Consequently, the focus of our work is on automatic detection of the presence of a dominant building object in a digital photo, i.e. classification of photos into *building* and *non-building* images.

Our goal is two fold. Firstly, we aim to identify a simple and computationally cheap, low-dimensional and low-level feature representation that could be a basis for detection of large buildings in natural images, captured by a ground-level camera, at a short to medium distance from the camera. Assuming an implicit presence of *indoor/outdoor* information, i.e. only outdoor images as an input, we identify an edge orientation based, multi-scale feature representation that, when evaluated on a constrained dataset of 1720 images, reasonably well captures the coarse building geometry/shape.

Secondly, we aim to make use of the available alternative, secondary sources of information and apply multi-modal fusion of low-level visual features with information from other modalities, that could facilitate automatic *indoor/outdoor* discrimination. We examine the impact of fusion of content-based information with contextual information in the form of the digital camera metadata, on the performance of the detector, and show that integration of the content and context of a photo positively affects the image classification rates. We implement early fusion and late fusion schemes to examine how each benefits the classification performance. The evaluation is performed on an unconstrained dataset of 8000 digital photos.

This thesis is laid out as follows. Chapter 2 provides an overview of the research efforts to date in the area of data fusion in the field of image and video analysis. The work on detection of large buildings in outdoor images relying on a simple low-level visual feature representation is presented in Chapter 3. In Chapter 4, we examine the potential of digital camera metadata for data fusion for the task of large building detection in unconstrained photo collections. We also look at the discriminative power of a selected subset of digital camera metadata for *indoor/outdoor* discrimination. Chapter 5 presents the results of evaluation of metadata-enhanced building detector. Finally, Chapter 6, summarises the thesis and discusses possible extensions of the method.

1.7 Publications

Part of the work in this thesis has been presented in the publications listed below.

- J. Malobabić, N. O'Connor, N. Murphy, S. Marlow, "Automatic Detection and Extraction of Artificial Text in Video", *Proceedings of the 3rd International Workshop on Image Analysis for Multimedia Interactive Services*, WIAMIS 2004, Lisbon, Portugal, April 2004.
- J. Malobabić, H. Le Borgne, N. Murphy, N. O'Connor, "Detecting large buildings in natural images", *Proceedings of the 3rd Int. Workshop on Content-Based Multimedia Indexing*, CBMI 2005, Riga, Latvia, June 2005.
- N. O'Hare, H. Lee, S. Cooray, C. Gurrin, G. Jones, J. Malobabic, N. O'Connor, A.F. Smeaton, and B. Uscilowski, "MediAssist: Using Content-Based Analysis and Context to Manage Personal Photo Collections", *Proceedings of CIVR 2006-International Conference on Image and Video Retrieval*, Tempe, AZ, 13-15 July 2006
- P. Wilkins, T. Adamek, P. Ferguson, M. Hughes, G. Jones, G. Keenan, K. McGuinness, J. Malobabic, N. O'Connor, D. Sadlier, A.F. Smeaton, R. Benmokhtar, E. Dumont, B. Huet, B. Merialdo, E. Spyrou E, G. Koumoulos, Y. Avrithis, R. Moerzinger, P. Schallauer, W. Bailer, Q. Zhang, T. Piatrik, K. Chandramoul, E. Izquierdo, L. Goldmann, M. Haller, T. Sikora, P. Praks, J. Urban, X. Hilaire and J. Jose, "K-Space at TRECVid 2006", *TRECVid 2006-Text REtrieval Conference TRECVid Workshop*, Gaithersburg, MD, 13-14 November 2006.

• N. O'Hare, H. Lee H, S. Cooray, C. Gurrin C, G. Jones G, J. Malobabic, N. O'Connor, A.F. Smeaton, and B. Uscilowski, “Automatic Text Searching for Personal Photos”, *SAMT 2006 - Proceedings of The First International Conference on Semantics And Digital Media Technology*, Athens, Greece, 6-8 December 2006.

Chapter 2

Data Fusion: State-Of-The-Art

Information fusion or the combination of information originating from different sources is one of the major areas of investigation in the analysis of visual content at present. In this chapter, we present an overview of some of the popular fusion-based approaches employed in the field of image and video analysis. The *late fusion* approaches we review here focus primarily on the issue of finding smart ways of combining data from different modalities at the decision level, i.e. ways of combining outputs of different classifiers into a single output. They also look at the issue of identifying the complementary modality that is best suited to a specific task (in both content-based image retrieval and image classification).

We begin by briefly introducing some of the fundamental concepts, such as data fusion itself and classification. Next, we present fusion approaches employed in image analysis, followed by those applied in video analysis, grouping them with respect to the types of features they combine. Finally, we conclude the chapter with a summary of the fusion methods presented here.

2.1 Fundamental concepts

2.1.1 Data fusion

Fusion is “a union by or as if by melting; merging of diverse, distinct, or separate elements into

a unified whole" [62]. Fusion of multiple inputs comes to humans naturally. In everyday life, humans combine visual, audio and tactile information for example in order to obtain a more accurate sense of the world and thus enhance their ability to act on it or react to it appropriately. In fact, most biological systems use some sort of combination of sensory information of different types so as to infer knowledge about the environment [24]. Thus the prime motivation for data fusion is obtaining more reliable knowledge about the surrounding world. The need for data fusion arises due to the following: (i) there is no perfect sensor nor perfect knowledge source and (ii) one modality is usually not sufficient to obtain an accurate and a complete picture of the observed phenomenon.

Hence, the need for the data fusion as a synergistic combination of information, which is performed in order to better understand the phenomenon under consideration, to achieve improved accuracy and obtain information of better quality [89].

Data from different types of sources may be combined. The following combinations of modalities for fusion in image and video analysis were encountered in the literature:

- low-level visual with textual features (i.e. accompanying text, captions, ASR);
- visual features with semantic features (detected using selected semantic object detectors);
- visual features with camera metadata and GPS information (i.e. data recorded in the EXIF header of each image).

Early vs. Late fusion

With respect to the stage at which the fusion is performed, fusion methods can be divided into early fusion and late fusion methods. Block diagrams of typical early and late fusion schemes are shown in Figure 2.1.

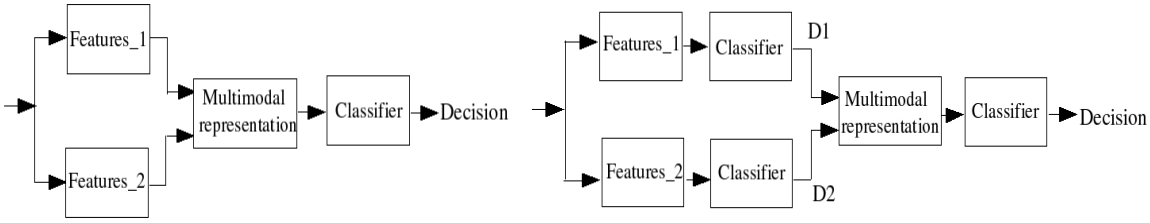


Figure 2.1 Block diagrams of typical early fusion and late fusion schemes.

In the case of *early fusion*, the fusion is performed at the feature level, i.e. the features are combined before classification (usually based on concatenated feature vectors for SVM, Bayesian networks and so on). In the case of *late* or *decision fusion*, the fusion is carried out later on in the process, that is at the decision level, by combining scores, ranks or probabilities obtained using classifiers on individual features or feature sets. In late fusion, individual "opinions" are combined in order to arrive at a final decision. The combination can be performed in parallel, serial, hybrid or in a multi-stage fashion.

Combination strategies for late fusion

A combination strategy determines how judgments from different classifiers are combined in order to arrive at the final decision. Two types of combination strategies are usually considered: *probability-based* strategies and *meta-classification* strategies.

- *Probability-based strategies.* Given a feature vector (that may belong to one of p classes) and k classifiers, its class will be determined as the class with the highest posterior probability. The final decision is based on the product *(product rule)* of posterior probabilities of all classifiers for a given class. The approach assumes that feature vectors from different modalities are conditionally independent and that prior probabilities for all classes are equal. In probability-based fusion an appropriate function (e.g. the sigmoid function) is used to map the classifier scores into (pseudo-)probabilities.

- *Meta-classification strategies.* The idea here is to treat the judgment from each classifier for each class as a feature in its own right. Based on feature vectors constructed in that way, another classifier, i.e. a meta-classifier, makes the final decision. So, unimodal classifiers first classify an image or video from the perspective of individual modalities then their judgments are concatenated into a new feature vector and the final decision is reached by the meta-classifier. A meta-classifier "observes more information" when making the final decision than any of the individual unimodal ones did in the first place [45,46].

2.1.2 Classification

A classification is a process whereby an entity, an object or entire image, is assigned to one of a set of classes, where classes are disjoint subsets whose elements have some common

properties. The classification objective determines how the set of entities are divided into classes [81]. A prerequisite for a good classification rate is a good quality feature representation [32]. Provided the feature representation is well chosen, vectors corresponding to a particular pattern should form a cluster, as intra-class variance will be small while the inter-class variance will be large.

Supervised vs. Unsupervised classification

The main distinction between the supervised and unsupervised approach to classification is that, in *supervised* classification, the number of classes as well as class labels (i.e. the classes themselves) are predefined and known in advance of classification, whereas in an *unsupervised* classification (aka clustering) patterns are assigned to hitherto unknown classes. As both the number of object classes and the classes themselves are unknown in unsupervised classification, the aim of the classification process is limited to grouping or clustering similar patterns together. In either case, the goal is to partition the feature space in a such a fashion that misclassification of objects is minimised [31].

In both supervised and unsupervised classification, the classification process is preceded by a *learning* step, during which a classifier is trained to distinguish between entities belonging to different classes. In the case of supervised classification, the training step involves presenting the classifier with a set of available examples along with their correct class labels (i.e. the expected output of classification) [81]. In contrast, examples presented to an unsupervised classifier are unlabeled examples (which is cheaper in terms of time and labelling effort needed), and the classifier then attempts to discover "natural" groupings without any knowledge of the class labels. The desired number of clusters may then either be decided in advance heuristically or learned during the training.

Exemplar-based vs. Model-based approach to classification

The task of scene or image classification is usually approached in two ways: *exemplar-based* or *model-based.*

- All *exemplar-based* approaches are based on learning, i.e. the knowledge inferred from the training examples is incorporated in the design of classifier [19]. Pattern recognition techniques are typically applied to image representations in the form of feature vectors, either based on low-level features (such as colour, texture, shape) or medium-level semantic

features (such as sky, grass, face). The inherent structure or partitioning of the feature space in the exemplars is learned during the classifier training and, in the next step, novel images are then classified based on the learned partitioning of the feature space. In *exemplar-based systems,* semantic "understanding" is achieved based on the similarity between the novel images and the exemplar images that were used in the training.

- *Model-based* approaches involve "building a model of the scene based on the expected scene configuration"[6]. The term *model* refers to a description that is both the simplest and the most general one, but still capable of describing the observations with minimum deviations. From this definition, it follows that a model can only describe the observation to some degree of accuracy and correctness [32]. The classification of novel images is performed based on the expected layout of the scene and expected relative sizes and positions of the objects. Explicit models of the scene are built using expert knowledge of the scene.

Clearly, the main limitations of the *model-based* approach lie in the fact that it only works in constrained and well-defined domains, requires expert knowledge and lacks generalisability. The *exemplar-based* approaches, on the other hand, require sizeable training sets, collection (as well as labelling, in the case of the supervised approach) all of which may be expensive, time-consuming and labour-intensive.

2.2 Fusion methods in image analysis

The image analysis fusion methods reviewed here employ the following combinations of features: (i) low-level visual features with textual features, (ii) low-level visual features with mid-level semantic features, and (iii) low-level visual features with digital camera metadata features.

2.2.1 Fusing low-level visual features with textual features

In [95], Yavlinsky *et al* conducted a comparative study of evidence combination strategies. They investigated the effectiveness of SVM meta-classification, CombMIN, CombMAX,

CombSUM and BordaFuse evidence combination strategies, combining visual features with textual annotation obtained by ASR (automatic speech recognition). A retrieved image is considered relevant if it comes from the same category as the query image. The *Comb-SUM* method is a simple linear combination of individual relevance scores. The *Borda Fuse model* is based on the Borda Count, an optimal voting procedure. It works as follows: each voter ranks a fixed set of *m* candidates in order of their preference. The top ranked candidate receives *m* points, the second-ranked received *(m-1)* points, and so on. The last on the list receives one point. In the case of unranked candidates the remaining points are evenly shared between the remaining candidates. The final ranking of the candidates is based on the total number of points. The Borda fuse model requires neither training data nor relevance scores, and is considered to be very simple and effective. *CombMIN, CombMAX* and *CombSUM* are the unweighted minimum, maximum and sum, respectively, of normalised relevance scores for each image over all classifiers.

The following low-level features were considered in [90]: the HSV global colour histogram, the HSV focus colour histogram, a colour structure descriptor, marginal RGB colour moments, a thumbnail feature, convolution filter-based features, texture variance, smoothness and texture uniformity.

The "bag-of-words" feature used in [95] is based on textual annotation obtained from the ASR transcripts. Each image is represented by a set of stemmed words and their normalised weights. The weights were determined using the TF-IDF formula [73] in which weights sum to one. The *term frequency* (TF) in a document is the number of times a given term occurs in a document, n_i, normalised by the number of occurrences of all terms in the document. The *inverse document frequency* (IDF) measures the overall importance of the term across all documents. The IDF is calculated by taking the logarithm of the ratio of the number of all documents, n_d, and number of documents containing the term, n_{di}. A high TF*IFF weight is thus associated with less common terms and measures the term relevance.

$$TF_i * IDF_i = \frac{n_i}{\sum_k n_k} * \log \frac{n_d}{n_{di}} \qquad (1)$$

The distances between the corresponding feature vectors are used to compare similarity of two images. The L1-norm is used throughout for visual features while the L1-norm raised to the power of three is used for the "bag-of-words" feature. A variant of the distance-weighted k-

nearest neighbour (k-NN) approach is used to retrieve images with respect to a given feature. This results in clustering of images that are similar with respect to a given feature.

The combination strategies in [95] use the similarity scores (i.e. distances of k-NN neighbours from query images) computed for each image to determine the overall ranking. A score vector for an image is defined as the vector containing *n* similarity scores for the images, each corresponding to one of *n* features. A linear SVM, acting as a meta-classifier, classifies the scores corresponding to each feature as provided by the k-NN algorithm. Novel images are ranked based on a single relevance score returned by the SVM. The SVM is trained using the scores and relevance information obtained by running the retrieval test on the training set. Relevance is defined as the distance of the score vectors from the hyperplane (i.e a linear weighted sum of the vector's components).

Performance of the SVM-meta classifier is compared against the standard combination strategies such as *CombMIN*, *CombMAX*, *CombSUM* and *BordaFuse* as well as the performance based on individual features on two test collections: the Corel collection and the TRECVID 2003 collection. The main conclusions drawn from the experimental evaluation are: (i) performance on the Corel data set is not indicative of the performance of the evidence combination on a real world data set such as the TRECVID set (greater care should be taken in identifying an appropriate benchmark collection), and (ii) the SVM meta classification appears to be a promising approach for specific topics in the TRECVID collection.

2.2.2 Fusing low-level visual features with mid-level semantic features

Luo *et al* [49] investigated the effectiveness of a Bayesian network-based framework for fusion of low-level features (colour, texture, shape) with semantic features (objects) for the task of semantic image understanding. The efficacy of the approach is demonstrated on three applications concerned with semantic understanding of digital photographs: detection of the main subject of an image, selection of the most appealing image from an event, and classification of images into indoor and outdoor scenes. It is argued in the paper that satisfactory results can be achieved in a specific image understanding task by using low-level visual features complemented by a small number of well chosen semantic object detectors. In other words, there is no need to detect every object in the scene, as selective object recognition suffices for the particular tasks selected.

Colour and texture low-level descriptors were extracted (either on pixel or block basis) for indoor/outdoor classification. The colour features were based on quantised colour histograms (3x64 bins) in the Ohta colour space [70] while the texture features were based on an MRSAR (Multiresolution Simultaneous Autoregressive) model [58]. A bank of filters was applied to an image in order to extract the low-level representation. However, the paper does not provide a detailed description of the features used.

The semantic features (in the form of semantic objects) were extracted using a bank of object detectors "of reasonable accuracy", two semantic features were extracted in this case: sky and grass. Regions containing grass and sky were identified using colour and texture features.

The proposed Bayesian network-based integration framework is considered to be a hybrid approach between *exemplar-based systems,* that use low-level features, and *model-based approaches,* that are built on the expected configuration of a specific type of scene. The choice of the Bayesian network as an instrument of feature integration is justified by its ability to incorporate domain knowledge both in the network structure and its parameters. The approach relies on both low-level and semantic features, and a probabilistic knowledge integration network allows all data to be both expressed and integrated in common terms of probabilities. The low-level and semantic evidence from a hybrid stream is fed to a Bayesian Network-based inference engine. Bayesian belief networks offer the capability of representing such diverse feature sets in a common modality (i.e. probability space) as well as of fusing those probabilities in order to obtain a final decision. The output of a Bayesian network may either be in the form of semantic labels for the entire image or "importance maps indicating different scene content".

The approach is evaluated on three tasks: main subject detection, emphasis image selection (the most appealing image in an image set pertaining to the same event) and indoor vs. outdoor classification. The MLBN (multi-level Bayesian network) system is benchmarked against versions of the system built using one naive and two different Neural Network-based (NN-based) classifiers. The authors claim that the major advantages of Bayesian Networks over Neural Networks are extreme stability in the case when some feature detectors are missing or faulty, as well as good generalization ability on novel data and ease-of-use.

Paek *et al* [68] present an approach for an *indoor/outdoor* scene classification, which combines image-based and text-based methods (i.e. image classifiers based on information originating

from the different modalities). They used a combination of the text accompanying images in multimedia documents (in a TF*IDF vector-based approach and the novel OF*IIF vector-based approach) in order to automatically annotate photos with content descriptions. The OF*IIF approach is a variation of the TF*IDF adapted for use in image analysis domain, where OF is the object frequency and IIF is the inverse image frequency.

The method makes use of the text accompanying an image (e.g. web pages, news articles with images and corresponding captions). The amount of text taken into consideration (e.g. full article, image caption, first sentence of the caption) and types of text information extracted is varied.

Two alternative approaches for *indoor/outdoor* classification based on text are examined: (i) TF*IDF scores and (ii) machine learning of words that distinguish well between the classes to categorise images using corresponding captions and articles. In the *first text-based approach* TF*IDF scores were computed for each document and class TF*IDF scores for all documents in each of the classes (indoor and outdoor): TF (term frequency) for a single document is the number of times a word appears in the document, TF for a given class is the number of times a word occurs over all documents in a given class. The inverse document frequency of a word (IDF) is the log of the ratio of total number of the documents and the number of documents that contain a given word. The product TF*IDF is higher for a word that is characterised by higher frequency within a document and low overall spread over the collection. Two scores were assigned to each image: one measures similarity with a prototypical indoor image and another similarity with an outdoor one.

The *second text-based approach* is based on automatically locating words (they selected 80 words) or phrases whose presence is a strong indicator of membership of one of the classes.

The image-based classifier used in [68] is analogous to the TF*IDF approach for text-based classification of images: they use OF*IIF scoring of images. *Object frequency (*OF) for a single image is the number of times an object occurs in the image, while an object frequency for a given class (indoor or outdoor) is the number of times an object occurs in all the images in a given class. The IIF of an object, *inverse image frequency*, is defined as the log of the ratio of the total number of images and the number of images that contain that object. The set of objects is predefined for a given set of training images in a cluster-based approach to defining and detecting of objects. In their experiments each image is divided into a 8x8 grid and a set of

colour and texture features is generated for each block (166 bin HSV colour histogram, 73-bin edge direction histogram of each block). The feature vectors associated with each block are clustered based on a single feature and the cluster centroid for each of the clusters defines an object. The next step involves detection of the objects defined in this way in both the training and test images. Similarly to their text-based approach, in [64] the authors computed OF*IIF scores for each image in the test set and class OF*IIF scores for the same objects. The dot product between the OF*IIF vectors for each image and each of two class vectors was computed in order to obtain two scores that measure the similarity of the image with the prototypical images of the classes.

Integration of the two approaches, the approach that draws on textual data (TF*IDF score of image captions) and the one drawing on visual information (OF*IIF score of image) is achieved by combining their respective scores. The aggregated score for an image is a weighted sum of dot products TF*IDF and OF*IIF with their respective class vectors. The probability density of the difference of two scores is empirically estimated by applying a rectangular smoothing window over the histogram of the difference. The integration of approaches results in a significant improvement in classification accuracy.

The method in [68] was evaluated on the task of *indoor/outdoor* classification. The raw data consisted of news articles containing images and associated captions. Images were labelled by 14 volunteers with each image being labelled by at least two volunteers. Only images for which the judgments of both subjects were in agreement were used in the experiments. In order to determine the upper bounds on the performance of classifiers, their experiments during labelling included restricting the amount of information available to humans: labeling based on text only and labelling based on the images only was performed. A conclusion drawn from these experiments is that humans make the *indoor/outdoor* distinction more easily from the image data, while the opposite is true for the automated classification system.

The experimental results show that for the first text-based approach, the most effective strategy involves the following: restricting the analysis to the first sentence of the caption, using normalised class frequency vectors and empirically estimating the probability density of the difference of the two scores. Experiments show that captions are much more closely related to images whereas the inclusion of the text from associated articles brings in background noise. The experiments demonstrated that by fusing the information from the two modalities, a classification accuracy of 86.2% could be obtained. This represented an improvement of

approximately 12% over image classifiers of the time, an improvement of approximately 3% over their text-based classifier alone and 4% over their image-based classifier alone.

2.2.3 Fusing low-level visual features with camera metadata

Boutell and Luo [8] investigated the use of camera metadata ("data about data") for improved semantic scene classification of digital photographs, through fusion of visual features with camera metadata. They considered three different classification problems (*indoor/outdoo*r classification, sunset detection and *manmade/natural* classification) and analysed the camera metadata statistics for each of the classes. To identify the most discriminative cue for a given problem, they calculated the average Kullback-Leibler (KL) divergence [39] of each cue for the two scene classes: the maximum average divergence corresponds to the most discriminative cue for a given task. The same analysis was applied to the cue combinations for joint distributions of variables. The analysis of camera metadata statistics shows that the following metadata fields are the most discriminative for each of the classification problems listed: `exposure time`, `flash fired` and `subject distance`.

Among the hundreds of metadata tags contained in the EXIF[1] header of a JPEG image, there are a number of tags that relate to image capture conditions such as: `Flash`, `FocalLength`, `ExposureTime`, `ApertureValue`, `FNumber`, `ShutterSpeed` and `SubjectDistance`, `ISOSpeedRatings`, etc. The authors categorised the relevant metadata tags into four families they considered useful for scene categorisation and valid beyond the applications they address in the paper: *Scene Brightness (*includes exposure time, aperture, f-number and shutter speed), *Flash, Subject Distance*, and *Focal Length.* It is claimed that the above features are mutually independent. However we note that scene brightness and flash are actually dependent as the use of flash affects the scene brightness. Similarly, camera focal length is dependent on the subject distance.

The authors proposed a probabilistic approach to the fusion of low-level content-based evidence (i.e. the output of a classifier) with the evidence based on the camera metadata. Following a statistical discriminant analysis conducted in order to identify the most discriminant cues for each of the classification problems, the visual content-based features and metadata cues were fused in the probability domain using a Bayesian network. All evidence

1 The EXIF components are explained in Chapter 4 in detail

was expressed in common terms of probabilities. Evidence fusion was performed in two stages: firstly, two separate classifiers were employed in order to classify metadata evidence on one side and low-level visual evidence on the other side. Low-level input was of a pseudo-probabilistic type (e.g. SVM output mapped into probabilities), while metadata input was either of Boolean type (e.g. flash used) or discrete type (e.g. quantised exposure time). At the second stage a Bayesian network completes the integration of both types of evidence.

The model was successfully applied to tasks of *indoor/outdoor* classification, sunset detection and *manmade/natural* scene detection. The experimental results for the three classification problems demonstrate that integration of content-based features and metadata contributes to improved classification performance. The experiments also showed that their classification scheme still works even in the absence of some cues and, more particularly, even in the absence of all content-based cues (in the case of *indoor/outdoor* classification). The results of *indoor/outdoor* classification based solely on metadata (no content-based cues) are comparable to those based on integrated content-based and metadata features. In fact, in the case of the *indoor/outdoor* classification task, metadata features on their own outperform the content-based features alone. In the case of *indoor/outdoor* detection, the best performing metadata cue combination results in detection accuracy of 92.2%, whereas the accuracy of the detection based on low-level visual features alone stands at 81%. By integrating low-level visual features with camera metadata, the detection accuracy is raised to 94.1%. In the *manmade/natural* scene classification task, use of metadata along with visual features improves the classification accuracy for an average of 2% over the entire operating range.

In [7], Boutell and Luo presented a probabilistic approach for *indoor/outdoor* scene classification of images in home photo collections based on integration of the image temporal context with the image content. The approach exploited the fact that, in a personal photo collection, unlike professional stock photos, each photo has a temporal context in the form of other photos that have been captured in close temporal proximity. They assumed that this contextual relation will be strongest in the closest proximity of an image.

In the Boutell and Luo approach, context and content were combined using a probabilistic model: a Hidden Markov Model (HMM) was used to model a sequence of images. A hierarchical classification method was employed. At the first level, each image was classified on its own based on image content using a SVM classifier. At the second level, classification

was performed using the combined evidence: classification based on content-based features plus the evidence gleaned from contextual relationship with the surrounding images. The approach assumes the following: (i) image classification at the first level, using content-based evidence, only depends on that image itself, (ii) the context of a node *i* only depends on the previous image/node *(i*-1*)* in the sequence.

Content-based evidence included colour and texture features: block-based colour histograms and wavelet texture features. Initial *indoor/outdoor* classification was performed using a Support Vector Machine classifier. The SVM's output was then transformed into pseudo-probabilities using a sigmoid function.

Image timestamps were used to determine the elapsed time between each two neighbouring images in an image sequence, thus providing a temporal context for every image in a sequence. The temporal context of an image was modelled using a HMM model, where the class of each image was represented by a node, and temporal dependencies between two nodes were represented by an edge. The HMM requires that two types of probabilities, transition probabilities and output probabilities, be either learned from data or set by an expert. *Transition probabilities* denote the probabilities of an image belonging to a class *i* given the classes of the images in its immediate temporal neighbourhood. For instance, an image surrounded by two indoor images can be expected to be an indoor image itself. Transition probabilities determine the strength of the class relationships between temporally close images: class relationship between the neighbouring images taken in a shorter space of time is expected to be stronger, i.e. the strength of the class relationship is inversely proportional to the the elapsed time between two images. In this work, transition probabilities were learned from the training set under the assumption that the strength of the relationship drops off exponentially.

The *output probabilities* are likelihoods of evidence (i.e. output of content-based classification) being observed given a true scene class. A sigmoid function was used in order to convert the real-valued output of the content-based SVM classifier into a pseudo-probability of an image belonging to a class. To maximise the probability of classifying an entire sequence of images correctly whilst keeping the algorithm complexity low, a dynamic programming algorithm (the *Viterbi algorithm* [22]) was used to perform optimisation (iteration through a sequence searching for the optimal path to each state from the start). The transition matrices (i.e. transition probabilities $P(C_{i+1} \mid C_i)$ as a function of elapsed time) were obtained by discretisizing the elapsed time between each two adjacent images in the test sequence and

mapping it back to learned probabilities that correspond to those elapsed-time bins in the training sequence. The classification was performed in two stages: the first stage was based on low-level feature representation (SVM) and the second stage consisted of imposing the temporal model (using metadata) on the results of the first stage using the HMM.

The comparisons of the performance of the temporal context model against the content-based classification only, shows a substantial increase in recall (approximate gain of 5% for ROC curve outdoor recall vs. indoor recall obtained by varying the bias of the SVM classifier).

2.3 Fusion methods in video analysis

Although not directly relevant to the work presented in this thesis, the core of which deals with classification of still images, we include the overview of the fusion methods in video analysis not only for the sake of completeness, but also due to the fact that, in many instances, video analysis could be viewed as still image analysis enriched with temporal information and audio. Furthermore, as discussed in the previous section, even the sequences of still images in personal photo collections do have a temporal aspect to them. The work of Boutell and Luo [7], demonstrates how the temporal context of a still image can be exploited for image classification. Lastly, any information on data fusion methods, even in the different domain, may be relevant to some degree.

Fusion methods in video analysis employ, amongst others, the following combinations of features: (i) low-level visual with textual features, (ii) audio-visual with textual features and (iii) audio-visual features using different statistical models for each feature type.

2.3.1 Fusing low-level visual features with textual features

K. Mc Donald and A.F. Smeaton [57] conducted a comparison of score-based, rank-based and probability-based fusion methods for video shot retrieval. They investigated a range of standard late fusion methods for combining classification results based on (i) multiple visual features (colour, edge and texture), (ii) multiple visual examples in the query and (iii) multiple modalities (text and visual). The comparisons were performed on three TRECvid collections

(2002, 2003 and 2003) as part of the TRECvid search task. They empirically established effective fusion methods that are suitable for different types of video search.

The premise for their work was that the retrieval results of a system can be presented either as a *ranked* or *scored result list.* While both lists are ordered lists of retrieved items sorted in order of their relevance to the query, the scored list also provides additional information in the form of scores. The scores measure relevance of the given item to the query, i.e. the degree of match. The score is also an index for the quality of the decision and shows more details about the relationship between the classes, whereas the r*anking* is a simple linear ordering of a set of items. The features used in their work were four low-level features combined with text-based features.

In [57], the authors represent the content of a video shot using four features: text obtained using automatic speech recognition techniques (i.e. ASR text), Hue-Saturation-Value colour space based histogram (16x4x4 bins), Canny edges (64 bins) and DCT-based texture (with 5 coefficients quantised into 3 bins each), using a 5x5 grid. This particular choice of image partitioning provided a limited, but still beneficial amount of spatial information. A *discrete language modelling approach* was used for the low-level features: shots were ranked in order of the probability of their language model generating the query (known as a query-likelihood approach). In order to be able to deal with low-frequency events (zero frequency in particular) and to reduce the impact of frequent events on the empirical distribution of features, smoothing was performed using a *collection model.*

A Hierarchical Mercer-Jelinek smoothed language model [35] was used for the ASR (automatic speech recognition) text feature: each shot was smoothed with the text from adjacent shots, from the whole video and from the entire video collection.

In the late fusion methods of McDonald and Smeaton in [57], matching was performed on individual features first and then the matching scores were fused, with the aim of improving upon the best individual retrieval result. *Rank-based methods* combine separate search results by way of summing the rank position of a document in different search result lists (e.g. Borda count and weighted Borda count [12]) *Score-based combination methods*, on the other hand, either sum the multiple retrieval scores or sum the scores from truncated result lists (e.g. the top 1000), and then multiply the average by the number of models that returned it (CombSUM [79]). When heterogeneous retrieval or feature models are combined it is necessary to perform

some sort of normalisation of retrieval scores in order to make the scores comparable (usually by mapping the scores to a [0,1] range).

In their multi-modal fusion experiments, retrieval results for the text and visual features were combined using variations of data fusion methods that were originally developed for combining the results of multiple text search engines. The methods were based on normalised score and rank and used one of the following combination functions: the average or the weighted average of the maximum of the individual search results. When features were combined, the result list for each feature was truncated to the top *N* results (*N*=1000). In the case of queries with multiple visual examples, the result list for each of the visual examples was truncated to *N=M/num_vis_examples*, where *M* was an empirically selected value in the range [1000, 3000].

Their results suggest that the normalised score fusion should be preferred over the normalised rank fusion as they noted that the distribution of scores holds valuable information that is lost when normalisation is performed based on rank alone. The result of the experiment involving fusion of the ASR text retrieval and the retrieval results of multiple visual examples show that CombSumWtScore (i.e. weighted average of the normalised scores of the top N results) is the best multi-modal fusion strategy.

Lin and Hauptmann [46] address the problem of classification of broadcast news video into *weather/non-weathe*r reports using SVM-based classifiers, and investigate different combination strategies for combining text features from closed captions and visual features from the image. Specifically, they compared a meta-classification combination strategy using SVM with probability-based strategies. The choice of classifiers (all SVM) for both text-based and image-based classification is justified by high dimensionality in the text feature space and the known ability of SVMs to work well in high-dimensional feature spaces.

In [46], news transcripts extracted from closed captions are used as sources of text features. In the text categorisation domain, a document is viewed as a "bag-of-words" where the order of words is considered irrelevant. Each individual word is treated as a feature, the document is represented as a feature vector (i.e. relative frequency vector) whose dimensionality equals the size of the vocabulary. Each feature value is a normalised word frequency (i.e. the number of times a word from the vocabulary appears in the documents normalised by the length of the document). *Word frequency* measures the significance of a word in the document. Stop word

removal and stemming are applied in order to reduce the dimensionality, which normally results in better performance. The feature vector is very long (vocabulary size is 19895), but it is sparsely populated i.e. most of its elements are zero.

Visual features in the form of colour histograms were extracted from a keyframe representing each shot. The keyframe was partitioned into a 5x5 rectangular grid and colours were mapped into a 5x5x5 cube in RGB colour space resulting in 125-dimensional feature vector for each block. By concatenation of feature vectors associated with each of 25 blocks, a long feature vector with 3125 elements was formed to represent an image.

Two types of combination strategies were considered: *probability-based* strategies and *meta-classification* strategies. The authors aim to achieve an improved classification performance by building several weak (as well as cheap) classifiers and then combining their scores. A *weak classifier* is a classifier whose accuracy is only slightly better than pure chance classification.

A performance comparison of unimodal classifiers shows that the text-based classifier has higher precision. However, the image-based classifier has higher recall, as well as a smaller difference between recall and precision values. A comparison of probability-based classifier (product rule) and meta-classifier demonstrate the superior performance of the meta-classifier as shown by an improved precision. The authors believe that naive estimation of prior class probabilities (i.e. all prior probabilities are equal) may be one of the reasons for inferior performance of the probability-based classifier. Also, in contrast to probabilities-based classifier, a meta-classifier "observes more information" when making the final decision. Namely, unlike the product rule which treats all classifiers equally and assigns them equal weights, the SVM meta-classifier learns the weights associated with different classifiers and in that way recognises that one classifier may perform better in recognising a particular class, but not all the classes. The authors also note the remarkable stability of SVM-based meta-classifiers even in a high dimensional environment, when using noisy data and simple features.

2.3.2 Fusing low-level audio-visual features with textual features

In [1], Adams *et al* presented a *learning-based* approach to semantic indexing of multimedia content based on cues derived from multiple modalities: audio, visual and textual features. They defined a lexicon of atomic concepts (e.g. sky, water, music, speech) and developed a set

of corresponding statistical models. They used Gaussian Mixtures (GMM), Hidden Markov Models (HMM) and Support Vector Machines (SVM) to model the atomic concepts in the lexicon. New concepts were defined in terms of concepts in the lexicon. High-level concepts were expressed in terms of the concepts contained in the lexicon and information in the annotated training data, using Bayesian Networks and SVMs as an integration framework for a late fusion approach. The paper investigates two main approaches to modelling of semantic concepts and events: probabilistic approaches (GMM, HMM and Bayesian Nets) and discriminant approaches such as SVMs.

Visual properties of a region are represented using a total of 56 visual features: normalised, linearised 3-channel histogram in HSV colour space (6-6-12 bins), 24-bin histogram of edge directions (Sobel detector applied on 3x3 windows) and moment invariants to describe the shape of each region. The low-level audio features are 24-dimensional Mel-Frequency Central Coefficients (MFCC).

The lexicon contains over 50 semantic concepts which are used to describe events, scenes and objects. The experiments used only a subset of those: visual concepts (rocket object, fire/smoke, sky, outdoor), audio concepts (rocket engine explosion, music, speech, noise) and multi-modal concepts (rocket-launch). For instance, the high-level concept of rocket-launch is inferred from the detected visual (atomic) concepts in multiple modalities.

Integration or fusion of information from different modalities may occur at different levels: (i) at the low-level features level, (ii) within atomic concept models or (iii) through the combination of several atomic concept models into multi-modal high-level concept models. The focus of the work presented in [1] is on modelling of atomic concepts using low-level features, and representing the high-level concepts using atomic concepts modelled across different modalities.

This fusion scheme obtains over 10% relative improvement in comparison to the best unimodal concept detector. Retrieval performance is measured using precision-recall curves and an overall-figure-of-merit (FOM), i.e. the average precision over the top 100 documents.

2.3.3 Fusing different statistical models for different audio-visual features

In [79], Smith *et al* described a *model-based* classification system that integrates features, models and semantics for automatic and interactive content-based video retrieval. Audio-visual descriptors were extracted on the shot level from a keyframe selected to represent the video shot. Using the audio-visual descriptors, they developed and applied models of scenes and events for classifying video shot content (i.e. assigning semantic labels) into broad categories such as: indoor *vs*. outdoor, nature *vs*. man-made, face detection, sky, land, water and vegetation. Statistical models for semantic concept modelling that they developed using the training data, were subsequently used to semi-automatically assign labels to novel video shots. All semantic labels were contained in "a lexicon for describing events, scenes and objects". The labels are automatically propagated to similar shots.

The following descriptors were extracted from each keyframe: colour histogram (166-bin HSV colour space), grid-based colour histogram (4x4 grid of the HSV histogram), texture spatial-frequency energy (variance measure of each of 12 bands) and edge histogram (using Sobel and quantisation to 8 angles and 8 magnitudes).

Statistical models were developed to model the following groups of concepts: events (fire, smoke, launch), scenes (greenery, land, outdoors, rock, sand, sky, water) and objects (airplane, boat, rocket, vehicle). Descriptors extracted from the video were modelled by a multidimensional random variable. Each semantic concept was modelled by two Gaussian Mixture Models (GMM) with 5 Gaussians each: a positive model for a given label (i.e. concept present) using positive examples in the training set and a negative model or garbage model (i.e. concept absent) for that label using negative examples. Parameter estimation was performed using annotated examples in the training set. The likelihood ratio (i.e. ratio of likelihood of being in a class Ω_1 and likelihood of not being in a class Ω_1 given a feature vector x) was chosen to be the measure of the confidence of classifying a test image correctly.

The objective of feature fusion in this paper is to combine multiple statistical models for the different video features. Statistical models generate semantic labels, each with an associated confidence score. Each of the descriptors is modelled using a separate GMM (whose parameters were determined using the annotated examples in the training set) and an image is

classified based on each of the descriptors, resulting in a number of individual classification confidences. Individual confidences associated with each descriptor are combined in a straightforward manner, by taking a sum, maximum or product of the individual confidence values to compute the final classification confidence. An alternative to this late fusion method corresponding to a concatenation of different descriptors into a single feature vector and building a single GMM model was not pursued due to the large dimensionality.

Using the ASR retrieval as a benchmark, the authors report that, although the performances varied for different search topics, the integrated interactive retrieval approach resulted in an overall performance improvement. The average number of hits per query over 46 general searches for an integrated approach was 4.3, whereas it was only 1.9 for the ASR-based approach.

2.4 Summary and Conclusions

In this chapter we present a number of fusion approaches employed in the analysis of the semantics of visual content, both in video and still images. As these fusion methods were evaluated on different datasets it is difficult to make direct comparisons in terms of performance. Moreover, some of the approaches were tested on consumer photographs, generally considered a very challenging dataset due to its unconstrained nature and high degree of ambiguity with respect to any predefined scene categories, whereas others were evaluated on less-real-world datasets such as the Corel dataset. This also makes direct comparison difficult.

Finally, various domains, from which the images were drawn, offer different alternative sources of information to complement the low-level visual feature representation. Examples are audio cues and temporal context in video, associated text in multimedia documents, temporal context for still images in personal photo collections, easily detectable semantic concepts, partial annotations and camera metadata in consumer photographs. A summary of the conclusions drawn as applicable to the work reported in the remainder of this thesis are:

- The power of low-level visual feature representation as a means to infer the semantics of an image is limited and it is thus necessary to explore and utilise data originating in

complementary modalities. Overall, individual performance comparisons between multi-modal and single mode approaches show superior performance of the multi-modal approaches to that of single-modal approaches.

- Successful approaches that overcome the limitations of low-level features, usually combine low-level visual descriptors (e.g. colour, texture, shape) with complementary information such as text [1,46,57,95], low-level audio features [1,79], mid-level semantic features such as grass and sky [49,68] or camera metadata in the case of digital photographs [7,8].

- Combination strategies employed in the approaches presented here are of two types: *probability-based* strategies and *meta-classification* strategies. In the case of *probability-based* methods, the final decision is based on the product *(product rule)* of posterior probabilities of all classifiers for a given class. On the other hand, a *meta-classification* is based on the idea of treating the judgments from each classifier for each class as a feature and concatenating individual class judgments into a new feature vector. The final decision is reached by the meta-classifier. A comparison of the *probability-based* method and the *meta-classifier* in [46] demonstrate the superior performance of the *meta-classifier* in terms of improved precision. The improvement is attributed to that fact that, compared to *probabilities-based* classifier, a *meta-classifier* "observes more information" when making the final decision.

- The majority of the methods are either pure *exemplar-based* or *hybrid*, and thus involve some degree of learning performed using annotated examples. This is a reflection of the fact that in the face of unconstrained image scenes, in consumer photographs in particular, building scene models becomes a virtually impossible task.

- The high dimensionality of feature spaces that sometimes result from feature fusion may be prohibitive for inference engines other than SVMs, which are distinguished for their ability to work well in high dimensional spaces. For this reason, the SVMs are expected to be particularly well suited for early fusion approaches. The authors in [46] also note remarkable stability of SVM-based meta-classifiers in a high dimensional environment when using noisy data and simple features. Overall, SVM appears to be the learning technique of choice along with the Bayesian networks.

- Broad-topic digital photographs in general, and consumer photographs in particular, constitute a rather challenging subset of still images due their their unconstrained nature.

However, digital photographs have an advantage of bringing with them a valuable and easy-to-obtain complementary source of information that also comes cheaply: namely camera metadata. The potential of camera metadata is well demonstrated in [8]: the results of *indoor/outdoor* classification based solely on camera metadata are comparable with that based on fusing content-based features with metadata. Moreover, metadata features on their own outperform the content-based features alone in the task of *indoor/outdoor* classification.

In summary, the following is highlighted as being of particular relevance to our objective here:

- superior performance of the multi-modal approaches over single-mode approaches;
- good discriminative power of some of the camera metadata features for the task of *indoor/outdoor* classification of consumer photographs;
- the SVM's ability to work well in high-dimensional spaces and "observe more information" when acting as a meta-classifier makes the SVM suitable for use as an integration device in both the early and late fusion approaches.

Hence, it seems feasible to pursue a multi-modal approach that combines low-level visual features with selected camera metadata, using a SVM for data integration, in order to achieve an improved detection of buildings in consumer photographs, which is our objective here.

Chapter 3

Detecting Large Buildings in Natural Images Using Visual Features

3.1 Introduction

Semantic concepts, such as objects, people, etc., are the main "instruments" that humans use to navigate through and retrieve examples from large image/video databases [64]. Semantic annotation of large image/video databases is thus essential if ease of access and use is to be ensured. Inferring the presence or absence of high-level semantic concepts from low-level visual features is a research topic that has attracted a considerable amount of interest lately.

Our objective in this chapter is to detect the presence of a large *building* object (i.e. *outdoor architecture* according to [64]) in an outdoor colour image within a general purpose collection of digital photos. All photos are taken by a ground-level camera in an otherwise unconstrained environment. In the image of interest, a *building* is either a single dominant object or one of the dominant objects. We aim to show that a feature representation based on a few carefully selected and physically meaningful low-level features, coupled with the high generalisation ability of an SVM classifier engine, may be sufficient to detect some high-level concepts, such as buildings. As there exist a number of methods that reasonably successfully address the issue of *indoor/outdoor* classification of consumer photographs [51,82], we assume the availability of contextual information in the form of an *indoor/outdoor* label.

In section 3.2, a short review of relevant work is presented, while section 3.3 provides details of the approach. The results of performance evaluation and conclusions are given in sections 3.4. and 3.5 respectively.

3.2 Literature review

A significant portion of research work in the area of building detection focuses on building detection in a constrained environment using multiple images of a scene (e.g. building detection in aerial photography). The majority of researchers, addressing either aerial or ground-level photography, utilise some sort of edge distribution-based feature as a low-level descriptor. In the following, we review some of the work on detection of buildings, and human-made structures in general, in ground-level photographs.

Vailaya *et al* [87] developed a procedure to qualitatively measure the saliency of a feature towards a particular classification problem based on a plot of the intra-class and inter-class distance distributions of that feature. They show that a specific high-level classification problem can be solved using relatively simple low-level features geared for the particular classes. The edge direction coherence histogram was found to have sufficient discrimination power to distinguish between cityscape and landscape images (an edge pixel is considered *coherent* if it belongs to a connected component in a given direction whose size is at least 0.1% of the image size). This feature is geared towards discriminating structured edges from arbitrary edge distributions. The presence of human-made objects or structure in an image results in an edge direction histogram that exhibits peaks at or around the significant edge directions, whereas the edge distribution for "natural" images appears to be of random nature, i.e. the distribution usually appears to be flat.

The Dorado and Izquiredo [18] approach is based on the MPEG-7 edge histogram descriptor (an 80-bin histogram representing the local distributions of directional edges within an image: 0°, 45°, 90°, 135°, and non-directional) and on the local and global distribution of edges. The approach exploits rough matching and problem domain knowledge through user relevance feedback, while classification is performed based on rule-based fuzzy inference. The image is spatially divided into 16 equally sized sub-images, each of which is further divided into a

given number of non-overlapping small square blocks. The blocks are divided into 4 sub-blocks and passed through 5 filters to assign them to a corresponding edge category. The edge distribution in an image is summarised by a 80-component feature vector (16 sub-images x 5 bins each). Fine-tuning is performed through relevance feedback.

The approach of Iqbal and Aggaraval [27,28] for detection of large man-made objects, such as buildings, bridges, towers, etc., is based on the perceptual grouping of image primitives according to Gestalt principles of perceptual grouping [72] (continuity, closure, proximity, co-linearity, co-circularity, symmetry, parallelism). Lower-level primitive image features, such as line/edge segments, are grouped hierarchically into higher-level structures aiming to reach a meaningful semantic structure. The goal of grouping is to identify image features that are likely to have arisen from some scene properties rather than accidental arrangements ("the principle of non-accidentalness"). For building images, a 3-component feature vector is used to represent an image to be classified into 3 classes: *building, intermediate* and *non-building.* Features used are: number of "L" junctions, "U" junctions and "significant" parallel lines in the total number of "retained" lines. In [29], they combine features based on perceptual grouping, colour features and texture features into a 66-dimensional feature vector to represent an image. Their experiments confirm the intuitive expectation that colour information does not have sufficient discriminative power for *building/non-building* classification on its own. Their method achieves good classification performance for broader classes such as man-made structures, but performs modestly on subclass classification within the man-made class.

Common to all three approaches outlined above is the focus on edge/line segments features and the use of orderliness or regularities that the presence of human-made objects in a scene generates in terms of edge distribution. An important limitation of an edge distribution based representation is the fact that the edge distribution in building images is a function of the perspective distortion. To some extent, perspective distortion can be dealt with by widening the histograms bin.

The work of Mojsilović and Rogowitz [64] describes a set of psychophysical experiments conducted in order to gain an insight into the broad semantic categories that govern human perception of image similarity and to understand the way in which users judge the similarity of photographs. Based on the experiments, they (i) identify the 20 most important (broad

semantic) categories in human similarity perception, (ii) model these semantic categories in terms of combinations of computable images features (low-level), and (iii) develop an appropriate similarity metric for classification and search of photographs. Among their qualitative findings they list that semantic cues such as “water”, “sky/clouds”, “snow” “and mountains” are very important and that the strongest cue is the presence of people. Regions in an image of the “human-made” class usually feature straight lines, sharp edges or geometric shapes, while regions in “natural” images feature rigid boundaries and more random edge distributions.

Boujema *et al* [5] present a proposal for an improvement of the commonly used colour histogram descriptor (a first order colour distribution with low associated computational complexity) by way of an adaptive weighting of each pixel's contribution and accumulative histogram. The pixel weighting schemes they propose are related to a local measure of non-uniformity computed in a pixel's neighbourhood and are based on evaluation of perceptual cues (e.g. corners, isolated corners), statistical colour area distribution and local colour relevance. The magnitude of all listed measures increases with the increase of local colour variability. By computing histograms on different windows and then combining them in “different ways of accumulation” (e.g. additive, multiplicative accumulation) to improve information of geometric distribution of colours, they demonstrate that the novel colour distribution is both easy to compute and achieves superior retrieval performance irrespective of the colour representation used.

In the work described here, we approach the problem of *building/non-building* classification of the whole image using simple low-level features suited for the classification problem at hand, resulting in a low-dimensional feature space. Our approach for detecting the presence of large buildings in consumer photographs is based on multi-scale analysis, from global to local level, and it relies on explicit edge detection. An SVM classifier engine is employed to infer the information about the presence of a large/dominant *building* object from the edge orientation-based features. We show that a few simple features with physical meaning coupled with the high generalization ability of the SVM can yield satisfactory classification performance comparable to that of the existing approaches. The key aspects of our approach are low-dimensionality and simplicity.

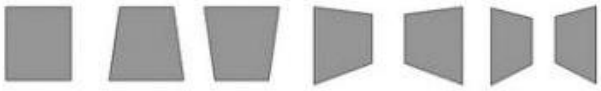
3.3 Proposed approach

3.3.1 Overview

Our objective is to detect the presence of a large *building* object in an outdoor colour image in a general purpose collection of digital photos. All photos are taken by a ground-level camera, at a close or medium distance, in an otherwise unconstrained environment. In the image of interest, a *building* is either a single dominant object or one of a few dominant objects in a possibly cluttered scene, with a complex background and frequently occurring occlusions. A *building* is a human-made structure, defined as "a structure with walls and a roof such as a house or factory", or "a usually roofed and walled structure built for permanent use (as for dwelling)" [62].

Figure 3.1 Variety of building shapes and views.

We approach the task of building detection as a classification problem, i.e. the assignment of an image to one of two classes: *building* or *non-building* using an inductive-learning method (a training set of labeled examples is used to learn the classification function automatically). Our approach, based on the classification of low-level feature representation of an entire image, is motivated by a simple observation: the most commonly occurring views of a *building* in non-artistic, amateur, general purpose consumer photographs can be summarised into six main types as shown in Fig 3.2.



a) b) c) d) e) f)

Figure 3.2 A building projection as a function of common viewing angles: (a) frontal view, (b) frog's view, (c) bird's eye view, (d) view from right, (e) view from left, (f) "street".

The presence of a dominant human-made object in a scene generates strong low-level evidence in the form of straight or elliptical line segments and edges [27,28,29]. Given that there is huge variation within the *building* class in terms of possible shapes that different types of buildings may take (as illustrated by Figure 3.1), we take the view that a coarse modelling of building shape/geometric properties is an appropriate approach. Dominant edge orientations of *building* object boundary edges and edges due to windows, doors, etc., are in most cases a combination of near-vertical and near-horizontal with near-45°, or near-135° degrees. Examination of the 36-bin edge orientation histograms of nine randomly selected typical images of *building, nature,* and *structure* in Figure 3.3 shows that "interesting events", which distinguish between *building* and *non-building* images, (e.g. large peaks), occur at around angles such as 0°, 45°, 90°, 135° depending on the viewing angle. This indicates that it may be sufficient to base our representation on relevant subsets of the edge histogram instead of the entire histogram. The edge segments are, in accordance with the Gestalt principles, which are discussed in more detail in section 3.3.2, expected to obey the rules of good continuity and co-linearity.

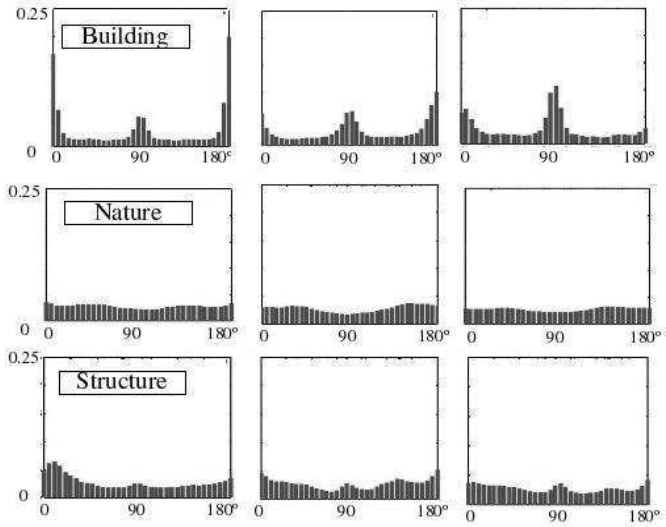


Figure 3.3 Comparison of normalised smoothed 36-bin edge orientation histograms for 9 randomly selected typical *building*, *nature* and *structure* images.

It is assumed that concepts that are large are also semantically important and that they usually, in consumer photographs at least, occur around the centre of the scene. We therefore incorporate localised information based on an analysis of a central rectangular region comprising 25% of the image size.

3.3.2 Low-level feature representation: edge orientation

Edge orientation histogram

The edge histogram is a first-order global shape descriptor characterised by its simplicity. It captures the general shape information in the image and it has been shown to be suited for use in a general purpose database [5]. The fact that it does not require segmentation as a prerequisite is a significant advantage considering that object segmentation is still a difficult problem. Other advantages of the edge histogram include its invariance to translation in the image and robustness to partial occlusion. However, edge histogram features are inherently neither scale nor rotation invariant. Scale invariance, which in this context means invariance to the absolute size of the object, is achieved by normalising the histogram by the sum of weighted contributions of all edge pixels considered. In this way we are able to deal with images (and buildings) of different sizes, avoiding the need for preprocessing.

Edge orientation histogram vs. edge direction histogram

The use of edge orientation histograms instead of edge direction histograms allows us to effectively reduce the number of bins considered, while retaining the relevant information (e.g. on parallelisms, co-linearisms) by reinforcing the relevant peaks in the 0° to 180° range. Both the orientation and direction are cyclic quantities. However, the direction is defined over the full angle range of [0°,360°], whereas the orientation is defined over the angle ranges of [0°, 180º].

In this application we aim to examine the edge angle distributions resulting from the presence or absence of a large building in an image. We are particularly interested in the presence of near-horizontal, near-vertical and near-diagonal edges etc. Obviously, such an objective does not necessitate making a distinction between the edges with 0° direction and that with 180° direction as for the purpose of this application it is sufficient to consider them both horizontal edges. Consequently, by choosing to “fold over” the edge direction histograms in order to

create edge orientation histograms we are able to maintain 5 degree quantisation while halving the number of bins. It is clear that this transformation causes no loss of relevant information, but allows for a reduction in the number of bins used to represent an image by factor of two.

Multi-scale generic object detection using Histogram of Orientation Gradients (HOG) has been studied in Scale Invariant Feature Transform (SIFT) by Lowe [48]. In SIFT, an image is represented by a number of keypoints or "key locations", each characterised by a local gradient orientation histograms (evenly spaced over 360° range). This results in a 128-dimensional feature vector describing each keypoint. Scale-space analysis is conducted on a pyramid of Gaussian smoothed images. "Key locations" or stable points in scale-space images are identified as local extrema of difference-of-Gaussian (DoG) filters at different scales (i.e. difference of Gaussian blurred images at adjacent levels in the image pyramid). The keypoint orientation is determined as a peak in a smoothed gradient orientation histogram covering the full 360° range in the local neighbourhood of the keypoint. Each pixel's contribution in a Gaussian window (with σ of 3 times that of the current smoothing scale) is weighted by the gradient magnitude and by a corresponding Gaussian window weight. A keypoint feature descriptor is computed as a set of 8-bin HOG histograms on 4x4 pixel neighbourhoods. As before, each pixel's contribution is weighted by the gradient magnitude and by a Gaussian with σ of 1.5 times the scale of the keypoint. Rotation invariance of the SIFT descriptor is achieved by constructing HOGs with orientations relative to the keypoint orientation.

In retrospect, the multi-scale, 24-dimensional EOH-based image descriptor used in our work could be thought of as consisting of two parts: a global, entire image based descriptor, and a single keypoint based descriptor. For each, a 4-bin EOH-based descriptor is extracted at 3 scales. The keypoint is assumed to be the centre of an image, the window is an unweighted rectangular one and its size is relative to the image size (i.e. 25% image size). A pixel's contribution is weighted by the relevance and coherency of its orientation in its 8-pixel neighbourhood and only relevant subsets of the EOH are used in our descriptor.

Block-based vs. object-based feature representation

Ideally, prior to feature extraction, an image is divided into meaningful regions corresponding to real world objects, features are extracted from each object in an image and then, based on the extracted patterns, each object is recognised or classified. However, after decades of research, there remains little disagreement in the research community that the accurate

segmentation of an image into objects is still a difficult task and the solution remains elusive. The alternatives to object-based feature representation are *global* and *block-based* representations:

- The simplest alternative that sidesteps the segmentation issue is *global* extraction of features i.e. extraction of features from an image as a whole. However, the main disadvantage of the global approach to feature extraction is the complete loss of local information.

- A *block-based* approach to feature extraction was identified as a good compromise between the aforementioned approaches. Here, an image is usually divided into a number of non-overlapping rectangular blocks or tiles, features are extracted from each block and then usually concatenated into a single feature vector. While avoiding the difficult issue of image segmentation, this approach still encodes some local information (as the position of the block in an image is implicitly encoded). For instance, blueish coloured blocks along the upper edge of an image may be an indication of an outdoor image, just as predominantly blue-coloured or greenish-coloured blocks at the bottom or around the centre of an image may be an indication of a water surface, etc. The limiting case of the block-based approach is actually a global feature extraction where the entire image is viewed as a single block [6].

In work of Dalal and Triggs [17], which is contemporary to our work presented in this chapter, a block-based Histogram of Oriented Gradients (HOG) representation for pedestrian detection (object scale known) is investigated. Their HOG descriptor is computed on a dense grid consisting of overlapping 16x16 pixel blocks of four 8x8 cells. Due to a 8 pixel block spacing, each 8x8 cell belongs to 4 different blocks, thus participating 4 times in gradient magnitude weighted voting. However, the cell's contribution to a block descriptor is each time normalised with respect to a different block. Pedestrian detection is performed using a 64x128 detection window.

In comparison, our approach can be thought of as having cells of size one pixel and two large, partly overlapping blocks: one corresponding to an entire image and another corresponding to a rectangular region in the centre of the first block, 25% of its size. Each pixel's contribution to either a global edge orientation histogram (EOH)-based descriptor or a single central block EOH-based descriptor is weighted based on relevance and coherency of its orientation in a 8-

pixel neighbourhood. Furthermore, our EOH-based block descriptor at each scale is characterised by only 4 orientation intervals being considered relevant, and is normalised by the sum of all EOH bins.

Scale-space or multi-scale representation - Pyramid approach

The optimal scale at which to conduct an analysis depends on the size of the object we aim to detect. "The detection of certain features in an image is optimal at a certain scale"[32]. However, the scale at which the object in a given image is observed, is unknown *a priori*. Therefore, it is necessary to obtain a representation of an image at different scales and to conduct the analysis on a number of scales simultaneously so as to be able to glean the evidence at different scales.

In a multi-scale representation, an original image $I(0)=I_0$, is associated with a sequence of simplified images $I(x,y,t)$, where t is a scale parameter. As the scale parameter t increases, the spatial resolution of an image in the sequence decreases. To obtain images from the fine resolution of the original image to those at coarse resolution, smoothing by convolution with Gaussian kernel is usually applied. In this way the so called *Gaussian pyramid* is generated. The Gaussian kernel and its variants have been shown by Lindeberg [47] to be the only smoothing kernels suitable for scale-space analysis. Smoothing in steps gradually reduces the level of detail, as shown in Figure 3.4. Multi-scale edge detection is usually better able to discriminate between texture edges and the edges that correspond to object boundaries.

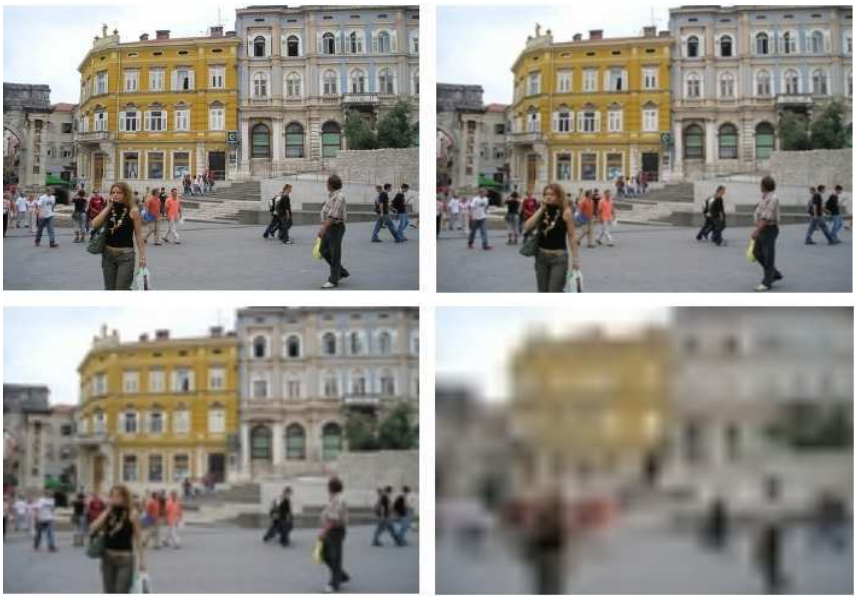


Figure 3.4 An example of multi-scale image representation (scaling by factor 2).

In Scale Invariant Feature Transform (SIFT)[48], the input image is incrementally smoothed with a Gaussian kernel $\sigma = \sqrt{2}$, and a scaling factor of $\sqrt{2}$ for subsequent smoothing, in order to achieve scale invariance. Gaussian-smoothed images are used both for i) identifying so called "key locations" as locations of minima and maxima in difference-of-Gaussian images, and ii) extraction of local descriptors to characterise "key locations".

Gestalt principles of perceptual grouping

The Gestalt theory of perceptual grouping as applied in computer vision has its roots in the Gestalt movement in psychology and the theory that our brains perceive "configurational wholes", instead of a collection of components. A Gestalt is a shape or "a whole form", "a structure, configuration, or pattern of physical, biological, or psychological phenomena so integrated as to constitute a functional unit with properties not derivable by summation of its parts" [62]. It refers to "a way a thing has been put together" [10]. This holistic operational principle is applicable to human visual recognition as well: instead of sets of curves and simple lines, we perceive figures and "whole forms". Under the Gestalt theory, our senses have an innate capability to search for forms: we perceive an object as a particular orderly collection of basic elements, or a specific arrangement of its elements. In the Gestalt interpretation of the mechanism of perception, our mind registers and comprehends the basic/primitive elements in the first stage, and then, whilst searching for some regularity and order among other things, recognises "the whole" in the particular arrangement of the basic elements such as lines, curves. Figure 3.5 illustrates the Gestalt principle of emergence of perception – the dog is recognised as a whole and as a specific configuration of basic elements such as lines, curves, etc. [56]. In contrast, in the conventional view of visual processing, the dog would be perceived as a collection of its parts such as legs, head, tail, etc., each of which is perceived individually [56].

The Gestalt principles of grouping low-level image primitives, such as edges, include the following ideas [10]:

- *Proximity or contiguity*– elements which are close by tend to be grouped together and are seen as belonging together;
- *Similarity* – elements which are similar in some way and share some attribute tend to be grouped together into an entity;

- *Closure* – missing elements whose addition would complete some entity are filled in by the mind. Similarly, the present elements are re-organised in a way to make a whole;
- *Simplicity* – elements tend to be grouped together to form simple figures;
- *Symmetry* - elements exhibiting symmetry in their arrangement are grouped together;
- *Good continuation* – as a result of our preference for continuous figures, we tend to ignore/disregard some interruptions;
- *Law of common fate* – elements that move in the same direction are seen as a unit.

Figure 3.5 "The dog" – an example of emergence in perception [56].

The edge segments in images containing large buildings are expected to exhibit symmetry and co-linearity, and obey the principle of good continuation. In *building* images, this will be manifested in the tendency of edge segments corresponding to structural edges to be aligned along straight lines of specific directions that outline the shapes of common building projections as shown in Figure 3.2.

3.3.3 Algorithmic details

As the appropriate scale for detection is unknown (it is only known that a building is at a close or a medium distance from camera), we adopt a multi-scale approach to edge detection and apply a Canny edge detector [13] at three scales. Scaling is achieved by smoothing with the

Gaussian kernel, as shown in Figure 3.6, with values of σ = 1; 1.5; and 2; empirically selected. The thresholds for hysteresis thresholding were set to 0.3 and 0.9 so as to ensure that most of the edge evidence generated by the texture edges is discarded while that due to edges corresponding to boundaries is retained. Non-maximum supression ensures that all edges are one pixel wide.

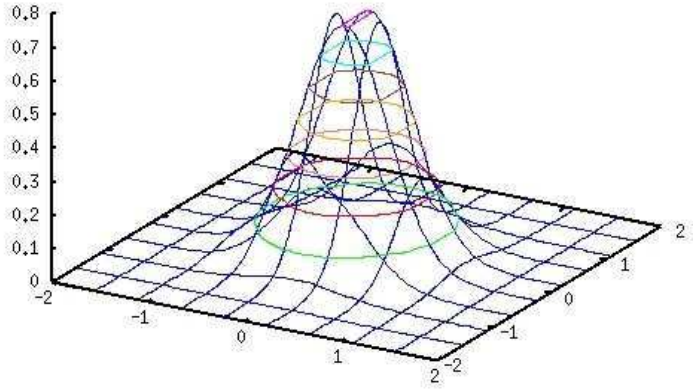


Figure 3.6 Gaussian function.

In addition to global edge detection, we extend our search for evidence to a sub-block corresponding to the central 25% of the image as we assume that if a building is really a dominant object, there must be strong evidence of human-made structure in the centre of the image. Based on the relevant edge intervals as shown in Figure 3.7, we construct a 5-bin histogram at each scale, globally and locally. Four bins correspond to the following edge orientation intervals: F_0=[0°,10°]∪[170°,180°], F_1=[35°,55°], F_2=[80°,100°], F_3=[125°,145°], and one bin is used for non-relevant edge pixels (i.e. all other edge pixels). Edge pixels contributing to the first four bins are referred to as "relevant" in the following. Each 5-bin histogram is then normalised by the sum of all five bins. A 24-dimensional feature vector is then formed by discarding the fifth bin and by concatenating the remaining 4 bins for each of 2 zones at each of 3 scales. Figure 3.8 illustrates the contributions of each of the significant edge orientation intervals to the total edge magnitude image for an image containing *non-building structure.* The comparison of contributions of relevant edge orientation intervals for *building, nature* and *non-building structures* images is shown in Figure 3.9. An underlying assumption of our method is a horizontal horizon line.

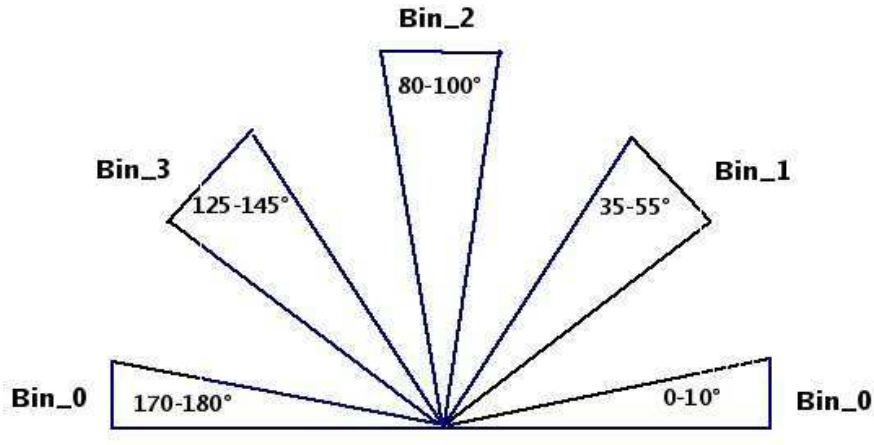


Figure 3.7 Histogram bins corresponding to relevant edge orientation intervals.

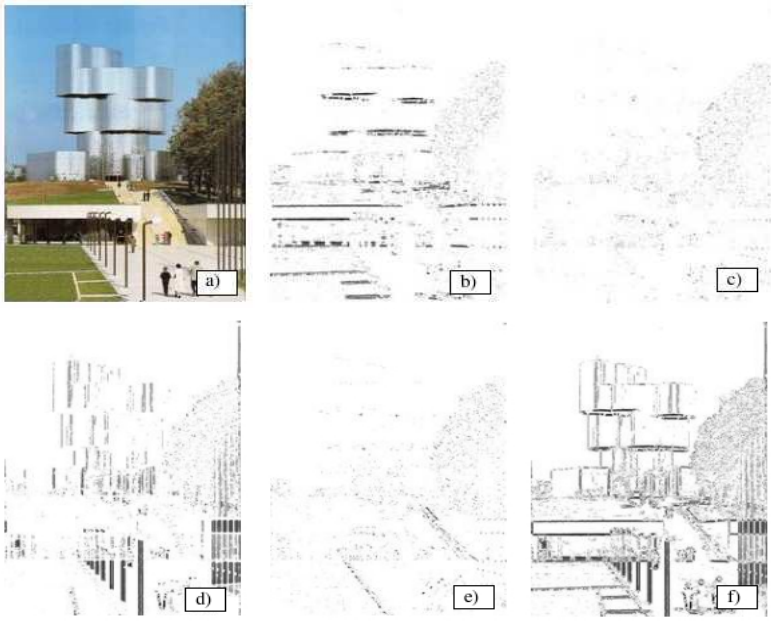


Figure 3.8 An example of outdoor *non-building structure* edge orientation contributions of relevant edge orientation intervals: (a) original image, (b) near-horizontal, (c) near-45˚, (d) near-vertical, (e) near-135˚, and (f) all relevant edge orientations.

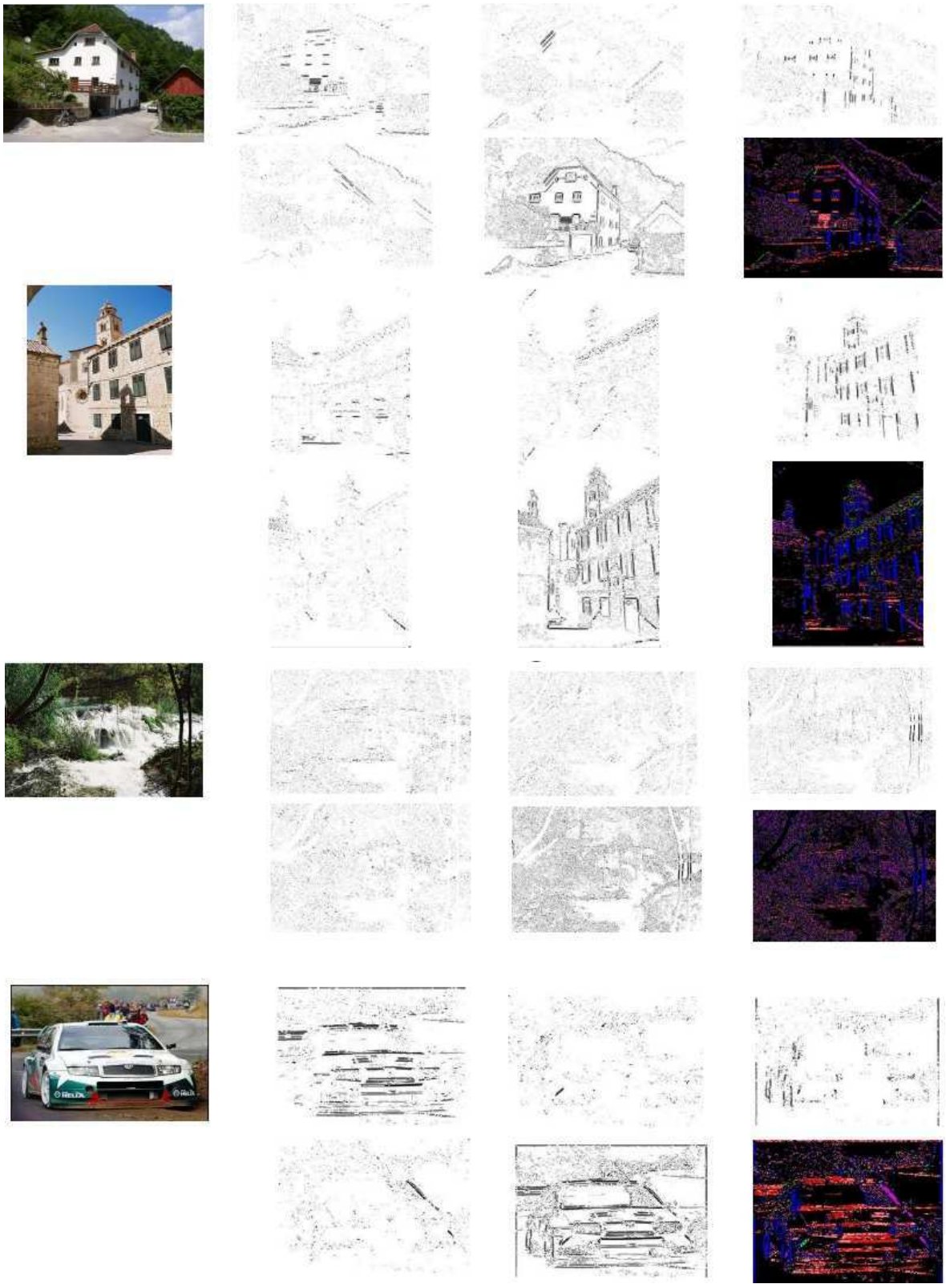


Figure 3.9 Contributions from different edge orientation intervals to the edge magnitude image for two *building* images, *nature* and *non-building structure* images: (a) original image, (b) near-horizontal, (c) near-45˚, (d) near-vertical, (e) near-135˚, and all relevant edge orientations (f) in black, and (g) colour-coded relevant edge contribution (from the top left to the bottom right).

Three versions of the approach, using different weighting schemes [9], were implemented by the author. We compute the 5-bin histograms, one for each region at each scale as follows:

$$H_e(5(j-1)+i) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} W_{ei}\, I_{ei}(m,n)\, I^{j}_{zone}(m,n) \qquad (2)$$

$$i=1,2,3,4,5; \quad j=1,2$$

where $H_{ej}(i)$ is an edge histogram bin corresponding to orientation i and region j (the first region is the entire image and the second region is the central 25% of the image), W_{ei} is the weight assigned to the contribution of an edge pixel with orientation i, I_{ei} (m,n) is an edge image component for orientation i, and $I^{j}_{reg.}(m,n)$ is a binary zone image (with value 1 for pixels in the region of interest, value 0 elsewhere).

In coherency weighting, a coherency check for relevant linear edge segments in a 8-pixel neighbourhood is chosen as a simple and computationally cheaper alternative to line detection using the Hough transform.

Gradient Magnitude Weighting

In the first version implemented, the edge pixel contribution to a given bin is weighted by the gradient magnitude, and the five-bin histogram is normalised by the sum of all edge pixel contributions in the image region being analysed so as to account for different image sizes.

Coherency Weighting: Weak coherency weighting and Strong coherency weighting

In the second version implemented, a weighting scheme which favours contribution of edge pixels more likely to belong to linear lines is introduced. The idea is to increase the importance of the relative contribution of the pixels that obey the good continuity rule. As illustrated in Figure 3.10, the 8-neighbourhood is examined for edge pixels with the same quantised orientation, termed *coherent* pixels, and the highest weight W_{ei}=1.3 is assigned to an edge pixel contribution, both of whose neighbours lie in a direction perpendicular to its gradient direction (in the case of one such neighbour weight W_{ei}=1.2 is assigned, and in the case of two such neighbours weight W_{ei}=1.3 is assigned).

In the third version implemented, a stronger weighting is used and the weights for *coherent* pixel contribution are increased to W_{ei}=2 and 3 respectively.

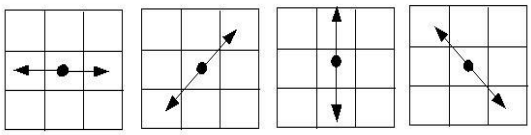


Figure 3.10 Coherency check in 8-neighbourhood for the edge angle θ of the central pixel: (a) $\theta \in [0°,10°] \cup [170°,180°]$, (b) $\theta \in [35°,55°]$, (c) $\theta \in [80°,100°]$, (d) $\theta \in [125°, 145°]$.

3.3.4 Low-level feature classification

The Support Vector Machine (SVM) [11] is a popular learning algorithm which has been extensively used in a number of applications, such as text classification, feature selection and hand-written digit recognition [37]. The SVM is characterised by high generalisation ability, and based on the idea of finding the hyperplane that best separates two classes after mapping the training data into a higher-dimensional feature space via some kernel function Φ. SVM classifiers are based on the hyperplanes of the type:

$$(\boldsymbol{w}.\boldsymbol{x})+b=0\,, \quad \boldsymbol{w} \in \mathbb{R}^N,\ b \in \mathbb{R} \tag{3}$$

where $\boldsymbol{w}$ is a weight vector, $\boldsymbol{x}$ is the training data, and b is a threshold. The corresponding decision function $f: \mathbb{R}^N \rightarrow \{\pm 1\}$ is:

$$f(x)=sign((\boldsymbol{w}.\boldsymbol{x}) + b\ ). \tag{4}$$

where $\boldsymbol{x}$ is a feature vector to be classified. The hyperplane is constructed by solving a constrained optimisation problem whose solution, a weights vector $\boldsymbol{w}$, is expressed in terms of a subset of training examples that lie on the margin: $\boldsymbol{w}=\sum_i \alpha_i \boldsymbol{x}_i$. This subset of training examples, called Support Vectors, carries all the relevant information contained in the training set. Thus the final decision function, $f(x)=sign(\sum_i \alpha_i (\boldsymbol{x}.\boldsymbol{x}_i)+b)$, where $\boldsymbol{x}$ is a new feature vector to be classified and $\boldsymbol{x}_i$ are support vectors, depends only on the dot product of the feature vectors.

One of the advantages of SVM over other classifiers is its speed, as the number of points that the SVM evaluates when a new point is classified is equal to the number of support vectors (usually significantly smaller than the number of training examples).

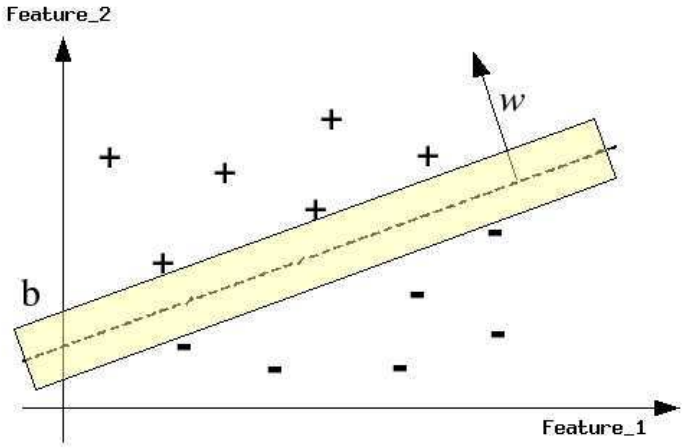


Figure 3.11 Geometric interpretation of Support Vector Machine in 2-D space.

We use the SVM^{light} [36,37] classifier which outputs a confidence measure for each test sample, the sign of which determines the class membership (if the score is positive, the example is labelled as a class member) while its absolute value gives an indication of the classification decision confidence, i.e. the distance from the separating hyperplane.

3.4 Experimental evaluation

3.4.1 Dataset

In order to evaluate the performance of the method, we use a diverse collection of 1720 images (consumer photographs), split into two sets: two different subsets of 200 images were used for classifier training/learning and the remaining 1520 images were used to evaluate the performance of the trained classifier. The dataset consists of images of arbitrary sizes in both portrait and landscape format. The images were collected from various sources:

- photo albums on the Internet,
- scanned from personal photographs, and
- donated digital photographs.

Non-building images include several sub-classes such as: nature (beaches, forest, field, water body, sunset, sunrise, etc.), large human-made-structure-other-than-building (boats, ships, cars, wheels, monuments, windmills, etc.), close-ups of flowers, fruit, animals and people.

Particular care was taken to ensure (i) that the data set is almost evenly split between *building* (769) and *non-building* (751) images, (ii) that the dataset includes images of objects that may easily be misclassified as buildings (113 *non-building structure* images or 15% of *non-building* images) and (iii) that the intraclass variance of the *building* images is sufficiently large (churches, cottages, skyscrapers, castles, huts, family houses, etc).

For the creation of a groundtruth, we apply a single label model assuming that all images can be singly labelled. Each image was labeled by two human subjects and a class was assigned based on the subject's perception of the dominant class in a given image.

3.4.2 Classifier training

Leave-one-out validation[2] [48] on the training set of 200 images (100 *building*, 100 *non-building*) is performed in order to determine the classifier parameters. The SVM with linear kernel is trained with different values of *cost factor* (which controls the ratio of misclassification penalty for the class and non-class members and corresponds to translation of the separation plane). As a criteria for selection of the SVM model we use the break-even-point (BEP) on the training set and a classifier with cost factor of 1.3 was selected. The *BEP point* [15] is defined as the point for which the values of precision and recall are equal. As shown in Figure 3.12, the BEP value of cost factor is identified as an intersection of precision and recall curves as functions of cost factor, on the training set.

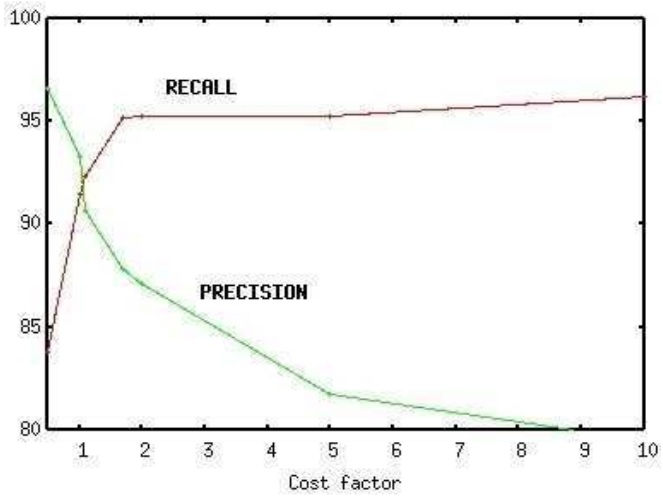


Figure 3.12 Determination of recall/precision break-even-point on the training set for classifier selection.

2 Given a set of *N* images, (*N*-1) images are used for training and the remaining image is used for testing

The projections of the training set patterns, *building* and *non-building*, into the 2-D and 3-D feature spaces are shown in Figures 3.13 and 3.14 respectively. As can be seen from the projection into the near-horizontal/near-vertical plane, shown in Figure 3.13, *building* patterns tend to have more edges aligned along both horizontal and vertical orientation than *non-buildings*. Furthermore, the number of edges aligned along the vertical direction is larger than that of edges aligned along the horizontal direction, i.e. the vertical edges dominate over horizontal ones in *buildings* images. The separation between the two classes in the pattern projection into the near-45º/near-135º plane is less evident. However, it is obvious that *non-building* patterns tend to exhibit large number of edges aligned along diagonals.

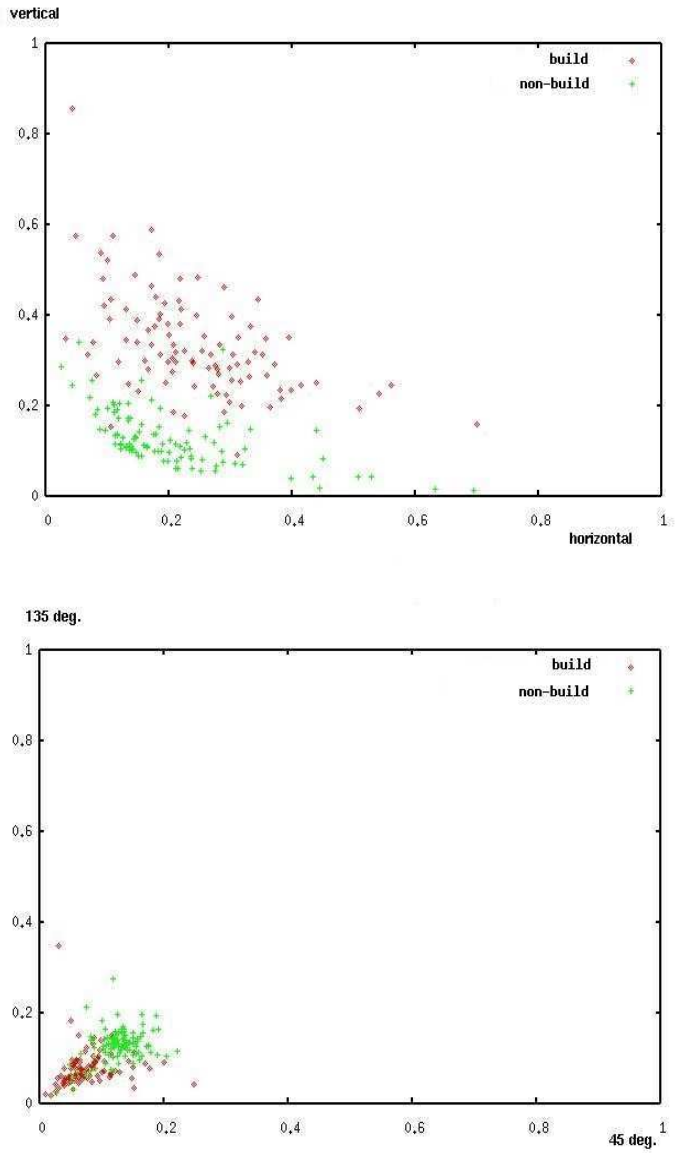


Figure 3.13 Projections of the training patterns into 2-D feature space: (a) near-horizontal/ near-vertical plane and (b) near-45º/near-135º plane.

The separation of patterns corresponding to the two classes are even more evident in the training patterns projections into the 3-D feature space as shown in Figure 3.14. Once again, dominance of vertically aligned edges in *building* images is demonstrated.

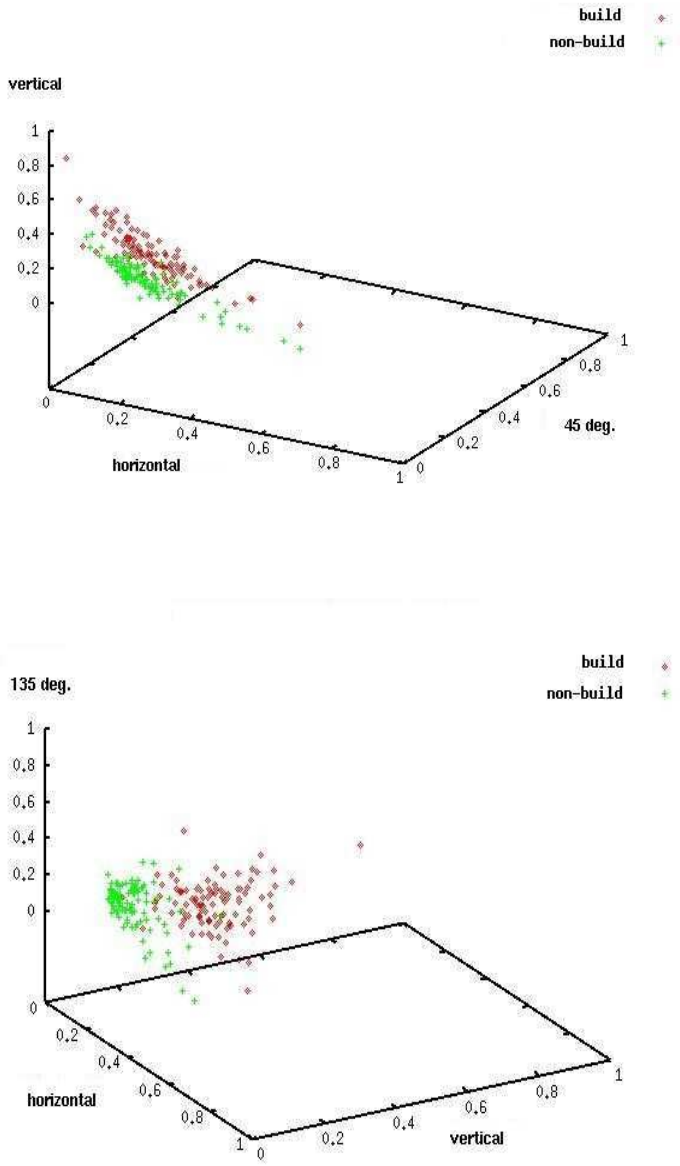


Figure 3.14 Projections of the training set patterns into 3-D feature space: (a) near-horizontal/near-45º/near-vertical plane, and (b) near-horizontal/near-vertical/near-135º plane.

3.4.3 Classification based on low-level features and discussion of experimental results

As a performance measure we use classification accuracy, recall and precision on the test set of 1520 images. *Classification accuracy* is the fraction of all images that has been assigned to a correct class. *Recall* is the fraction of *building* images that has been assigned to a *building* class, while the *precision* is the fraction of images assigned to the *building* class that actually belong to the *building* class. For each approach we evaluate here, we use a linear SVM classifier selected based on the BEP point and leave-one-out validation [37] on the training set.

Experiment 1 - the effect of coherency weighting

In order to determine the impact of weighting, we compare the performance of three different versions of the method: one with edge magnitude weighting, one with weak coherency weighting and one with strong coherency weighting with the MPEG-7 edge histogram descriptor [65]. The results presented in Table 3.1 show that the strong coherency weighting scheme outperforms both weak coherency weighting and edge magnitude weighting, as well as the MPEG-7 edge histogram descriptor.

Table 3.1 Comparison of experimental results for different methods (200 training images, 1520 test images).

	Accuracy	***Recall***	***Precision***
Grad. Magnitude Weighting	85.52	81.27	89.16
Coherency Weak Weighting	87.30	83.38	90.81
Coherency Strong Weighting	88.22	84.01	92.02
MPEG-7 Edge Hist. Descript.	84.93	79.45	89.59

Experiment 2 – the effect of local information

In order to verify the hypothesis that the inclusion of the localised edge information pertaining to the central 25% of the image actually improves classification performance, we compare the performance of the 12-component global feature representation and the 24-component feature representation (global+local information) for strong coherency weighting. The results in Table 3.2 confirm that, for this particular dataset at least, the incorporation of localised information positively affects the classification rate. The examination of misclassified images in both cases shows that this improvement is due to a reduction in the misclassification of *structure* images.

Table 3.2 Comparison of performance of 12-component and 24-component representation for strong coherent weighting (200 training, 1520 test images).

	Accuracy	***Recall***	***Precision***
12-component (global)	86.18	81.40	89.96
24-component (global+local).	88.22	84.01	92.02

By closely examining the misclassified images, we observe that most frequently misclassification occurs in the case of scenes containing dominant human-made structures other than buildings with edge distributions similar to that of buildings, such as those shown in the top rows of Figure 3.15 (a-c). In other cases, the misclassification occurs due to strong regular textures such as the presence of tree trunks in close proximity to camera, as can be seen in Figure 3.15 g).

Another difficult example is the Giant's Causeway (a naturally occurring outcrop of hexagonal basalt columns in Northern Ireland) shown in Figure 3.15 h). This natural feature exhibits an exceptionally high degree of regularity and attributes we normally associate with human-made objects.

Figure 3.15 Typical non-building images misclassified as buildings.

We also observe misclassification of *building* images due to the fact that edge orientation based features are not rotation invariant, as can be seen in Figure 3.16. The two building

images on the left were misclassified with a high degree of confidence. The images on the right side are those that are correctly classified.

The performance of our approach is comparable to that of existing approaches. However, we have to emphasise that we used our own dataset and a different number of training examples so that we are not in a position to make a direct comparison. Dorado *et al.* [18] report similar recall and precision on a test set of 3000 TREC images using 115 images for training. The user interaction improves the recall and precision to 86.31 % and 86.25% respectively. Iqbal and Aggarwal [27] validate their approach on 120 images (using 30 images for training) and report a recall of 80% and precision of 83.72%.

Since the image set used by Dorado *et al.* [18] was not available, we used our own dataset to compare the performance of a standard MPEG-7 edge histogram descriptor used in [18], with the performance of our edge orientation-based descriptor. As can be seen from Table 3.1, both of our coherency weighting approaches outperform the MPEG-7 edge histogram descriptor on a common dataset.

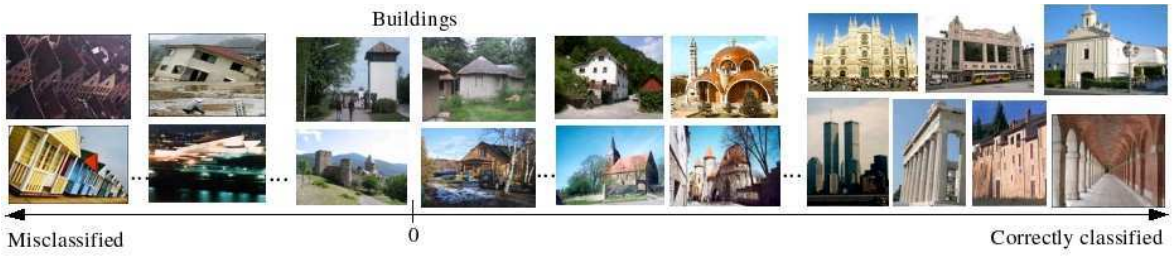


Figure 3.16 Classification results for *building* images in order of decision confidence, i.e. distance from the separation plane.

3.5 Conclusions

In this chapter, we present an approach to *building/non-building* classification of outdoor consumer photographs based on a few simple edge-orientation features with physical meaning, extracted at three scales, and used in conjunction with an SVM classifier engine. Experimental results on a diverse dataset of 1720 images show that the performance of our method is comparable to that of existing approaches. However, the results also show that an improvement is required in order to overcome the lack of rotation invariance and reduce misclassification between buildings and other human-made structures. Future work in this area should focus on extensive comparison with other techniques.

Chapter 4

An Improved Building Detection Using Camera Metadata

This chapter provides details of the proposed approach to applying data fusion methods to the task of large building detection in consumer photographs. The approach presented in Chapter 3 assumed all outdoor images. This is in fact quite a strong assumption considering that in real-world collections, indoor photos may constitute a significant portion of the collection. For instance, nearly 23% of the photos in the MediAssist [59] collection were taken indoors. The new approach we present here is an extension of the previous approach that (i) overcomes the strong outdoor assumption and (ii) provides better overall performance by reducing the misclassification rate of indoor photos containing human-made structures into *buildings*. We explain the motivation in section 4.1 and re-examine the efficacy of our edge orientation descriptor in this modified context. In section 4.2 we briefly touch upon the hierarchical approach to image classification, while in section 4.3 some approaches to *indoor/outdoor* classification are briefly reviewed. In section 4.4, we introduce a complementary set of features, based on camera metadata, and examine their discriminative power for the task of *indoor/outdoor* classification. Having established the most salient camera metadata tags, we select a subset to be used in experimental evaluation in Chapter 5, and propose an approach in section 4.5.

4.1 Introduction

An automated indexing of digital photographs with semantic concepts remains an important factor in improving the performance of the existing content-based image retrieval systems. However, semantic classification of (or, detection of high-level semantic concepts in) images in unconstrained, broad-topic, general purpose image collections is a challenging task and as such remains an open problem even after some years of research. Moreover, the existing approaches have mostly been evaluated in constrained environments (e.g. the Corel dataset). On the other hand, the review of the literature in Chapter 2 testifies to the fact that the power of a low-level visual feature representation as a means to infer image semantics, is limited. Furthermore, there exists broad agreement that using features from a single modality rarely provides enough information for the detection of high-level semantic concepts. The solution is, thus, increasingly sought in combining, or fusing, the evidence originating in different modalities, i.e. in a multi-modal approach to image understanding and semantic scene classification [20]. Overall, individual performance comparisons in the surveyed literature demonstrate superior performance of multi-modal approaches over single mode approaches.

The huge increase in the number of digital photos generated in recent years has put even more emphasis on the task of image classification of unconstrained datasets. Consumer photographs, a typical example of an unconstrained dataset, comprise a significant portion of the ever increasing digital photography corpus. Due to their unconstrained nature and inherent diversity, consumer photographs present a greater challenge for the algorithms (as they typically do for image understanding) [86]. Fortunately, digital photographs usually offer a valuable additional piece of information in the form of camera metadata that complements the information extracted from the visual image content.

Our approach to the detection of large buildings in unconstrained photographs aims to combine the low-level visual evidence with the camera metadata evidence. A set of features based on the complementary information available with digital photographs, which is embedded in the EXIF header, is fused with the low-level visual features (based on edge orientation histograms). The approach is evaluated on a diverse, unconstrained photo collection comprising 8000 genuine consumer or non-professional photographs.

The classification problem of interest here, i.e. the large building detection, can be defined as follows: given an arbitrary digital photo (but with proper orientation), determine whether there is a large/dominant building present in it. In Chapter 3 we present and evaluate an approach to classification of images into *building* and *non-building* images. However, the building detector is developed under the assumption that all input images are outdoor images, and this may not be true in reality. While demonstrating an overall satisfactory performance on outdoor images, with 88% accuracy and 84% recall rates on outdoor images, it has been already shown that the approach has difficulties in disambiguating *buildings* from large *non-building structures* (i.e. other human-made objects that exhibit similar edge orientation distributions). Likewise, indoor images may frequently contain human-made structures, such as large pieces of furniture, shelves, etc. The presence of such objects in a scene generates edge orientation distributions which somewhat resemble those generated by the presence of large buildings in outdoor images. As illustrated by Figure 4.1, and in contrast to the *nature* images characterised by a relatively flat histogram with low, wide and round peaks, i.e. more random distribution of edge orientations, the presence of human-made structures tends to generate narrow spikes in the edge orientation histograms. Consequently, the patterns corresponding to both the *outdoor non-building structures* and *indoor non-building structures* are also located in close proximity in the feature space to patterns of *buildings.* As such, they are difficult to separate from *buildings* and this results in frequent misclassification. In any case, this conforms with the view that a single image attribute usually lacks sufficient discriminatory information [34]. A comparison of typical edge orientation distributions for a *building*, an *outdoor non-building structure,* an *indoor* scene and *nature* images are shown in Figure 4.1.

Due to the fact that buildings can only occur in an *outdoor* photo (apart from a photo of a building picture captured indoors, or looking through a window to the outdoors), we expect that the capability of distinguishing between *indoor* and *outdoor* images will contribute to an improved building detector accuracy. One of the main differences between *indoor* and *outdoor* daylight photos is in their scene brightness levels: natural lighting in *outdoor* photos is very significantly stronger than artificial lighting present in the images captured indoors. The usefulness of some of the camera metadata, such as exposure time, flash use and subject distance, for *indoor/outdoor* classification has already been demonstrated in [51]. By fusing the selected camera metadata with the low-level visual features, which were, on their own, sufficient to detect buildings in outdoor photos, we aim to enable the detector to effectively handle all images, thus improving the overall classification rate.

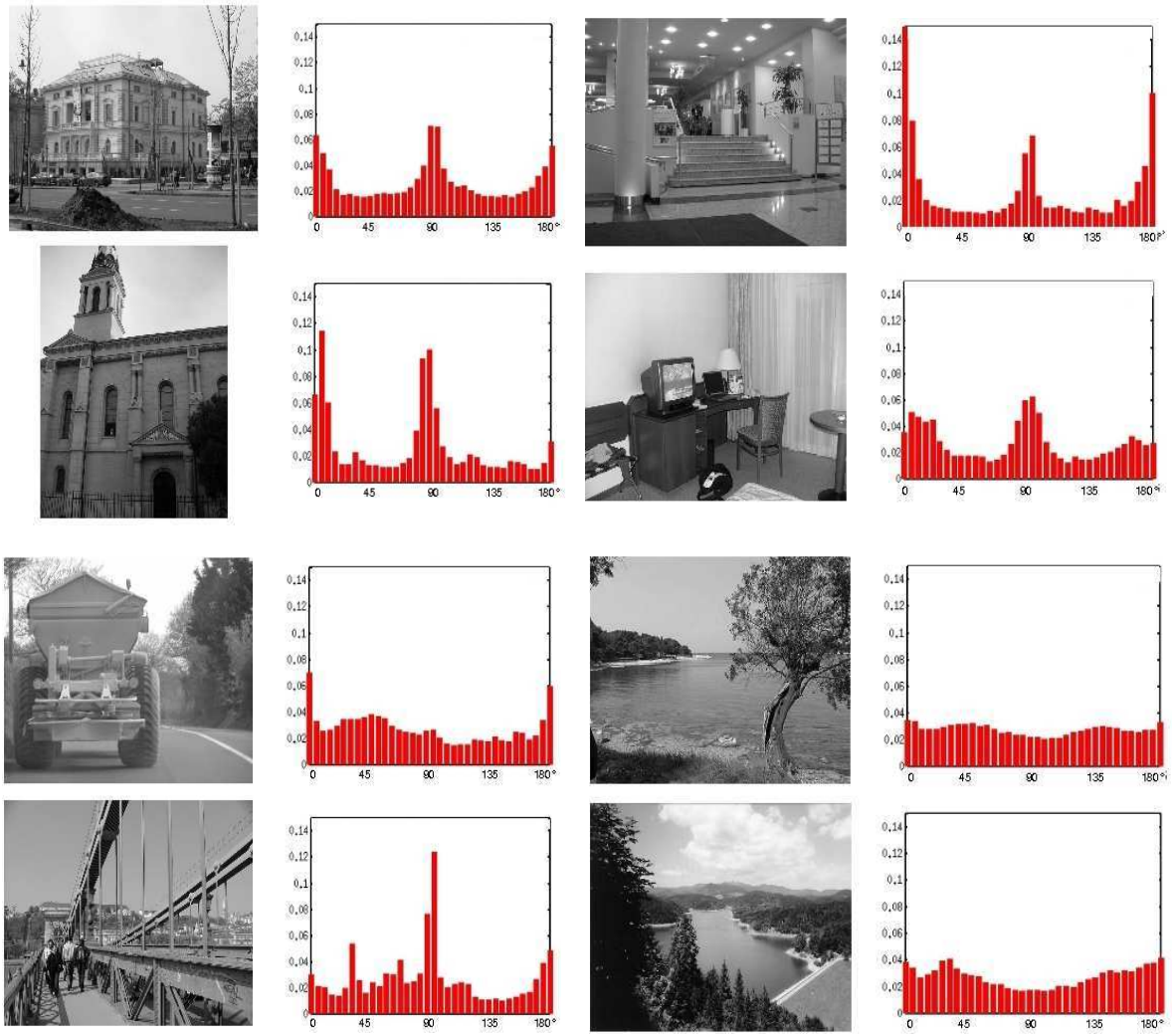


Figure 4.1 An example of edge orientation histogram distributions for *building*, *indoor,* and *outdoor non-building structures*, and *nature* images.

4.2 A hierarchical approach to semantic image classification

One of the popular approaches to image classification found in the literature, is based on a "divide-and-conquer" strategy: the classification problem is decomposed into a set of two-class classifications for each of which a particular feature set of high discriminability is identified. Essentially, the strategy implies a hierarchical approach, whereby the images are categorised in a multi-stage fashion, starting with a coarse classification into broad and rather abstract categories at the highest level. Further down the classification tree, the classes are subdivided into finer and more specific, more refined subsets. In [87], Vailaya *et al* approach the

classification of vacation images using such a hierarchy of high-level classes as shown in Figure 4.2.

As the research work so far suggests, there is no single universal feature representation that is suitable to every task. Thus, the main advantage of the hierarchical approach to image classification lies in the fact that it allows the use of simple and relatively low-dimensional feature representations, which are most appropriate for a given stage in the classification process. While colour features and camera metadata may be the most suitable features for *indoor/outdoor* classification, texture features would better serve the subsequent classification of *outdoor* images into *city* and *landscape* images.

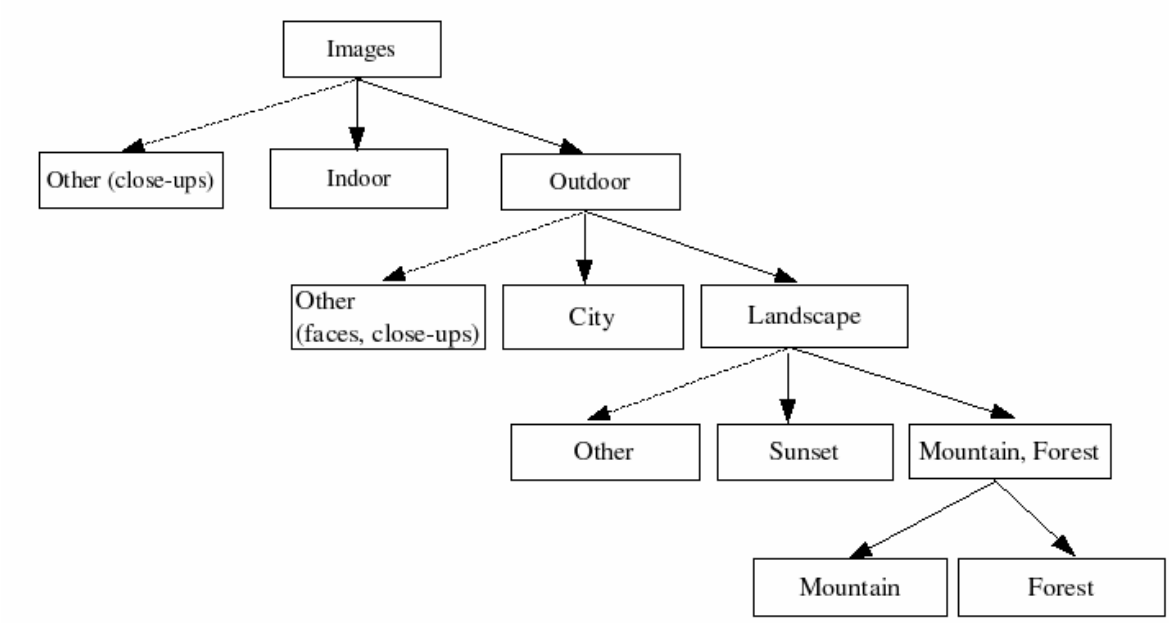


Figure 4.2 The Vailaya's image classification hierarchy [87].

We employ a modification of the above hierarchical classification method in our late fusion approach: at the first level, each image is classified on its own based on image content using a SVM classifier. At the second level, classification was performed using the combined evidence: initial *building/non-building* classification based on content-based features plus the *indoor/outdoor* classification based on camera metadata.

4.3 Indoor/outdoor classification

At the top level of the image classification hierarchy is the discrimination between indoor and outdoor images. The *indoor/outdoor* classification can either be a task in itself or a step towards a classification into more refined categories. The task has long been studied and colour and texture remain the low-level features of choice for the task.

In [87], Vailaya *et al* address the classification of vacation images to *indoor* and *outdoor* using low-level features: colour features in the form of 10x10 sub-block colour moments in LUV colour space. The approach exploits the fact that outdoor images tend to feature uniformity in spatial colour distributions, while indoor images, on the other hand, tend to feature more varied colour distributions as well as more uniform lighting. Their experiments, which used combined texture and colour moments, did not show any improvement in accuracy when compared to the colour moment only approach. Misclassification was reported for *outdoor* images which are either close-ups (so they feature uniform lighting across the image) or low contrast images.

The approach of Szummer and Picard [82] also relies on colour and texture: it combines colour histograms in the Ohta colour space with MSAR-based texture features (multiresolution simultaneous autoregressive model) for the whole image and for each sub-block of a 4x4 image tesselation.

In the work of Luo and Savakis [51], *indoor* vs. *outdoor* categorization of consumer photos is approached by using low-level visual features combined with some mid-level features so as to infer high-level information. A Bayesian network is used as a framework to integrate knowledge from low-level (the quantised colour histograms in the Ohta colour space and MSAR-based texture features) and mid-level features (such as sky and grass). The authors report an improvement over the classification results based on colour and texture alone, which they attribute to incorporation of mid-level feature such as sky and grass. However, the results of classification based on computed mid-level features alone are as good as those based on a combination of colour, texture and computed mid-level features, which suggests that any relevant information that may be encoded in either texture or colour information has already been captured by the grass and sky detectors.

A further advance in the *indoor/outdoor* classification is presented in [50] by Luo and Boutell. They utilise the camera metadata and low-level features, and report that "the metadata was so effective, it could be used in place of image content and still achieve high (that is 90%) classification accuracy". The most salient metadata cues were identified to be `flash` fired, `exposure time`, `aperture value`, `subject distance` and `focal length`.

In our work, *indoor/outdoor* classification is used as an intermediate step towards *building/non-building* classification and as a component to enhance the building detector. The approach is described in detail in the following sections.

4.4 Digital camera metadata

Metadata is "data about data"[16]. Metadata is information, usually highly structured, about documents, photos, books or other items, generated in order to facilitate organisation and access to the primary information. The three broad categories of metadata include administrative, structural and descriptive metadata [84]. Camera metadata refers to the information embedded in the EXIF header of the JPEG image created by the digital camera. It is a potential source of valuable information on the camera settings, capture conditions, etc., and, indirectly, the environmental lighting. Camera metadata can be used to infer the context of an image, such as location, date, time, brightness of the scene, and so on.

4.4.1 The EXIF standard

The Exchangeable Image File Format (EXIF) is one of the most commonly used image file formats and metadata standards at present [21]. The standard defines the format of both images and sound captured using digital still cameras and provides a standard specification for storing metadata pertaining to images and audio. The image metadata is stored in the image file header and is identified by unique tags. The metadata tags, as provided by the EXIF standard, include a large number of image-related metadata such as those pertaining to:

- the image data structure (e.g. width, height, numbers of bits per component, compression scheme, image orientation);

- the recording offset (e.g. image data location, number of rows per strip, bytes per compressed strip, bytes of JPEG data);
- image data characteristics (e.g. transfer function, white point chromacity, colour space transformation matrix coefficients);
- other/general tags (e.g. file change date and time, image title, image input equipment, copyright holder);
- picture taking conditions (e.g. exposure time, f-number, exposure program, ISO speed rating, shutter speed, aperture, brightness, exposure bias, subject distance, flash, lens focal length, exposure index, scene type);
- GPS (Global Positioning System) attribute information (e.g. GPS time, GPS date, speed of GPS receiver, GPS satellites used for measurement, altitude, latitude, longitude).

Digital cameras produced by different manufacturers, as well as different camera models may record different metadata. The quality and reliability of the metadata recorded may also vary for different brands of digital cameras. Some metadata, such as the location information, can be obtained using a separate GPS device and then added to the EXIF header in post-processing by matching the image timestamps with those in the GPS device's log. This is how the location information is obtained for images in the MediAssist collection which we use in our experiments. An example of EXIF header content is shown in Figure 4.3.

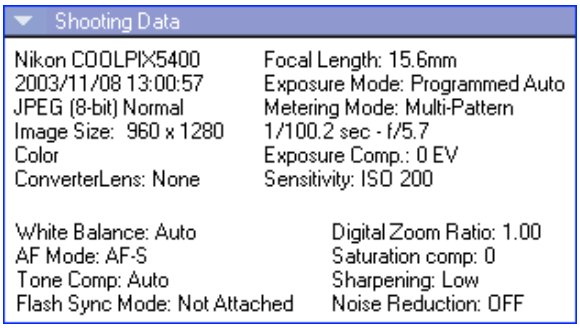


Figure 4.3 An example of the EXIF header content

4.4.2 Camera metadata potentially useful for *indoor/outdoor* classification

Among other tags, the EXIF standard for JPEG images [21], specifies a large number of tags related to capture conditions, camera settings, etc. that can be included in the EXIF header of

the JPEG image. Of these, there are 27 tags that are related to image capture conditions, such as: `ExposureTime`, `ExposureBiasValue`, `FNumber`, `ShutterSpeedValue`, `ApertureValue`, `BrightnessValue`, `FocalLength`, `SubjectDistance`, `SceneType`, `Flash`, etc. This kind of information could be useful for gaining knowledge about the image context. It may complement the content-based information, and thus help distinguish between images belonging to different sematic classes. Using camera metadata provides a way of introducing contextual information to an image classification task. For example, an effective approach to *indoor/outdoor* classification could be based on environmental light levels. We consider the following metadata, divided into original and derived, as being potentially of interest, and use them either directly or indirectly in our experiments. Original metatadata is raw metadata as contained in the EXIF header, whereas derived metadata is calculated/derived using two or more original metadata values.

Original EXIF metadata:

- `BrightnessValue` *(Bv)* – indicates the scene luminance or brightness. Larger values of *Bv* indicate greater scene brightness. Not all commercial camera models record the value of scene brightness. Brightness values in our dataset fall within the following range: [-6.04, 12.10].

- `ShutterSpeed` (`ExposureTime`) – is the length of time the shutter is kept open during the photo capture. The standard sequence of shutter speed values is as follows:

  8 4 2 1 ½ ¼ 1/8 1/15 1/30 1/60 1/125 1/250 1/500 1/1000 1/2000 ...

- `ApertureValue` – is the size of the lens opening. The size of the aperture is measured in *f-stops* where the *f-stop* value is inversely proportional to the aperture size (the ratio of the lens' focal length to the diameter of the lens diaphragm opening). A smaller aperture value corresponds to more light entering the camera. The standard sequence of values (*f* is the focal length) is as follows:

  f/1.0 *f*/1.4 *f*/2.0 *f*/2.8 *f*/4 *f*/5.6 *f*/8 *f*/11 *f*/16 *f*/22 *f*/32 *f*/45 *f*/64 ...

- `ISOSpeedRatings` – a standard value used as an indication of the light sensitivity of a film or electronic sensor. The higher the value of ISO speed, the more sensitive to light it is.

- `Flash` – indicates whether an auxiliary light was used during the capture to supplement the available lighting, be it natural (in an outdoor scene) or artificial (usually indoors). The use of a flash usually results in better exposure and better colour, as well as improved sharpness of the photo. The EXIF `FlashValueCode` values in the dataset are {0, 1, 9, 16, 24, 25, 31, 73, 89}. The odd values indicate that flash was fired, while the even value indicate that flash was not used. The other bits indicate the status of the returned light.

- `Timestamp` – provides time and date of the picture capture. Using the original image timestamps along with the GPS location information, which is recorded separately, we derive the `TimeOfTheDay` tag (i.e. dawn, day, dusk, or night).

- `FocalLength` – indicates the focal length of the lens or a lens' angle of view. In order to make the values comparable over the entire collection, where images were taken by different camera models, the value is normalised with respect to focal length of a 35mm film camera. The values in the dataset fall within the range: [1000, 250 000].

- `SubjectDistance` – this is a rough estimation of the main subject distance from the camera, quantised to four ranges and expressed as values: {0=`unknown`, 1= `macro view`, 2=`close view`, 3=`distant view`}.

Derived metadata:

- `Exposure Value` *(Ev)* – Camera exposure determines the amount of light that falls on the image sensor. For a given ISO Speed, the exposure is controlled by the combination of shutter speed and lens aperture. A given exposure value defines all combinations of the lens aperture and shutter speed that result in the same exposure. The effect of a varying shutter speed is shown in Figure 4.4. A larger value of *Ev* denotes less exposure. Larger exposure values are appropriate for photography in more brightly lit environments, or for higher film speeds. The following formula is used to calculate the exposure value from shutter speed, lens aperture and ISO Speed values [61]:

$$Ev = \log_2 \left| \frac{ApertureValue^2}{ShutterSpeedValue} * \frac{100}{ISOSpeedRatings} \right| \qquad (5)$$

When the ISO Speed is unknown, the following alternative formula is used (essentially assuming an *ISOSpeedRatings* of 100):

$$Ev = \log_2 \left| \frac{ApertureValue^2}{ShutterSpeedValue} \right| \quad (6)$$

The Exposure Value is a function of both the environmental light as well as any artificial light produced by the camera flash. It is calculated as an alternative measure of environmental light levels and to act as a replacement for missing Brightness Value that many cameras do not record. Figure 4.5 illustrates the relationship between the Brightness Value and Exposure Values. Exposure Values in our the dataset fall into the following range: [0.97, 17.29].

Figure 4.4 The effect of a varying shutter speed on night photography (captions indicate the number of seconds the shutter was kept open) [77].

Using the Exposure Value calculated as described above and assuming an ISO Speed of 100 (which corresponds to Speed Value, *Sv,* of 5), we calculate the corresponding Brightness Value for photos created by cameras that do not record the *Bv*:

$$Bv = Ev - 5 \quad (7)$$

As Figure 4.5 shows, based on the subset of image data for which *Bv* was recorded by the camera, this relationship can indeed be reasonably well approximated by the straight line, *Ev=Bv+5.*

- `TimeOfTheDay` – this value is obtained using standard astronomical algorithms based on the time and location [60]. The position of the sun in the sky is calculated at a certain time and place: if the sun is above the horizon it is daytime, if it falls below the horizon it is twilight (dawn or dusk), and even further below the horizon it is a night time [66]. The following labels are derived in that way: dawn. daylight, dusk, and night.

- `Photo's GPS location` – represents the Global Positioning System coordinates of the location at which the photo was taken. GPS is a system of 24 satellites that orbit 11,000 miles above the earth, and only 3 or 4 of those are needed to facilitate navigation using a GPS receiver. Due to the fact that radio signals, used for communication between the satellites and the receiver, cannot reach deeply into solid objects, such as buildings, it does not work well indoors. GPS enabled cameras are still confined to the high-end of the still camera market and only a handful of cameras on the market today, such as some of Nikon's models, support direct encoding of GPS information into the EXIF. As none of the camera models used in the collection process for the MediAssist image collection had an integrated GPS navigation, a separate GPS device (a Garmin Geko) was used during the photo capture. In post-processing, which involved matching the GPS logs and timestamps with those from the camera, the approximate GPS coordinates for each photo were obtained.

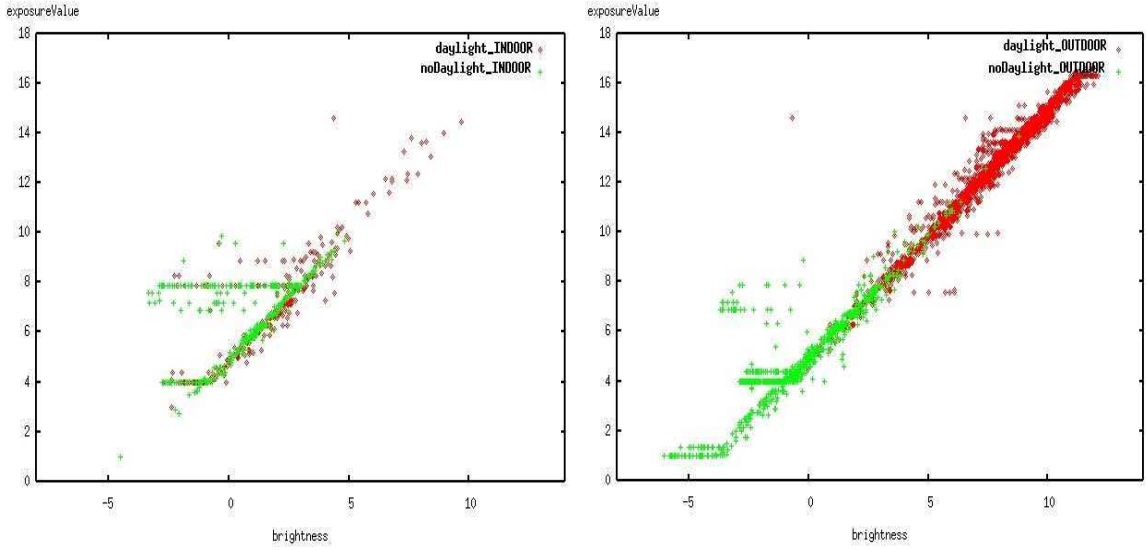


Figure 4.5 Relationship between the calculated exposure values and recorded brightness value for: a) indoor, and b) outdoor images.

4.4.3 Metadata discriminatory power for *indoor/outdoor* classification

A feature is considered to have good discriminative power for a given classification task if its inter-class variance is large while its intra-class variance is small. A small feature variance for a certain class implies a small extension of the class cluster in the corresponding direction in the feature space (i.e. the compactness of the cluster in the given direction).

In order to establish the usefulness of a particular metadata value for the task of *indoor/outdoor* classification, we examine the distribution, over the entire dataset, of different metadata values for the two classes of interest. We plot the distributions of the following features for *indoor* and *outdoor* classes, so as to empirically determine their discriminative power: `BrightnessValue`, `ExposureValue`, `Flash`, `FocalLength`, `TimeOfTheDay` and `SubjectDistance`. The aim is to identify the most informative ones and look for features that maximise the class separability. In doing so, we distinguish between images taken under *daylight* and under *no-daylight* (i.e. taken at dawn, dusk, night). All metadata values are linearly normalised to the [0,1] range. The results of empirical evaluation of distributions of metadata values are as follows:

Brightness Value

A distribution of the scene brightness values for *indoor* and *outdoor* images using different numbers of bins (i.e. different bin sizes or widths) is shown in Figure 4.6. For this purpose, the entire dataset is considered irrespective of the *daylight* status. These plots suggests that the brightness value is useful for the task, although there exists a large overlap such that nearly a half of the outdoor class distribution overlaps with the indoor distribution. In part, this may be due to the fact that no distinction was made between images with different daylight status (daylight vs. no-daylight photos). A long, slow-dropping tail on the left side of the outdoor distribution probably includes images taken at dawn or dusk. The small peak on the right side of the indoor distribution is likely to have been caused by the indoor images with reasonably high light levels combined with the use of flash. Lastly, it is possible to see that, although outdoor photos are spread across the entire range of brightness levels, in general, outdoor photos are characterised by higher scene brightness levels.

Exposure Value

A distribution of exposure values for the two classes is shown in Figure 4.7. A visual comparison of the distributions of brightness and exposure values indicates that a better class separability is exhibited by exposure (which is especially evident in the case of the 20-bin distribution).

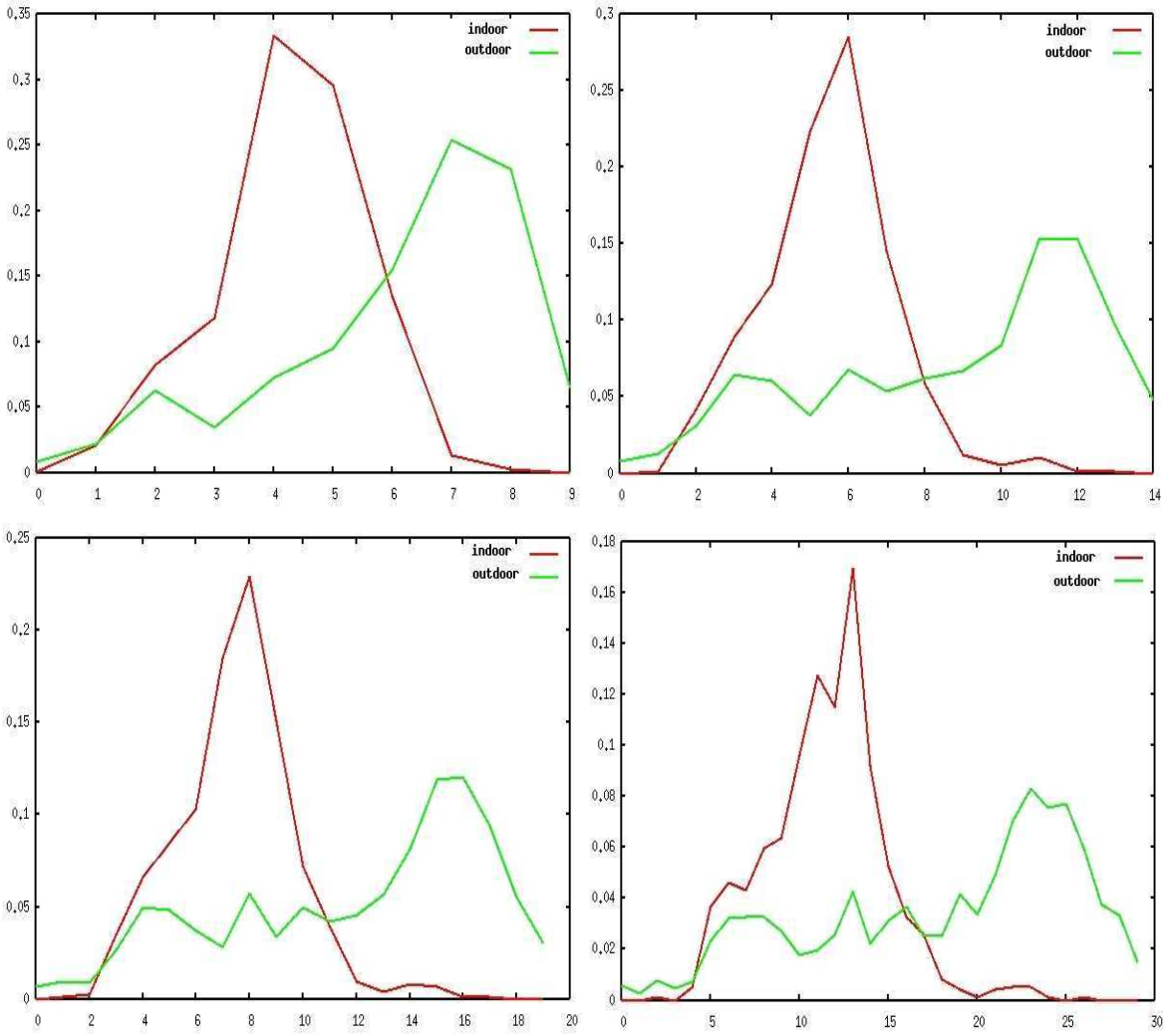


Figure 4.6 Distribution of brightness values of *indoor* and *outdoor* images: (a) 10 bins, (b) 15 bins, (c) 20 bins and (d) 30 bins.

Judging by the exposure value distribution, even a relatively simple threshold-based classification could yield reasonable classification accuracy (at the 6th or 7th bin). The relationship between the camera recorded brightness value and calculated exposure values for *indoor* and *outdoor* images under daylight and no-daylight (dawn, dusk and night) is illustrated in Figure 4.8. It can be observed that, in the case of *outdoor* photos taken under *daylight,* the relationship between the *Ev* and *Bv* values can be reasonably well approximated by the straight line, *Ev=Bv+5.* On the other hand, a significant number of outliers occur in both *indoor* and *outdoor* photos, taken at *no-daylight* time. In *outdoor* photos, both, *Ev* and *Bv* are spread over a larger interval.

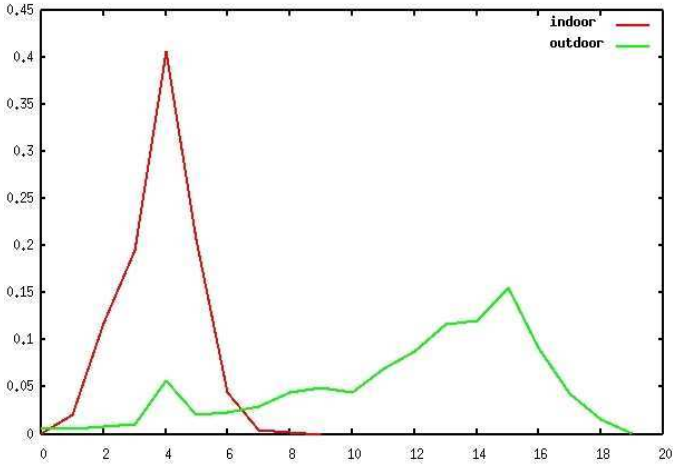


Figure 4.7 Distribution of exposure values of *indoor* and *outdoor* images using 20 bins.

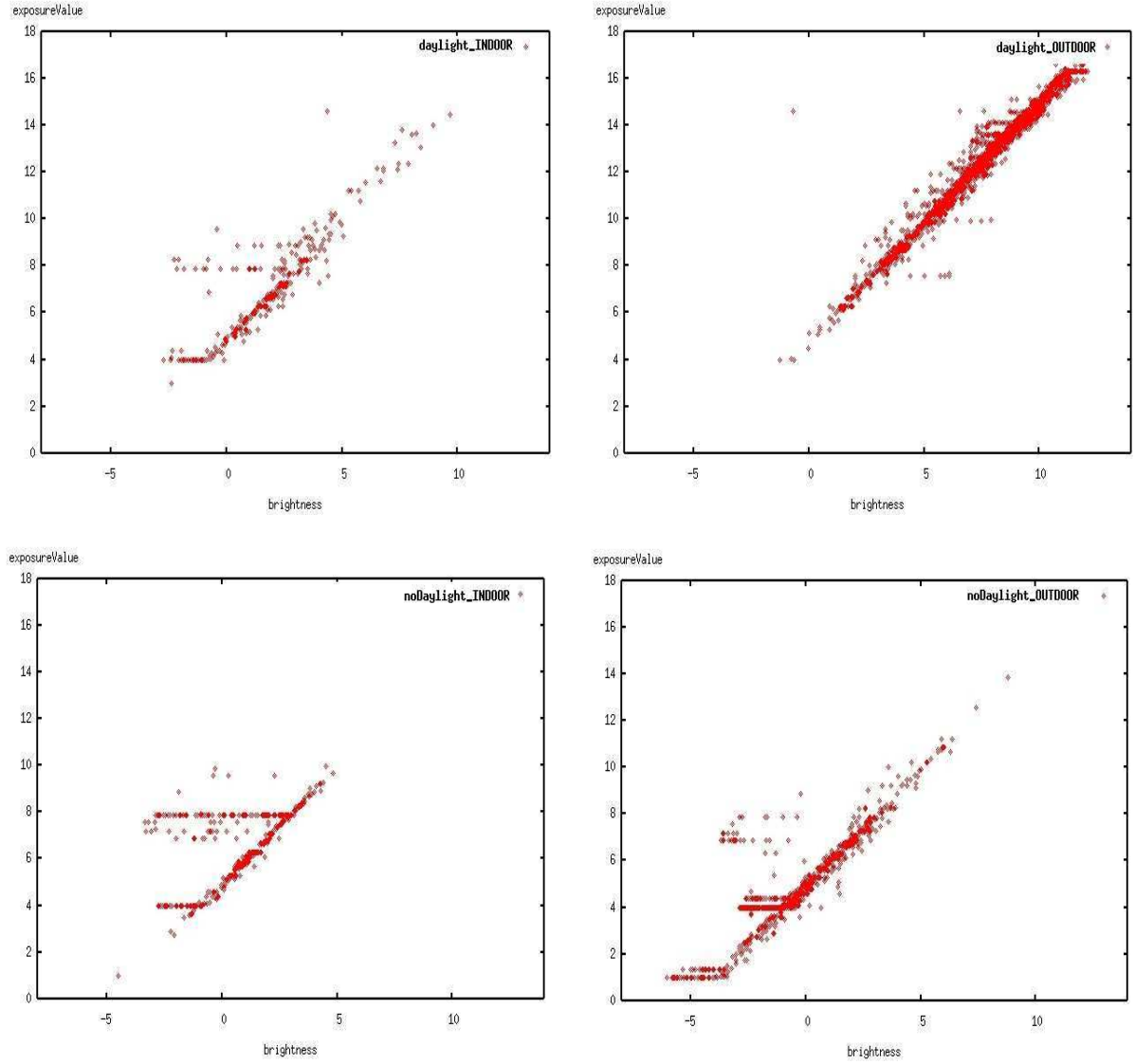


Figure 4.8 Recorded brightness and calculated exposure values of *indoor* and *outdoor* images under *daylight* and *no-daylight* (dusk, dawn, night): (a) *daylight indoor*, (b) *daylight outdoor*, (c) *no-daylight indoor,* and (d) *no-daylight outdoor*.

Flash Used

Overall, the likelihood of using a flash in an *indoor* image is much higher than using it *outdoor* as can be seen in Figure 4.9. The likelihood of using and not using a flash in an *indoor* image in our dataset is nearly even: the flash is used in some 53% of all indoor images. In contrast, a flash was used in less than 10 % of all *outdoor* images. In this respect, our dataset significantly differs from that used in [7], where a flash was used in approx 90% of *indoor* photos and in 19% of *outdoor* photos, and as such was a highly discriminative cue for *indoor/outdoor* classification. This difference in flash use in our *indoor* photos is quite significant. This may be user-dependent as some users may be aware of the concept of "fill-in-flash" and others not.

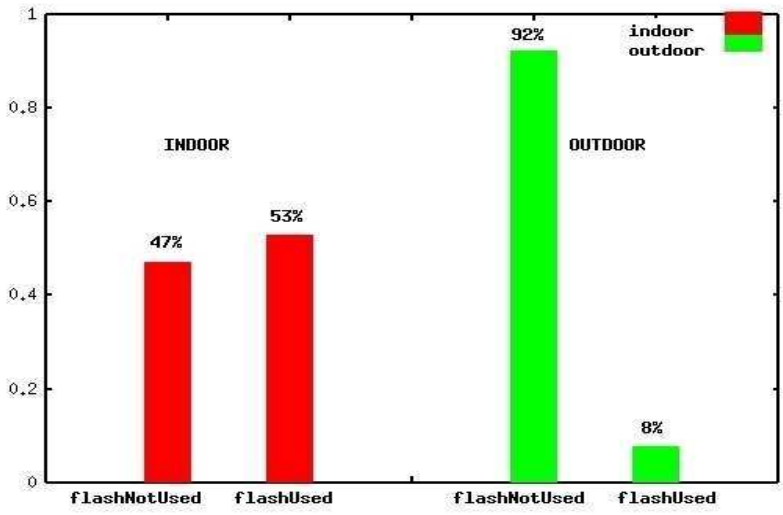


Figure 4.9 Distribution of flash value of *indoor* and *outdoor* images.

TimeOfTheDay (or Daylight Status)

The distribution of *indoor* and *outdoor* images with respect to the time of the day the photo was captured, is shown in Figure 4.10. As we can see, by far the largest proportion of *outdoor* photos, nearly 81% of them, were taken during the day time, while only 12% were captured at night time. On the other hand, there is not such a huge difference among the *indoor* photos as the number of those captured at day and night time stand at 43% and 55% respectively. Therefore, if we know a photo from our collection was captured at night time it is far more likely that the photo in question is an *indoor* photo. Likewise, a photo captured at daytime is likelier to be an *outdoor* photo. Overall, very few photos were captured at dawn: none of the *indoor* and only 0.07% of the *outdoor* photos. The number of photos captured at dusk were slightly higher, and those are mostly outdoors: 2.5% of the *indoor* and 12.1% of the *outdoor*

photos.

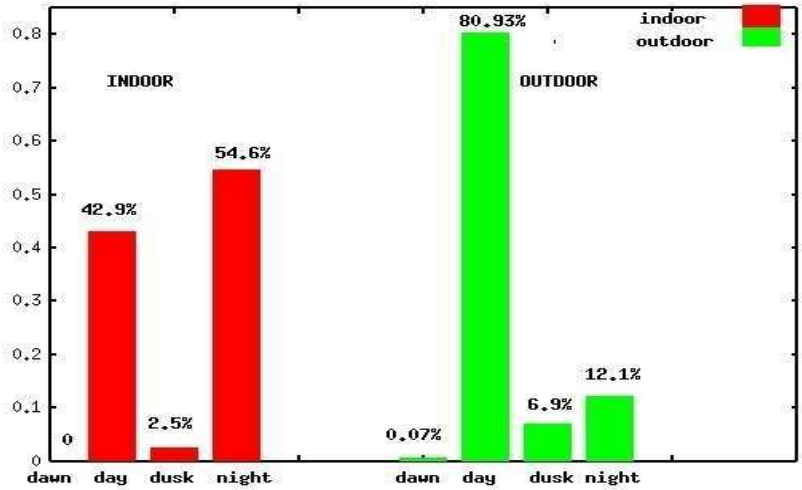


Figure 4.10 Distribution of TimeOfTheDay value of *indoor* and *outdoor* images.

Focal Length

The focal length distributions for the two classes are shown in Figure 4.11. While we can observe differences in the distributions of the two classes of interest, the class separability appears to be weak. The focal length distributions closely match for both classes for low values of focal lengths. Also, the largest number of photos have a focal length which falls in that interval[3]. The differences in distribution are more significant for photos with larger focal lengths.

Subject Distance Code

As the subject distance distribution shown in Figure 4.12 illustrates, the majority of our camera models do not record the subject distance code (i.e. the rough estimation of subject distance). For the photos captured by the cameras that do, we see that more *indoor* images are shot at a close distance range to the camera, i.e. `macro view`, while the opposite is true for the longest subject distance: even greater majority of photos with that subject distance are *outdoor* photos. A similar proportion of indoor and outdoor photos were taken at the medium range (i.e. `close view`). While this indicates that Subject Distance Code is a potentially useful cue, it is unfortunate that a large proportion of photos in our collection lack it. Figure 4.13 (a zoomed-in version of Figure 4.12) shows the subject distance range distributions for the known subject ranges: the proportion of photos captured at `close view` (or at medium distance) is nearly

3 Given a large range and distribution of focal length values, using log values or variable bin sizes may be more appropriate.

even for both classes.

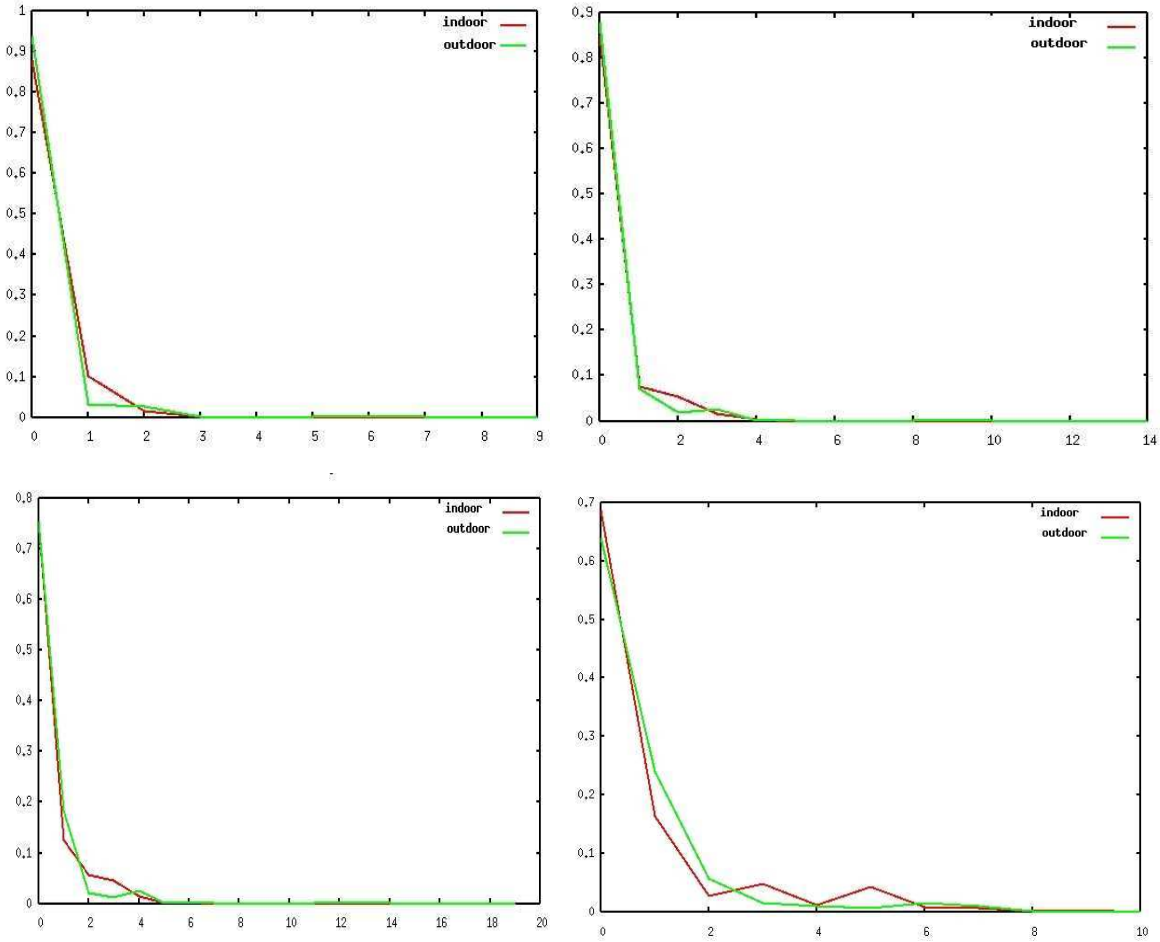


Figure 4.11 Distribution of focal length of *indoor* and *outdoor* images: (a) 10 bins, (b) 15 bins, (c) 20 bins, and (d) 30 bins (zoomed-in version).

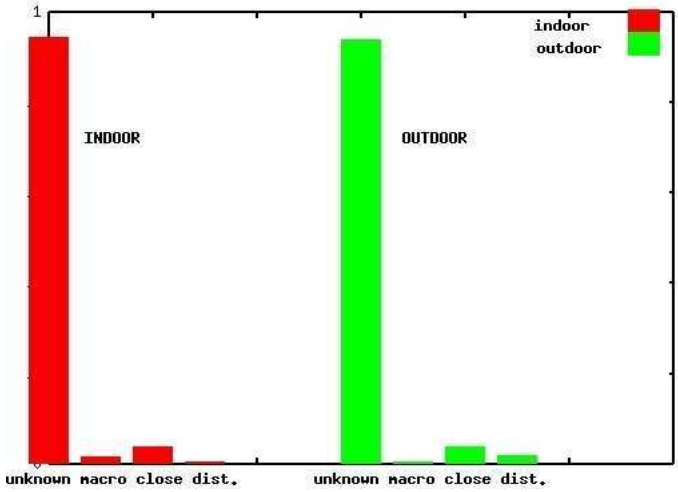


Figure 4.12 Subject distance range distribution of *indoor* and *outdoor* images (`unknown`, `macro view`, `close view`, and `distant view`).

Unsurprisingly, the photos captured as distant views were overwhelmingly *outdoor* photos, while most of the close-ups (i.e. `macro views`) were *indoor* photos. Unfortunately, for the largest majority of our photos the subject distance range value is "unknown".

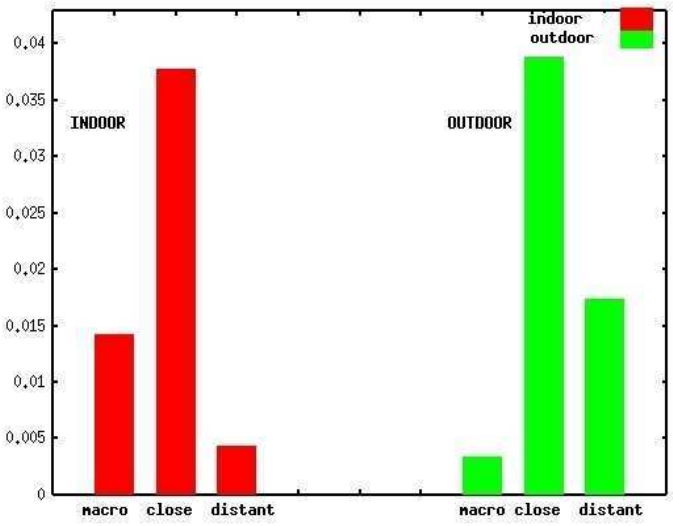


Figure 4.13 Subject distance range distribution of *indoor* and *outdoor* images for known values (`macro view`, `close view`, and `distant view` only).

Summary of metadata evaluation

The empirical evaluation of the metadata subset presented thus far suggests that, to a varying degree, all of the examined camera metadata could assist in discriminating between the classes of interest: the *indoor* and the *outdoor* images. Based on the distributions, we can conclude that the most discriminative of the metadata features are the exposure value, brightness and, to a lesser degree than expected, the flash used value. Although the evaluation indicates that the subject distance code is potentially a valuable cue for the task, as can be seen in Figure 4.13, it is unfortunate that its value is known only for less than 10% of the photos in our collection. Hence, the exploitability of this feature is necessarily limited.

A metadata feature such as the TimeOfTheDay is useful for separating the images taken at daytime (for which we assume light-level-associated features can be used to discriminate between the classes) from those taken at other times of the day (such as dawn, night or dusk). The focal length distributions closely match for both classes for low values of focal lengths. Coincidently, focal lengths associated with the largest number of photos fall in the same interval.

4.5 Proposed approaches to fusing camera metadata with low-level features

Our objective is defined as follows: given an arbitrary digital photo, determine whether there is a large or dominant building present in the photo. The method we propose here is intended as an improvement of the method presented in Chapter 3. We intend to test the hypothesis that the fusion of camera metadata with the existing low-level visual features will improve the classification performance by means of its capability to also distinguish between *indoor* and *outdoor* photos. We propose to fuse a selected subset of camera metadata with the existing low-level visual features (i.e. edge orientation histogram-based), using the Support Vector Machine (SVM) as an inference engine so as to improve the accuracy of the existing method of building detection. The features fused are as shown in Figure 4.14.

We base the approach on the assumption that a reliable *indoor/outdoor* classification based on lighting levels (inferred from the camera metadata such as brightness levels, exposure value, flash used etc.) can only be performed for the photos captured during the daylight time. The main assumption underpinning the approach is that the natural lighting in *outdoor* photos is stronger than artificial lighting present in the scenes captured *indoor* and this is only valid during the daylight hours. Some researchers even go as far as asserting that *outdoor* photos captured at night time, due to the fact that such images lack depth, should be for practical purposes treated as *indoor* [66]. Thus, we restrict our experiments to *daylight* photos.

In our approach, each image is represented by features extracted from two different modalities -from the image visual content and the camera metadata. The low-level visual feature representation comprises a 24-dimensional feature vector, based on edge-orientation features extracted at three scales, on global and local level. The visual feature representation is described in more detail in Chapter 3. The camera metadata representation of each image comprises a 5-dimensional feature vector, generated using the following metadata values: brightness value, exposure value, flash used, focal length and subject distance. All metadata values are linearly normalised into the range [0,1], across all values of a given metadata tag in the entire dataset. We compare two fusion strategies: (i) late fusion which entails combining the initial decisions and (ii) early fusion of features (by concatenating feature vectors). In both cases, the SVM is used as an integration device. The details of the approaches are introduced

in the following two sections.

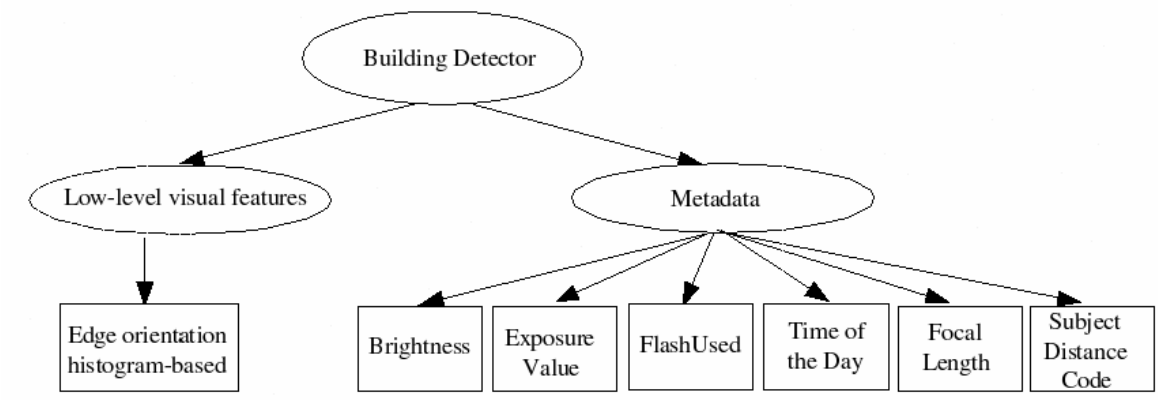


Figure 4.14 Low-level and metadata features used in fusion.

4.5.1 Early fusion

Early fusion, or feature level fusion, combines various features into a single feature representation. The features for fusion may originate in either different data sources or in the same raw data [80]. In our case, the features come from two data sources or two modalities: the image content and the camera metadata. The block diagram of our early fusion scheme is shown in Figure 4.15. Features are computed separately for each modality, each forming a feature vector. The two feature vectors are then combined into a single vector by concatenation. As a result, each photo is represented by a compact vector made of the multimodal data. The resultant feature representation, which is expected to be more informative than either of its components, is then classified using a single SVM classifier into either *building* or *non-building* class.

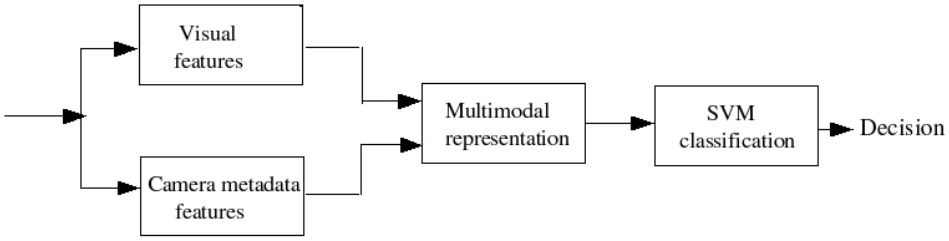


Figure 4.15 Block diagram for the early fusion scheme.

4.5.2 Late (or decision) fusion

Late fusion, or decision fusion, is a fusion of information from different sources at the decision level whereby the decisions reached by two or more inference engines are combined. A fusion of several initial classification results usually improves the quality of classification. Figure 4.16 shows the block diagram of our late fusion scheme.

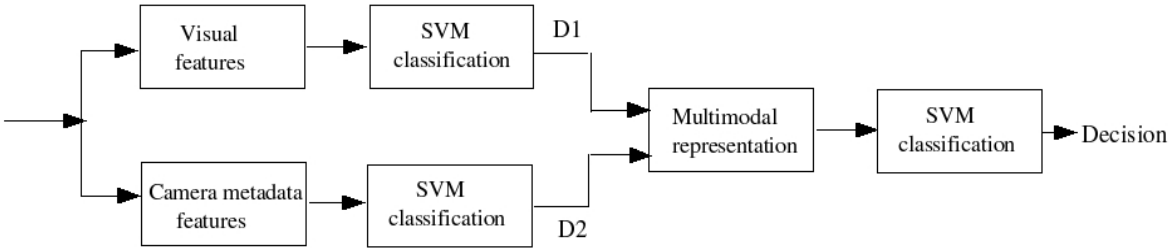


Figure 4.16 Block diagram for the late fusion scheme.

The two unimodal feature representations, the visual feature-based and the metadata-based representation, are classified separately using the SVM classifier. At this stage, the initial classification of an image into *building/non-building* and *indoor/outdoor* categories respectively, is performed. In order to make the SVM scores, D1 and D2, comparable, they are then linearly normalised to the [0,1] range. By combining the normalised scores into a new 2-dimensional feature vector, each of the decisions from the first stage of classification becomes a feature in the new feature vector used at the subsequent stage. A new classifier is trained and a final classification decision is obtained. In this way, the initial results of building detection are improved in refined classification when a fusion with the decision on the image's *indoor/outdoor* status, based on camera metadata, is performed.

4.6 Conclusions

We identify and examine different camera metadata for their potential to distinguish between *indoor* and *outdoor* images. Based on the empirical evaluation of the distributions of various camera metadata for indoor and outdoor images, we conclude that the exposure value and the brightness value are the most discriminative cues. Due to a different pattern of use of flash in our collection, the flash-used cue is not as highly discriminative cue as it is in [50]. However,

it is still useful. While the subject distance code also appears to be a very valuable cue, it is available only for a small portion of our photos. In the next chapter we evaluate our hypothesis and show the potential usefulness of fusing visual features with camera metadata for improving the performance of our building detector.

Chapter 5

Experimental Evaluation Of Metadata-Inclusive Implementation And Performance Comparisons

This chapter presents the results of the experimental evaluation of the metadata-inclusive implementation of the building detector approaches proposed in Chapter 4. The proposed multi-modal fusion schemes are intended to improve the classification accuracy by incorporating the features derived from camera metadata. In section 5.1 we describe the dataset used in the evaluation and also the annotation criteria. In section 5.2 we present the experimental results of both the outdoor detection and the building detection. The visual-features-based building detector is used as a baseline. The results are summarised and discussed in section 5.3.

5.1 Dataset

5.1.1 The MediAssist dataset

The dataset used in the experiments comprises 8000 genuine consumer photographs which were collected over the course of three years. This dataset is a subset of a larger image collection of over 17000 photos, 11000 of which have been annotated for various concepts such as indoor/outdoor, buildings, the presence and identity of people, etc. The images were gathered as a part of the MediAssist project, the aim of which was to develop tools for organising, managing and efficiently searching large personal photo collections [59]. A total of

16 people (amateur photographers) all members of our research group, contributed their photos to the collection, covering a total of 28 countries and 475 different locations. The photos in the collection span all seasons, and were taken at different times of the day and under different weather conditions. Some of the photos were taken from the air. All images in the collection contain metadata, and are timestamped and as well as GPS location stamped. The image database stores all images in three sizes: large (i.e. original), medium and small. The resolution at which images were processed for the work described here was mainly medium size (i.e. 640 x 480 pixels, 540x720, 720x540) both in portrait and landscape format (between 0.3-0.4 Mpixels). All processing was performed on grayscale JPEG images, and all images were presented to the algorithm in the correct orientation i.e. upright. Orientation correction was performed manually at the time of upload to the database.

Camera types

A number of different camera types were used in the process of image capture. Consequently, the number of metadata tags supported and their quality vary, as one would expect in a realistic large collection of consumer images. Table 5.1 lists the camera types used indicating which of the relevant metadata tags such as brightness value, shutter speed, aperture and flash were recorded.

Annotation

The groundtruth data was generated through a manual annotation by six members of our research group. The collection has been annotated for the presence of concepts such as large buildings, people, vehicles, animals and for indoor/outdoor status. Categorisation during annotation was crisp: ambiguous images and images that would have been more appropriately described using multiple labels were assigned to the closest category. We have not separated out close-ups, although we are aware that this has been done in the previous work [50]. According to these authors, "close-ups do not contain enough information for the algorithm (and sometimes even for the human) to decide the category". The issues overlooked at the time of annotation reflected on the performance results. The annotation taxonomy we use is shown in Figure 5.1 and the main categorical alternatives are discussed next.

***Indoor* vs *Outdoor*.** Strictly speaking, an *indoor* image is an image captured indoors, i.e. both the subject and the camera are indoors. Likewise, an *outdoor* image is an image of outdoor

scene captured by a camera located outdoors. However, there are many borderline cases in the collection such as those taken with both camera and subject located inside a roofed (with a non-transparent roof), but not walled (or only partially walled) spaces. In other instances, both camera and subject may be located in what in fact is an indoor space, completely enclosed, but both roof and walls may be fully transparent as is the case with a greenhouse or conservatory. So, a photo of an outdoor scene may be taken through a window of a house or a vehicle. While the photo is clearly an outdoor one, the location of camera at the time of capture will, to a varying degree, have bearing on the light levels falling on the light sensors, and thus will affect the values of camera metadata. Most often, the reference point for labelling was taken to be the position/location of the subject of the photograph.

Table 5.1. Metadata tags recorded by the different cameras models [66].

	Brightness	***Shutter Speed***	***Aperture***	***ISO Speed***	***Flash***
Canon PowerShot S40		√	√		√
Canon EOS DIGITAL REBEL		√	√	√	√
Kodak743		√	√		√
Kodak DX6490		√	√		√
Kodak DCS Pro 14n	√	√	√	√	
FujiFilm FinePix40i	√	√	√	√	√
FujiFilm FinePix S5000	√	√	√	√	√
FujiFilm FinePix F601 ZOOM	√	√	√	√	√
FujiFilm FinePix A203	√	√	√	√	√
Minolta DiMAGE X20	√	√	√	√	√
Nikon E775		√	√	√	√
Nikon D70		√	√		√
Olympus X400, D580Z, C460Z		√	√	√	√
Olympus X350, D575Z, C360Z		√	√	√	√
Olympus u20D, S400D, u400D		√	√	√	√
Sharp VE-CG30		√	√	√	√

***Building* vs *Non-Building*.** We defined a large *building* as a dominant object, a structure which is usually walled and enclosed, and built for permanent use. Some of the challenging images for annotation include images in which a significant degree of occlusion may be present due to the presence of other objects such as large non-building structures, vegetation, snow, etc. A photo of a large building completely covered with illuminated advertisement

panels, captured at night, is yet another example of building candidate to be annotated as *non-building structure.* Similarly, a photo of a building shot at night from a distance so that only illuminated windows are visible was labelled a *non-building structure*. Deciding on the point where an object ceases to be considered a large building and becomes a *non-building structure* was also a matter of subjective assessment. Generally, *non-building structures* included photos containing any human-made objects or their parts, such as monuments, furniture, vehicles, fences, road sides, swimming pools, light poles, traffic signs, electric lines and poles, ski paths, train tracks, etc. Close-ups of buildings depicting various parts of building such as wall texture, architectural details, ornaments, and so on, were also labelled as *non-building structure* images.

As shown in Table 5.2, out of the 8000 images, 2623 images (32.8%) contain a large *building* object, 3785 images (47.3%) contain some human-made structure, both outdoor and indoor, and 1481 (18.5%) are images of *nature*. The remaining 111 images (1.4%) are indoor images, mainly close-ups of people or pets that contain no human-made structures. Out of 5377 *non-building* images, 33.8% are indoor images, while 70% of all *non-building* images are *structure* images, i.e. contain some human-made structure other that buildings, such as bridges, monuments, ships and vehicles in outdoor images, or pieces of furniture, etc. in indoor images. Images of *nature* represent only 27.5% of all *non-building* images.

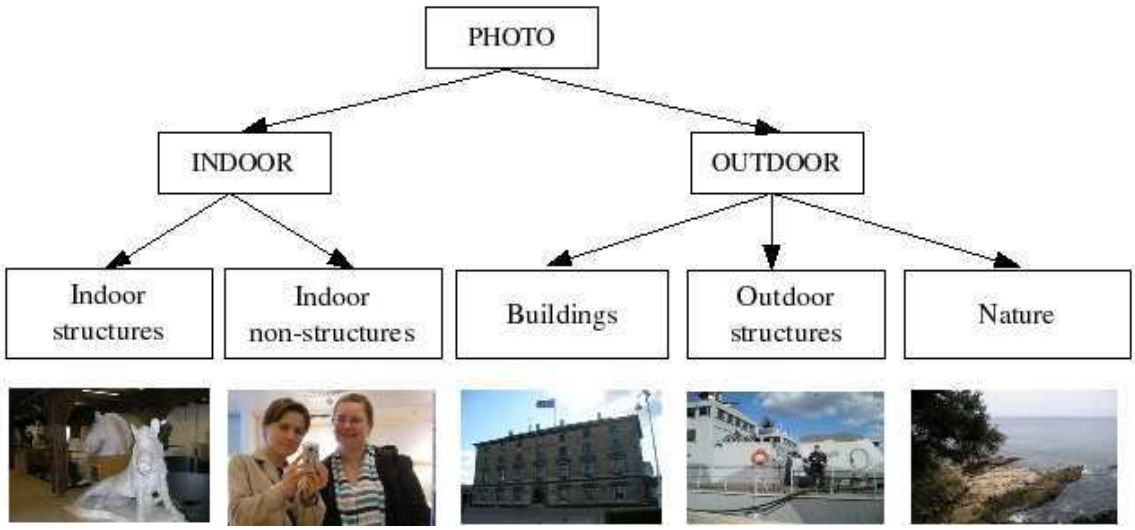


Figure 5.1 The annotation taxonomy.

As can be seen from Table 5.2, the proportion of outdoor to indoor images in the collection is 77% vs. 23% which is quite unbalanced. This simply reflects the fact that contributors to our

collection tended to take photos outdoors rather than indoor. The proportion of outdoor to indoor images in the *daylight*-section of the database is even higher, i.e. 86%. Again, this is due to the fact that more indoor photos were taken during night than daytime. The proportion of indoor and outdoor images in our collection that were taken at daytime is 43% and 81% respectively. While this may only be a feature of this particular photo collection, it may also indicate that people are actually more likely to take outdoor than indoor photos, and that majority of night time photos are taken indoors. Furthermore, we note that people feature more frequently in indoor photos due to a more limited choice of available or interesting subjects indoors. Example images from the MediAssist collection are shown in Figure 5.2.

Table 5.2 The MediAssist dataset structure.

OUTDOOR ***6179 (77.23%)***			***INDOOR*** ***1821 (22.77%)***	
Buildings	Non-building structures	Nature	Structures	Non-structures
2623 (2237[4])	2075 (1671)	1481 (1056)	1710 (732)	111 (51)

5.2 Experiments

To evaluate the metadata-enhanced approach to building detection, we conducted a number of experiments. We also compared the dataset used in Chapter 4 and the MediAssist dataset used in this set of experiments. The results verify the discrimination ability of metadata for outdoor detection and determine the upper limit of performance of a camera metadata-enhanced version of the building detector.

Each image was represented by a combination of visual cues and camera metadata-based features. The edge orientation based features (24) were extracted from grey scale JPEG images to represent image content, while the image context was represented by the following metadata features: brightness and exposure values, flash used, focal length and subject distance. All metadata values extracted from an EXIF header were linearly normalised into the [0,1] range so as to ensure that all the features, whether content-based or context-based, fall into the same range. Exposure values were calculated from shutter speed, aperture and ISO speed. The

4 Number of images taken under *daylight*

missing brightness values were imputed based on approximations using calculated exposure values [69] and further analysis was carried out as if these were observed or actual data.

Due to the storage space constraints and fact that only edge-based visual features are used, all images were processed as grey scale JPEGs. Thus the process of visual feature extraction entailed conversion from compressed JPEG to uncompressed PGM format before the edge orientation feature calculation. The majority of the photos processed were of the following resolution: 540x720 for portrait and 720x540 for landscape format. The average time taken for the extraction of visual features on a 1.5 GHz-Celeron processor running Linux RedHat 8.0 was 3.356 seconds per image.

In all experiments, the Support Vector Machine (SVM) was used as the inference engine. We experimented with different SVM kernels, which were available off the shelf. These included polynomial kernels of various degrees, and the radial basis function:

- Linear: $$K(\vec{x}_i, \vec{x}_j) = \vec{x}_i \cdot \vec{x}_j \quad (8)$$
- Polynomial: $$K(\vec{x}_i, \vec{x}_j) = [\vec{x}_i \cdot \vec{x}_j]^d \quad (9)$$
- Radial Basis Function: $$K(\vec{x}_i, \vec{x}_j) = \exp\left(\frac{-\left|\vec{x}_i - \vec{x}_j\right|^2}{\sigma^2}\right) \quad (10)$$

A kernel can be thought of as a similarity measure between the inputs [3]. This implies that, assuming a selection of suitable kernel, feature points representing objects of the same class should have high kernel value, whereas points representing different classes of objects should give a low kernel value. A kernel describes a mapping from the original or an input feature space into a higher-dimensional space. When the training data is not separable in the original feature space, it is mapped into a higher dimensional space and a separating hyperplane is defined there. In other words, a kernel function defines a new feature space in which the classification will take place.

Classification accuracy, precision and recall are used as performance measures. In all cases, a precision and recall break-even-point (BEP) during training serves as a criterion for selection of the SVM classifier. The dataset is divided into independent training and test sets and an effort is made to ensure that the learning examples are evenly distributed across different

Figure 5. 2 Example photographs from the MediAssist collection.

photographers' contributions. Approximately 1/3 of all *building* images are used as training examples along with as many *non-building* images, resulting in a training set of 1400 images. As the importance of a feature is monotonic in its absolute value [37], the larger valued features are more influential. In order to bring all the feature vectors onto the same scale and make them equally important, some sort of feature normalisation is required. We chose to linearly normalise the features.

5.2.1 Dataset comparison

To assess the difficulty of the MediAssist dataset we compare it with the dataset used in Chapter 4. The MediAssist dataset, being an unconstrained collection, is considered a far more challenging dataset. We compare the peformance of the SVM classifier trained on 200 images (i) on 1520 images as used in the evaluation in Chapter 4, (ii) on 4964 outdoor MediAassist images (daylight, outdoor only) and (iii) on 5747 MediAssist images (daylight, both outdoor and indoor). In the second set of experiments we use the entire image collection from Chapter 4 (1720 images) to train the SVM classifier and compare its performance on the MediAssist outdoor images only and on the entire MediAssist set. We use a linear kernel and select the classifier based on the recall precision break-even-point, i.e. the BEP criterion. As can be seen from Table 5.3, our classifier does not generalise well to consumer images, as there is a significant drop in all performance measures: accuracy, recall and precision rates on the MediAssist dataset, even when only outdoor images are considered. However, an increase in the training set size from 200 to 1720 images does affect the performance positively, although only slightly (an approximately 2% increase). Apparently, the initial training set size of 200 images, which performed very well on the Chapter 4 dataset (with recall and precision rates in mid-80-ies to early 90-ies respectively), is inadequate both in size and diversity to meet the requirements of the MediAssist dataset.

Table 5.3 Performance comparison of a building detector trained on different number of examples and on different datasets.

# examples	***Dataset***	***Accuracy***	***Recall***	***Precision***
200	Outdoor images	88.22	84.01	92.02
200	MediAssist (Outdoor)	70.89	68.45	68.45
	MediAssist (All)	66.02	68.45	54.42
1720	MediAssist (Outdoor)	73.35	70.23	70.54
	MediAssist (All)	68.25	70.23	57.57

We also observe that, irrespective of the training set size, an inclusion of indoor images in the test set further reduces the classification accuracy, and especially the classification precision. A large drop in precision indicates that many of the *indoor* images were actually misclassified as *buildings*. This indicates the need to enrich the feature representation with features that

would aid discrimination between *indoor* and *outdoor* photos. The recall rate, however, remains the same.

Another point well worth nothing is the fact that 71.8% of the 8000 MediAssist images were actually used in the experiments and only images taken during non-daylight time were removed from the dataset. As a comparison, in [50], which deals with the issue of natural vs. manmade classification, only 39% of the intial dataset was actually used in their experiments. All images whose scene content was considered such that the majority of it could not be unambiguosly put in a single class were removed, as well as all close-ups. The close-ups were left out on the account of not containing sufficient environmental information to determine the actual class. A further 45% of the images in [50] were considered either to be ambiguous or of low quality. As a result, a total of 61% of images were left out of the original dataset. In our collection, a significant number of photos contain large regions exhibiting characteristics of two different classes, and as such, could be better labelled with both. This degree of ambiguity inherent in the MediAssist dataset contributes to a further reduction in performance measures. In contrast, the dataset used in Chapter 4 was carefully selected so as to minimise ambiguity with respect to the class membership.

5.2.2 Visual features combined with the *indoor/outdoor* ground truth information

By utilising the groundtruth information for the *indoor/outdoor* categories, we determine the upper limit of performance of the building detector based on fused visual and metadata features. The SVM classifier is trained on 1400 images: 700 *building* (approximately 1/3 of all building images) and 700 *non-building* images. A linear kernel is used with the BEP criterion-based classifier selection. The comparison of performance on the entire MediAssist (MA) dataset and the outdoor images only is shown in Table 5.4.

Table 5.4 Comparison of building detector performances on outdoor and all images in the MediAssist dataset.

	Accuracy	***Recall***	***Precision***
MediAssist (Outdoor), lin, j=1.3	75.83	71.44	72.74
MediAssist (All), lin, j=1.3	69.68	71.44	55.57

It can be observed from Table 5.4 that both the accuracy and precison rates are significantly reduced by the inclusion of *indoor* images in the dataset, while the recall rate remains unaffected. This shows that the performance deterioration is caused by the misclassification of many *indoor* images as *buildings*, i.e. by mistaking the human-made structures, which are often present in *indoor* images, for *buildings*. Secondly, a considerable increase in the size and diversity of the training set (from 200 images in Table 5.3 to 1400 MediAssist images) has increased the recall rate on the outdoor MediAssist images by only 1.04% (from 68.45% to 71.44%), and slightly increased the precision (from 68.45% to 72.74%), while the classification accuracy increased to 75.83% from around 70.9%.

Another set of groundtruth experiments we conducted involves the early fusion by concatenation of visual features with the ground truth information for the *indoor/outdoor* class. A 25-dimensional feature vector is formed using 24 visual features, and the values 1 and 0 for outdoor and indoor respectively, as the 25th feature. The performance measures for fusion approaches presented in Table 5.5 represent the upper limit of the performance of the detector (i.e. given a perfect *indoor/outdoor* detector). It can be observed that, in general and up to a certain point, performance measures improve when more complex kerels are used. The best performing kernel is the polynomial kernel of degree 4.

Table 5.5 Comparison of the best performing approach using only visual features with the approaches based on visual features fused with groundtruth information for *indoor/outdoor* status.

Approach	***Accuracy***	***Recall***	***Precision***
Visual only, poly 3, j=1, BEP=73.3%	71.32	72.93	57.46
Fused, linear, j=1.25, BEP=78%	76.7	78.59	63.88
Fused, poly 2, j=1.17, BEP=79.3%	78.52	76.45	67.3
Fused, poly 3, j=1.09, BEP=79.14%	78.38	80.29	65.99
Fused, poly 4, j=1.05, BEP=79.2%	78.48	80.29	66.13
Fused, rbf, j=0.99, BEP=79.5%	78.11	79.83	65.69

The comparison of the SVM scores for visual feature-based and fusion approaches is presented in Figure 5.3. As expected, the most notable difference is the impact of feature fusion on the correct classification of *indoor* images as their SVM scores moved from the upper half-plane representing *building* class in (a) into the *non-building* half-plane in (b) and (c) i.e. the *indoor*

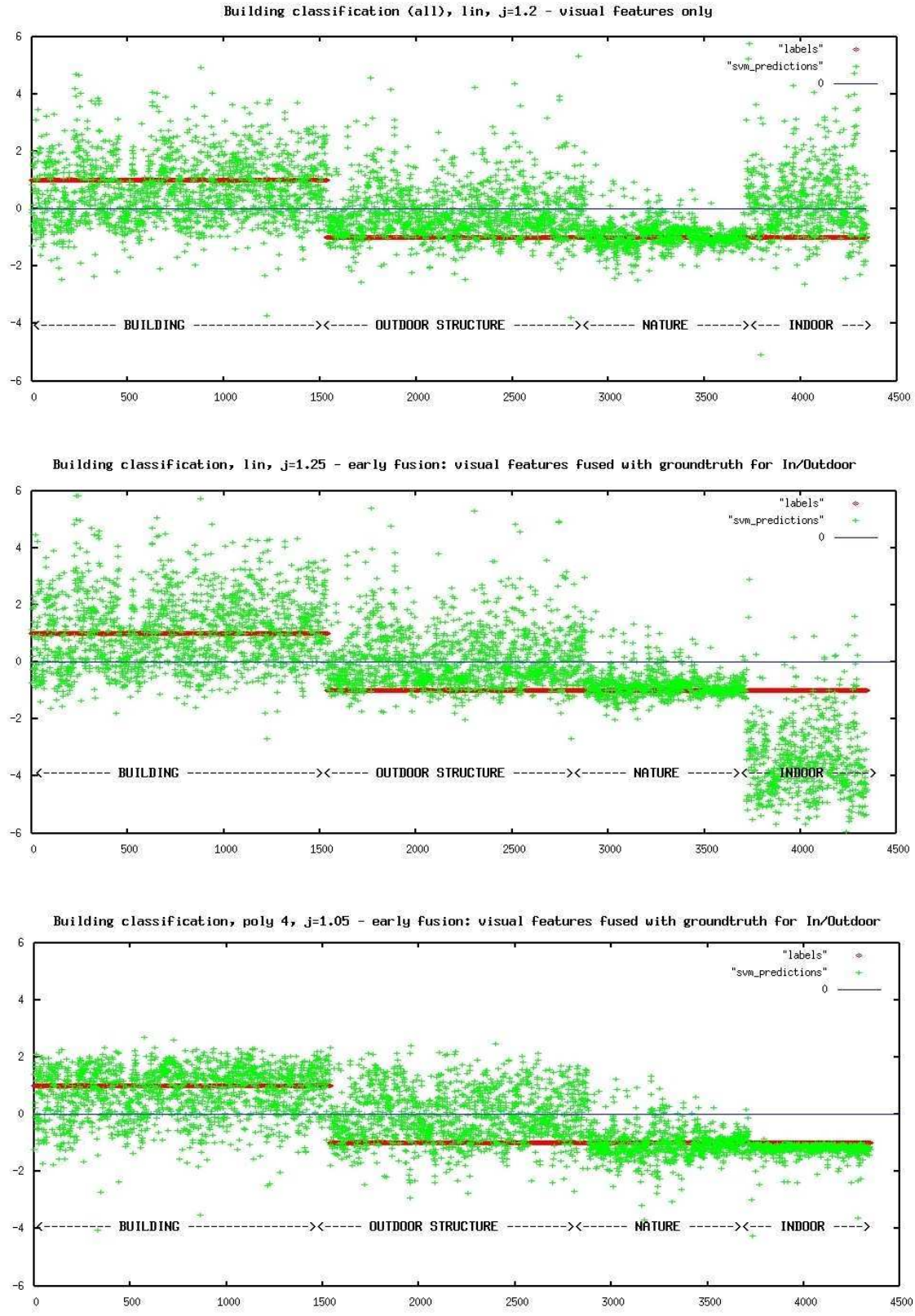


Figure 5.3 Comparison of SVM scores for (a) visual features only approach (linear, j=1.2) with (b) the early fusion approaches with groundtruth for *indoor/outdoor* class using linear, j=1.25, and (c) polynomial kernel of degree 4, j=1.21.

images' scores move towards the correct class label. A comparison of scores in (b) and (c) illustrates the effect of using a differrent kernel, as we can see the scores for the polynomial kernel of degree 4 are more clustered around the actual class label. *Nature* images were those with no human-made elements to them.

5.2.3 Indoor/outdoor classification based on camera metadata

Here we evaluate the ability of camera metadata on its own to discriminate outdoor from indoor scenes using 5747 *daylight* images (783 indoor, 4964 outdoor), utilising a different number of training examples and different combinations of metadata. In all instances, a linear kernel is used unless it proves to be too slow to train. In such cases, a polynomial kernel is chosen. The results of outdoor detection using different combinations of metadata features, such as brightness value (B), exposure value (E), flash used (F), focal length (L) and subject distance (SD), and utilising different numbers of training examples is shown Table 5.6.

Table 5.6 Outdoor detection using different number of metadata features and different number of training examples.

	# examples	***Accuracy***	***Recall***	***Precision***
B+E+F	200 (BEP: j=2.6), lin	91.64	96.07	94.48
	300 (BEP: j=1.5), lin	88.42	89.72	96.56
	784 (BEP: j=1.1), poly 2	85.75	85.63	98.79
B+E+F+L	200 (BEP: j=1.7), lin	89.65	98.54	90.5
	300 (BEP: j=2.1), lin	89.65	98.11	90.9
	784 (BEP: j=1.1), lin	86.28	86.2	98.75
B+E+F+L+D	200 (BEP: j=3), lin	91.45	96.09	94.27
	300 (BEP: j=1.3), lin	88.23	89.45	97
	784 (BEP: j=1.2), lin	86.9	86.88	98.76

A visual inspection of some of the images that are misclassified with high decision confidence shows that these images are actually ambiguous with respect to their *indoor/outdoor* status. Such examples include pictures captured through a window (i.e. camera indoors, subject outdoors), from an aircraft, or a vehicle, etc. Furthermore, we observe that the performance

using 784 training examples improves when the subject distance feature is included, even though only a small portion of images have a known subject distance. The distribution of the SVM outputs, for a classifier trained on 200 examples, is shown in Figure 5.4. The good class separability conforms well with the results of the empirical evaluation of the discriminatative ability of metadata cues for the *indoor/outdoor* classification presented in Chapter 4.

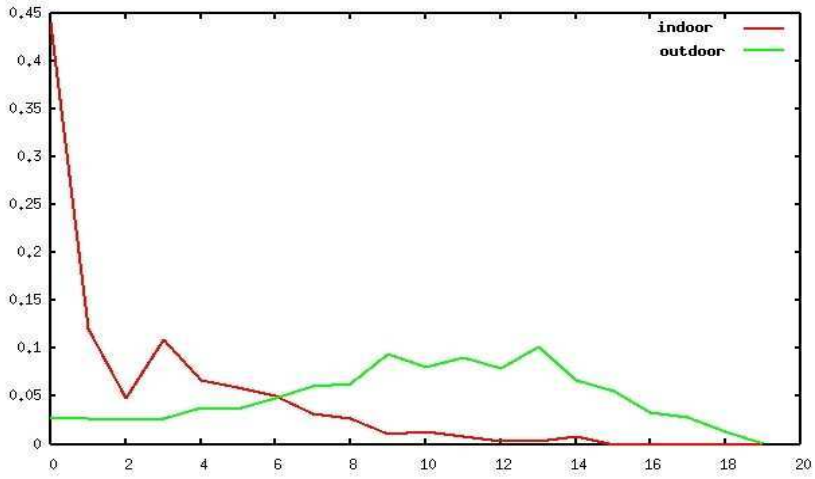


Figure 5.4 Distribution of SVM outputs for *indoor* and *outdoor* categories, based on the following metadata: brightness, exposure, flash, focal length and subject distance, using a linear kernel, j=3, with 200 training examples.

A conclusion can be drawn that the integration of metadata features with low-level visual features should improve the classification of building detection due to the ability of the camera metadata to discriminate between *indoor* and *outdoor* photographs. This is expected to reduce the misclassification of *indoor human-made structure* images into *buildings.*

5.2.4 Early fusion of visual features with the selected camera metadata

The early fusion approach entails fusion at the feature level. It relies on separate feature extraction from different modalities and their subsequent combination in the new feature space. In this work, visual features, based on edge-orientation histograms, are fused with five metadata-based features (brightness value, exposure value, flash, focal length and subject distance) by feature vector concatenation into a 29-dimensional vector. The results of the evaluation using 1400 training examples and different kerenels are presented in Table 5.7.

Table 5.7 Results of early fusion of visual features with selected camera metadata (BEFLD, using 1400 training examples).

	Accuracy	***Recall***	***Precision***
Visual only, poly 3, j=1, BEP=73.3%	71.32	72.15	57.46
Fused, linear, j=1.25, BEP=71%	71.16	72.74	57.27
Fused, poly 2, j=1.2, BEP=75%	73.69	75.08	60.29
Fused, poly 3, j=1.2, BEP=76%	74.33	75.67	61.08
Fused, poly 4, j=1.21, BEP=76%	74.36	75.67	61.11
Fused, rbf, j=1.1, BEP=76.6%	74.93	75.21	62.02

We observe that the performance of an early fusion approach using a linear kernel is comparable to the performance of the best performing visual features classifier (using a polynomial kernel of degree 3). All other early fusion-based classifiers outperform the visual features alone. As can be seen from Table 5.7, the usage of polynomial kernels improves all performance measures up to a point. After the polynomial kernel of degree 3, further increase in the degree and/or complexity of the kernel results only in a marginal improvement in performance.

Figure 5.5 illustrates the relationship between the correct class labels and the corresponding SVM scores for the visual features only approach, and the two early fusion approches using two different kernels. Firstly, we observe that, in all three plots, the *nature* images tightly cluster around the actual class labels (i.e. *non-building*), with quite a low number of images on the *building* side of the decision surface. This is as expected since our visual features are well-capable of discriminating between the images of purely natural scenes and images containing human-made structures.

Secondly, the introduction of new features does not significantly affect the classification scores of the *building* images, as can be seen comparing plot a) with plots b) and c). An exception to this is the first 200 *building* images on the very left, whose SVM scores appear to have moved deeper into the *non-building* side of the decision surface. This is possibly only a side effect of the displacement/adjustment of the decision surface caused by the introduction of the new features.

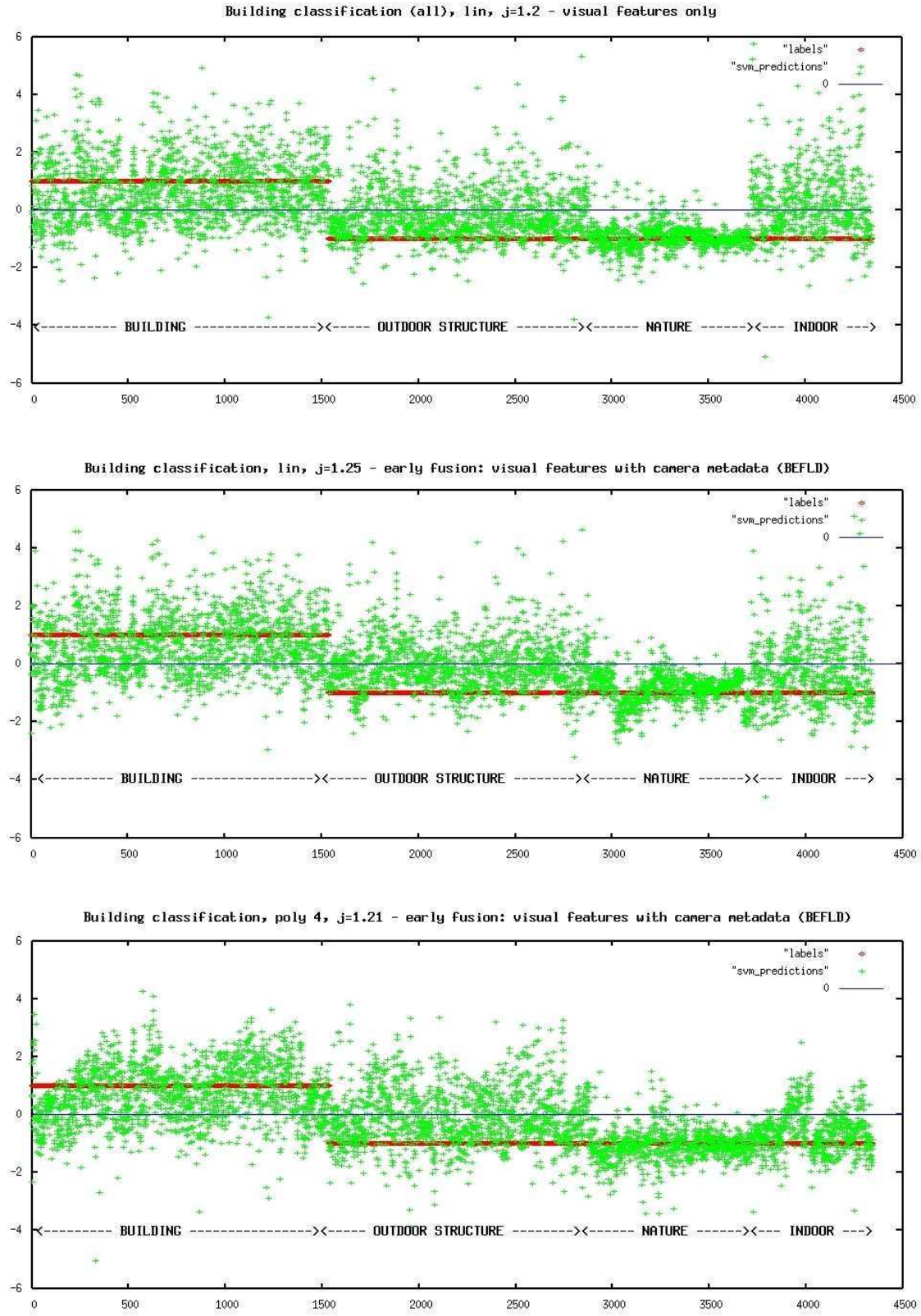


Figure 5.5 Comparison of SVM scores for (a) visual features only approach (lin, j=1.2) with (b) the early fusion approaches using linear, j=1.25, and (c) polynomial kernel of degree 4, j=1.21.

Thirdly, the SVM scores relating to outdoor *non-building structures* remain largely unaffected by the introduction of metadata features, as their scores remain scattered on both sides of the decision surface in all three plots. This is not surprising, since the new features do not bring any additional information that would affect their classification in either direction, and the existing visual features are not powerful enough to discriminate between the *buildings* and the other *human-made structures*. We note that the classification of outdoor *non-building structure* images in unconstrained photo collections remains inadequately resolved.

Finally, we observe that the *indoor* images are the most profoundly and positively affected by the introduction of the camera metadata features, since the majority of the SVM scores corresponding to *indoor* images move into the *non-building* half-plane in both b) and c) plots. However, there remains a number of *indoor* images that could not be correctly classified using the selected camera metadata features.

5.2.5 Late fusion of visual features with the selected camera metadata

The late fusion approach involves two stages, each of which includes a learning step. In our late fusion scheme, two classifiers are trained on visual and metadata features respectively. The visual features-based classification produces an initial *building/non-building* score. The metadata-based classifier generates a score on the *outdoor/indoor* status of the image. The best performing classifier is selected for each task, i.e. the initial *building* detection (using a polynomial kernel of degree 2, j=1, 1400 training examples) and the *outdoor* detection (BEFLD, using linear kernel, j=3, 200 training examples). The SVM scores, representing the classifier decisions at the first stage, are linearly normalised to the [0,1] range and then combined into a 2-D feature vector (i.e. a semantic representation of an image) associated with each image. These combined scores yield the final score. A new subset of 200 images is used as training examples for the SVM metaclassifier in the second stage. Results of the final classification are presented in Table 5.8.

Table 5.8 Results of late fusion of *building* detection decision, based on visual features, with the *indoor/outdoor* detection based on camera metadata (for 2 approaches that result in the same accuracy, one with greater recall-precision product is considered to be a better performing one).

	Accuracy	***Recall***	***Precision***
Visual only, poly 3, j=1, BEP=73.3%	71.32	72.93	55.46
Fused, linear, j=0.96, BEP=73%	72.87	70.63	59.11
Fused, poly 2, j=1, BEP=74%	73.55	70.49	60.12
Fused, poly 3, j=1.01, BEP=74.5%	74.1	71.61	60.74
Fused, poly 4, j=1.075, BEP=75%	74.1	71.26	60.81
Fused, rbf, j=0.9, BEP=76.5%	74	71.47	60.63

Again, we observe that the performance of the late fusion approach, using a linear kernel, is comparable with that of the best performing visual-features-only approach. All other late fusion-based classifiers outperform the visual features alone approach in accuracy and in precision. However, the recall rate of the late fusion-based classifiers is consistently inferior, and even the best fusion recall rate is still below the recall rate based on visual features on their own.

5.2.6 Result comparison

The comparison of the best performing classifiers for visual features, early fusion and late fusion approaches are presented in Table 5.9. These results show that the early fusion approach, using RBF kernel and cost factor of j=1.1, with BEP=76.6%, performs best overall on our dataset. As can be seen from Table 5.9, the early fusion approach improves all performance measures. The classification accurary is increased by 3.59%, while the recall and precision rates are increased by 3.05% and 3.53% respectively.

Table 5.9 Comparison of the best performing classifiers for each approach.

	Accuracy	***Recall***	***Precision***
Visual only, poly 3, j=1, BEP=73.3%	71.32	72.93	57.46
Early fusion, RBF, j=1.1, BEP=76.6%	74.93	75.21	61.11
Late fusion, poly 3, j=1.01, BEP=74.5%	74.1	71.61	60.74

The comparison of the SVM output distributions for the three approaches is shown in Figure 5.6. Although the class separation is weak in all three cases, it does appear to be the best in the case of the early fusion approach, which corresponds to the evaluation results.

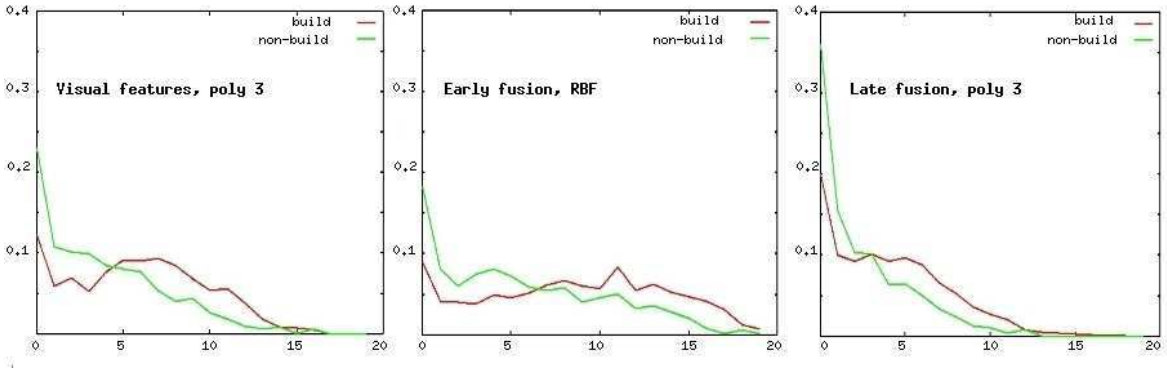


Figure 5.6 Comparison of SVM score distributions for (a) a visual features only approach, (b) early fusion and (c) late fusion approaches for *building* and *non-building* classes.

The comparison of the performance of the three approaches on our dataset for different types of kernels is presented in Table 5.10. As expected, the fusion approaches outperform the visual feature-based approach in all cases. For all kernel types used, apart from the linear kernel, the early fusion approach performs best on our dataset. Only in the case of linear kernel, does the late fusion approach outperform the early fusion approach. Overall, early fusion results in slightly higher precision and accuracy, but notably in higher recall. The difference in recall rates increases with the increased complexity of the kernel. Late fusion approaches give higher accuracy and precison rates than visual features approaches. However, the recall rates of the late fusion approaches are the lowest of all three approaches. On average the late fusion recall rates are 3.78% lower than those of the early fusion recall rates and 1.28 % lower than recall rates of visual features.

5.3 Discussion and Summary

In this Chapter, we evaluate three approaches to detection of buildings in consumer photographs, (i) a visual features-based approach, (ii) early fusion and (iii) late fusion approaches. Based on the experimental evaluation we note the following:

- The unconstrained, real-world dataset is far more challenging, mainly due to the addition of ambiguous photographs, hence there is a significant drop in performance of the visual feature-based method compared to its performance on the *ad hoc* dataset (outdoor images only) which is used in Chapter 3. In addition, our single-label approach to annotation, coupled with the presence of a significant number of ambiguous photos in the dataset that

belong to multiple semantic categories, present an additional challenge to the method. The ambiguous images and images that would have been much better described using multiple labels, were assigned to the closest category. Moreover, we have not separated out close-ups, although we are aware that this has been done in earlier work as "close-ups do not contain enough info for the algorithm sometimes even for the human to decide the category" [40]). Furthermore, we have a rather high image retention rate between the original collection and the dataset used in the experiments as ambiguous images and close-ups were retained in our test and training set. The fact that none of these image categories is uncommon in personal photo collections, makes the detection of buildings in unconstrained photo collections even more challenging. For instance, we kept 72% of the initial dataset, while only 39% of the original dataset was used for evaluation in similar work [40].

Table 5.10 Comparison of the classifier performances for each approach for the same kernel type (the best performing classifiers are highlighted).

Kernel type	***Approach***	***Accuracy***	***Recall***	***Precision***
Linear	Visual, j=1.24, BEP=71.1%	69.91	71.83	55.81
	Early fusion, j=1.25, BEP=71%	71.16	72.74	57.27
	Late fusion, j=0.96, BEP=73%	72.87	70.63	59.11
Poly 2	Visual, j=1, BEP=72.5%	71.34	72.15	57.58
	Early fusion, j=1.2, BEP=75%	73.69	75.08	60.29
	Late fusion, j=1, BEP=74%	73.55	70.49	60.12
Poly 3	Visual, j=1, BEP=73.3%	71.32	72.93	57.46
	Early fusion, j=1.2, BEP=76%	74.33	75.67	61.08
	Late fusion, j=1.01, BEP=74.5%	74.1	71.61	60.74
Poly 4	Visual, j=0.98, BEP=73.5%	70.99	72.54	57.09
	Early fusion, j=1.21, BEP=76%	74.36	75.67	61.11
	Late fusion, j=1.075, BEP=75%	74.1	71.26	60.81
RBF	Visual, j=0.91, BEP=72.86%	71.36	72.41	57.58
	Early fusion, j=1.1, BEP=76.6%	74.93	75.21	62.02
	Late fusion, j=0.9, BEP=76.5%	74	71.47	60.63

- Our approach to *indoor/outdoor* detection is most similar to the work of Luo and Savakis [51]. However, they used a single camera model which reflected significantly on the characteristics of the image collection. A total of 17 different camera types were used for capturing images in our collection. The use of a single camera type ensures uniformity in quality, reliability, and availability of camera metadata across the dataset. Considering the large number of cameras used, our collection may not be considered a typical personal

photo collection, although over time and within a family group the number of different camera models is likely to be more than one or two.

- The metatadata and content-based features complement each other by capturing context and content based aspects of image relevant information. The outdoor detection experiments confirm that the selected camera metadata on their own does have sufficient power to discriminate between *outdoor* and *indoor* images. Classification accuracy and recall rates for outdoor detection in our experiments are above 85.6%, while the precision is above 90%.

- We showed that use of metadata cues improves the accuracy of the building detector by an average of 2-3%, regardless of the operating point chosen. The performance measures are approximately 4% below the upper limit of the performance determined using ground truth information. We also show that for any given kernel type, with the exception of a linear kernel, the early fusion approach performs best on our dataset.

- However, the inadequacy of visual features selected appears to be the major obstacle to further improvement in the performance of the building detector. The major shortcoming of this approach is the fact that the visual features selected initially (i.e. edge-orientation based), which have been shown to perform well in a constrained dataset, do not have sufficient discriminative power to separate *buildings* from *outdoor non-building structures* in genuinely unconstrained photo collections. As our outdoor detection rate is above 85.6%, we conclude that, to an extent, the metadata resolves the issue of correct classification of *indoor structure* images. However, it is clear that additional visual features are required to disambiguate the *outdoor human-made structures* from *buildings*. We believe that the inclusion of colour features may improve the classification rates to a certain degree, but is unlikely to completely resolve the issue. We observe that, in general, most *buildings* are characterised by muted colours, shades of grey and brown, while on *outdoor non-building structures* such as vehicles, bridges, road signs, benches, fences, ships, etc., usually stronger, brighter colours prevail.

Chapter 6

Conclusions and Future Work

The purpose of this thesis has been to describe the work on detection of semantic concepts in digital photographs. In this thesis we detail the work on development of approaches to the detection of semantic concepts in still images, focusing on the detection of large buildings. The work also includes an approach to indoor/outdoor discrimination based solely on camera metadata. In the course of the work leading to this thesis, other related investigations and developments took place in the area of artificial text detection in video. Although this work is not directly a part of the main research activity presented here, it is nonetheless related to it. The work on detection of artificial text in video is presented in Appendix A so as not to interfere with the presentation of the main research work. This Chapter summarises the work presented in this thesis, analyses the results and points to directions which we believe may be worthwhile exploring in possible future extensions of the work.

Objective

The overall rationale for our work is an automatic detection of semantic concepts in digital visual content with the aim of providing semantic indices to the content. The ultimate objective of this work is to facilitate the organisation and efficient and user-friendly browsing and retrieval of visual content from large multimedia databases.

In this thesis we set out to explore the ways of extracting an appropriate set of features from both the image content and its context, and ways of integrating these features in order to obtain high-level semantic labels, i.e. to automatically generate indices to images in large photo

collections. The primary objective of the work was the identification of the appropriate low-level and other descriptors that could well capture the semantics of an image category e.g. buildings. In this thesis, we focus on the task of the detection of large buildings in consumer photographs by way of exploiting and integrating both the visual and the contextual cues.

Detecting Large Buildings in Natural Images using Visual Cues

In Chapter 3 we describe an investigation of a multi-scale approach to the detection of large buildings in images based solely on low-level visual features, i.e. edge orientation based features. In the image of interest, a building is either a single dominant object or one of the dominant objects. The approach (which assumes the implicit presence of contextual information in the form of an *indoor/outdoo*r label so that all input images are outdoor) is validated on an *ad hoc* dataset of 1720 images collected from various sources. The aim is to show that the feature representation based on a few simple and physically meaningful low-level features, combined with the high generalisation ability of the SVM classifier, may be sufficient to detect some high-level concepts such as buildings. Experimental results on our dataset of 1720 images show that the performance of our method, with accuracy of 88.22%, and recall and precision rates of 84.01% and 92.02%, is comparable to that of the existing approaches on the constrained datasets. However, it has to be emphasised that, in comparison to the MediAssist dataset that we use for evaluation in subsequent chapters, this dataset can be considered fairly constrained (all images were clearly categorised as outdoor).

We showed that, for constrained datasets, a simple single-feature representation coupled with the high generalisation capability of the SVM can be sufficient to detect high-level concepts such as buildings. However, we realise that the task of detection of buildings in an unconstrained dataset is a challenging task, given the large variations within the class and similarities with other classes, most notably with the *non-building structures*. The low-level visual features initially selected do not hold sufficient discriminative power to disambiguate between *building* and *non-building structures*. This is not surprising as those two categories are not easily discriminated even by humans in some instances, unless some contextual information is available[5]. Moreover, the overlap between *buildings* and other *human made structures* is reflected in the fact that some researchers actually include structures such as

5 An important value for benchmarking the system performance again human performance is called *interrater reliability*. Interrater reliability is the extent to which a human and a system agree in their decisions.

bridges etc. into *building* class [55].

Detecting Large Buildings in Natural Images – Fusion-based Approaches

In Chapters 4 and 5 we take a closer look at a photograph's contextual information, focusing on camera metadata in particular, and examine the ways of exploiting it towards an improved building detection in genuine, unconstrained and broad topic consumer photo collections. The purpose of using camera metadata is primarily to facilitate discrimination between indoor and outdoor photographs, and thus reduce the misclassification into buildings of indoor photographs that contain human-made structures. The discriminative power of a selected subset of camera metadata for the task of *indoor/outdoor* classification was examined. Our experiments on *indoor/outdoor* classification, based on camera metadata only, confirmed that the metadata alone has a sufficient discriminatory ability for the task. The accuracy of outdoor detection, using different number of metadata features and different training set sizes, ranged from 85.6-91.5 %.

Early and late fusion approaches to fusion of the existing low-level visual features with selected camera metadata are explored. In both cases we use SVM as integration device. The performance comparisons of the best performing classifiers for visual features only, early fusion and late fusion approaches show that the early fusion method, which combines 24 visual and 5 camera metadata features performs best on our dataset, with an accuracy of 74.93%, and recall and precision rates of 75.21% and 61.11% respectively.

We show that the introduction of metadata cues improves the classification accuracy of building detector by an average of 2-3% regardless of the operating point chosen. The performance measures are approximately 4% below the upper limit of performance determined using ground truth information. The early fusion approaches performed best on our dataset.

Integration of features from different modalities, i.e. a combination of complementary content-based and contextual information, does improve the classification performance as expected. The contextual information in the form of camera metadata is cheap and easily obtainable. However, different manufacturers support different sets of metadata tags, thus their availability and the degree of reliability varies.

Future Work

As regards the possible future work on the task of building detection in consumer photographs, the following directions may be worthwhile exploring:

- The inclusion of colour features is expected to bring further improvement in performance. Colour descriptors such as dominant colour, colour layout, and even heavily quantised histograms may be useful. The colour features are expected to aid the discrimination between the *buildings* and *outdoor non-building structures*.

- It is believed by the author, based on experiments, that for the late fusion approach, an introduction of additional visual cue-based classifiers trained on either different training sets or with different classifier settings, may be beneficial (i.e. to reduce the influence of the outdoor detector). This is because it appears that unweighted one-to-one approach may not be the most adequate. Also, further experimentation with the choice of SVM kernels and their parameters should be explored as the work presented here has dealt only with simple kernels using default values of most parameters.

- Extensive evaluation of the approach following the removal of ambiguous photos from the dataset in required in order to be able to compare the results with similar work. As an alternative, a multi-label approach to classification should be considered since the current single-label approach to photos that may belong to more than one class penalises the performance.

- The artificial text detection method in video may be adapted for detection of scene text in natural images. Text occurrences detected in natural scenes, such as road signs, names of streets, buildings are all useful pieces of information that could either help refine the results of the analysis of low-level cues on their own or provide an extra index themselves. At the very least, the presence of text indicates that a photo was captured in an urban(ised) location.

Appendix A

Artificial Text Detection in Digital Video

In this appendix we present the work [54], as published in WIAMIS'04, on developing an approach for detecting the artificial or overlay text in the MPEG-1 coded video, segmenting the text and enhancing it for further processing by the OCR (Optical Character Recognition) software. Unlike the work presented in the main body of the thesis, which deals with the detection of semantic concepts such as buildings in still images, here we present our work on detection of artificial text in video. The appendix is organised as follows. In section A.1 we introduce the problem. In section A.2, approaches upon which our work is based are summarised, and in section A.3 we give a brief overview of relevant compression standards. Section A.4 presents the algorithmic details of our approach for detection, localisation, enhancement and character segmentation. Section A.5 details the evaluation procedure and summarises the results obtained. Finally, section A.6 provides a conclusion.

A.1 Introduction

A significant challenge in large multimedia databases is the provision of efficient means for semantic indexing and retrieval of visual information. The need to handle large volumes of digital video data highlights the importance of the provision of efficient means for automated content-based indexing as the real value of the information stored in a large digital video archive is dependent on its accessibility.

Text appearing in digital video can be broadly categorised into two classes: *scene text* and *artificial text*. S*cene text* appears naturally as a part of the scene being recorded and is an

integral part of the image. Due to the accidental nature of its occurrence, scene text rarely carries significant information as for example a picture of a roadside sign containing a name of a town. Scene text may appear in almost any size, shape, colour and orientation, and as such it is often difficult to detect. On the other hand, *artificial text* (i.e. open caption or non-scene text) appearing in video is usually closely related to the visual content and is a strong candidate for high-level semantic indexing, thus offering an alternative or complementary approach to indexing based on low-level features extracted from the video or audio signal. The artificial text embedded in video frames frequently includes the most valuable information about the content of the video, such as scene location names, names of people, topics covered, sports scores, movie credits, etc. An index built by detecting, extracting and recognising the artificial text contained in a video sequence enables keyword-based queries in a manner similar to text-based retrieval.

Our approach for detecting and extracting artificial text regions in uncompressed video frames is essentially texture-based. We exploit the fact that text regions have different texture properties to the surrounding areas, such as alignment of edges along particular directions. Thus we localise all regions featuring a high concentration of short vertical edges that are horizontally aligned. Text regions are enhanced by smoothing and bi-linear interpolation, and are subsequently binarised by local thresholding in order to retain only pixels that exhibit high local contrast relative to the maximum contrast of the image, which is typical of pixels that form characters. In order to restore character fragments lost in the process of character segmentation, morphological processing is applied.

Figure A.1 Examples of (a) artificial text, (b) combination of artificial and scene text and (c) scene text in video frames.

A.2 Literature review

The vast majority of algorithms for text detection and extraction make use of typical characteristics of artificial text appearing in digital video, such as high contrast to the background, high density of short edges of varying orientation, horizontal alignment, various geometrical properties and temporal stability [92,93]. The first algorithms for detection/extraction of text from images were developed for still images. The methods used for still images had to be adapted for use with video given factors such as the considerable difference in quality, the low resolution of video frames, the presence of noise, the possibility of characters touching and complex backgrounds. Additional challenges to be addressed are the diversity of fonts, styles, colours, size and orientations that text occurring in video can exhibit.

Lienhart and Effelsberg in [42] used colour segmentation in the RGB colour space combined with edge analysis and empirically determined geometrical restrictions without making any assumptions about the text alignment (they calculated the direction of word's main axis in order to determine the writing direction). Temporal redundancy of text in video was exploited to eliminate non-text regions. Although they did not implement it in their approach in [42], the authors believe that temporal redundancy may be further exploited so as to improve the recognition results by means of a combination of recognition results pertaining to adjacent frames. The drawback of their approach is that it appears to work for large fonts only.

The approach of Lienhart and Wernicke [91], which used the properties of high contrast and high frequency to detect and localise the occurrences of artificial text, is capable of handling text sizes ranging from 8 pixels to half the frame, as well as estimating the text colour by colour quantization and comparison of colour histograms. Temporal redundancy was exploited to determine the colour, size and position of a particular text occurrence through comparison of colour, size and position of text located in adjacent frames.

Miene *et al* [63] adopted an approach based on region-growing methods in a colour-segmented image, followed by segmentation of characters from the background based on size (i.e. height, width and height-to-width ratio) and alignment constraints. Character candidates were clustered into word candidates by clustering regions of similar colour and height whose length

does not exceed a certain maximum value. The approach also includes a method aimed at restoring small character fragments which were lost during the segmentation step, thus improving the input for the OCR.

The approach of Wu *et al* [94] is a texture-based one: the text is treated as a distinctive texture which is characterised by certain frequency and orientation information. This feature was used so as to identify and segment initial text regions from the image. Furthermore, text also exhibits spatial cohesion: geometrical constraints, such as height similarity, spacing and alignment were applied to the segmented text regions in order to draw tighter rectangular bounding boxes around the text strings. In order to be able to detect text over a range of font sizes, a multi-scale approach was adopted: a pyramid of images was formed for each image and the detection algorithm was applied at each resolution. Subsequently, detection results at all resolutions were fused at the original resolution.

The approach of Li *at al* [40] is texture-based. The observation that text regions have different texture properties (i.e. similar frequency and orientation) than non-text regions forms the basis of their method. In [40], they view the problem of text detection and tracking in digital video as a multi-target detection and tracking problem: multiple text regions can occur in a frame and can move in different directions. Making use of the fact that the same text remains in the scene for a number of consecutive frames they performed the text detection only periodically whilst placing emphasis on the tracking process. As in [94], detection of different text sizes is facilitated using a three-level pyramid approach. Each instance of detected text kicks off the tracking module. Haar wavelets were used to detect the text regions characterised by line segments, and the results of text classification at all levels were integrated using a Neural Network.

Wolf and Jolion [92,93] applied a detection algorithm to each frame of the video sequence. All processing was performed on grey-scale images. Their approach makes use of the following properties of text in video: (i) grey level properties (high contrast in given directions), (ii) morphological properties (spatial distribution and shape), (iii) geometrical properties (height, width, height-to-width ratio) and (iv) temporal properties (stability). The main assumption that the method is based upon, is that artificial text regions are characterised by a high density of vertical edges which are horizontally aligned. In both, [92] and [93], temporal redundancy was

exploited to determine the final text bounding boxes and to obtain an enhanced image which is then binarised and passed to OCR for recognition. A combination of morphological processing and imposition of geometrical constraints was used to remove non-text regions. Segmentation was performed using a modified Niblack algorithm [92,93], which uses local thresholding.

The conclusions drawn from the aforelisted approaches can be summarised as follows:

- Overall, the approaches to text detection, both in still images and digital video, can be divided into two groups: connected component based and texture based. The c*onnected component based* methods [42,63] rely on the assumption that text pixels are characterised by similar colour or intensity, so they start with clustering of regions of similar colour that exhibit high contrast to their surrounding and then proceed to verify that those regions or components satisfy certain geometrical constraints, such as height, length, height-to-length ratio, etc. One of the drawbacks of connected component based approach lies in its inability to effectively handle text embedded in complex backgrounds.
- The *texture methods* [40,91,92,94], on the other hand, make use of the fact that text regions are characterised by their distinctive texture: i.e. particular frequencies and orientations, as well as their spatial cohesion. Various texture analysis methods, such as Gaussian filters, wavelets or simple edge detection filters, are used in order to locate text regions. In texture-based methods, imposition of geometrical constraints on initially detected regions is used as well in order to refine the detection results.
- Most approaches, unlike [42] which calculates the direction of word's main axis in order to determine the writing direction, assume horizontal alignment of artificial text.
- The temporal redundancy of video is exploited in different ways: to determine the colour, size and position of a particular text appearance through comparisons of text appearances in adjacent frames [91], to improve recognition results by combining recognition results in adjacent frames [42], to refine the results of text detection/localisation as in [92, 93] or simply to reduce computational demand by applying the text detection method only periodically as in [40].
- In order to facilitate the detection of text over a wide range of sizes, most methods adopt a multi-scale approach: the original image is decomposed into a number of images at different resolutions so as to form a pyramid of images. Detection of text of different sizes

is then achieved by applying the detection method to each of the decomposed images and eventually mapping and merging the detection results at different pyramid levels to the original image.

Our approach is closest in spirit to Wolf-Jolion approach. The principal differences between our approach and that of [92,93] are that our detection method is applied to every I-frame only, that we use the magnitude of the symmetrical horizontal difference as a measure of probability that a pixel belongs to a text region, that all text regions in frame are bounded by a single box, and that text segmentation is performed twice.

A.3 Relevant compression standards

The objective of compression is to reduce redundancy in data so as to facilitate more efficient storage and transmission of the data. Digital video is among the most information-intensive modes of communication and as such places huge demand on storage space and transmission bandwidth. As an illustration, a single frame decompressed from a 352x288 MPEG-1 stream into an RGB image takes 297 KB of storage space (i.e. 3 bytes for each pixel). With a frame rate of 25 frames per second, storing a single second worth of uncompressed video thus requires over 7 MB of storage space. In order to deal with the issue of storage and transmission, different compression standards for still images and digital video were developed, among others: MPEG-1, which is used for compression of digital video, and JPEG, used for compression of still images.

JPEG

JPEG stands for Joint Pictures Experts Group. JPEG is a standard for coding still images in a compressed format by means of exploiting the spatial redundancy in the image and limitations of the human eye (which is more sensitive to relative luminance changes than relative colour differences). JPEG is well suited for compressing full-colour or gray-scale images of natural scenes and is less suited for compressing synthetic (i.e. human-made or human-generated images). It is not suited for compression of black and white images such as line drawings, comics, etc. In general, images with abrupt changes in colour do not compress well with JPEG. JPEG is a block-based scheme that works on blocks of 8x8 pixels, on images in chrominance-

luminance colour space (YCrCb colour space), using discrete cosine transform to transform from spatial domain into frequency domain.

MPEG-1

MPEG stands for Moving Pictures Experts Group. MPEG is a family of standards used to code audio-visual information in a digital compressed format. In MPEG-1, video consists of a sequence of still images, each of which corresponds to a two-dimensional array of pixels. Each pixel in an array has three colour components: red, blue and green. MPEG is a block-based coding scheme which operates on images in YCrCb colour space (Luminance, Chrominance Red, Chrominance Blue) and exploits both the spatial, or intra-frame redundancy (discrete cosine transform coded 16x16 macroblocks), and temporal, or inter-frame redundancy (motion vector), of data in digital video.

The MPEG-1 sequence consists of three types of coded frames: I (intra frames), P (predicted frames) and B (backward predicted frames). The I-frames are coded as still images, i.e. As effectively equivalent to JPEGs (only the spatial redundancy is reduced), and thus contain all the data necessary to fully describe a particular frame. Therefore, an I-frame can be decoded independently of the other frames in the stream. The P frames are coded as differences between the given frame and previous I or P-frame. The B-frames are described as difference between the previous and the following reference frames in the sequence (i.e. I or P-frame). In oder to reconstruct a P-frame at the decoder end, the most recently reconstructed I or P frame is used. Decoding a B-frame requires the two closest I or P-frames, the immediate predecessor and immediate successor of the B-frame being decoded.

The main advantage of the MPEG-1 compression standard over other compression standards lies in the fact that, for the same picture quality, MPEG files are much smaller (higher compression ratio). The MPEG-1 video sequences used in this work have frame sizes of 352x288 and play at the rate of 25 frames per second.

A.4 Our approach

A functional diagram of the system is presented in Figure A.2. The detection algorithm, which operates in the uncompressed domain, is applied to every I-frame of the MPEG-1 video

sequence only, thus exploiting the temporal stability of artificial text over a number of frames. A single rectangular box that bounds all candidate text regions is defined for each I-frame. Following image enhancement and morphological processing these rectangular boxes are cropped and binarised, and subsequently passed to the OCR module. These steps are described in more detail in the following.

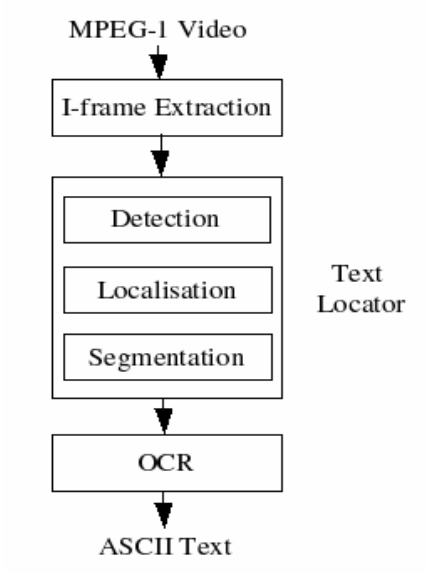


Figure A.2 System block diagram.

A.4.1 Detection and localisation of artificial text

Our detection method is based on texture analysis, relying on the property of Latin script to form a texture characterised by a high density of vertical edges aligned horizontally. The method operates on an uncompressed video frame in a YUV colour space (luminance, chrominance red, chrominance blue). Temporal stability of text in video is taken into account through an assumption that a particular text appearance has to remain visible for a certain minimum duration (i.e. approx. 1 second) in order to be readable. It is therefore sufficient that only I-frames be examined and analysed for the presence of text that appears over a number of consecutive frames.

Edge detection and processing

The magnitude of the symmetrical horizontal difference is calculated for each pixel in the luminance component of the frame. Each pixel value in the resulting image is a measure of the

probability that it belongs to a text region. Pre-processing prepares the edge map for binarisation by joining the vertical edges horizontally into clusters corresponding to words and text lines. The edge map is first smoothed using a 3x3 binomial filter, which is followed by blurring horizontal using a 3x1 mask. A small blurring mask is used in order to avoid connecting noisy areas to areas containing text. Erosion by a cross mask is then carried out to clear the top intensity layer in the greyscale image [4]. As a result, bright text areas slightly shrink in size, but so do the noisy edges in the background. Subsequent smoothing increases the size of text regions as does a final 3x3 dilation. Figure A.3 illustrates the effects of some of these processing steps.

Binarisation

Binarisation of the edge map is performed in order to separate text-containing regions from the rest of the frame using Otsu's global thresholding method as described in [92,93]. An optimal threshold is calculated based on the grey level histogram by assuming Gaussian distributions of text pixels and non-text pixels. The method aims to maximise the interclass variance. The optimal threshold is calculated using the following formula [92,93]:

$$t = arg\ max_t \quad \left(\omega_0 \omega_1 (\mu_1 - \mu_0)^2\right) \tag{11}$$

where ω_0 is the normalised mass of class 0 (i.e. the number of pixels in the class divided by the total number of pixels in the image), ω_1 is the normalised mass of class 1, and μ_1 and μ_0 are mean grey levels for each of the classes. Unlike [92,93], in this system thresholding is implemented based on a 64-bin histogram using a single threshold. Ideally, this step results in an image featuring clusters of white pixels in areas corresponding to the text regions. In practice, small clusters of white pixels may appear elsewhere in the frame. Binarisation is followed by post processing to remove these noisy areas. Figure A.3 also shows the result of the edge map binarisation before and after morphological processing.

Fitting bounding boxes

The aim of this step is to fit a single bounding box that encloses all text areas in the frame. This requires that as much noise as possible be removed beforehand, otherwise there is a risk that the bounding box may potentially grow to reach the size of the frame. As can be seen from Figure A.3, the binarised edge map contains some noise pixels. In order to remove these,

several steps are taken. The first step is to use a 3x3 median filter that deals well with the noise spikes whilst preserving the edges. In the sample frame the benefit of median filtering in removing noise is not obvious as the noise spot is not a single pixel. However, its averaging effect is clear as the tiny black spots were removed from the white regions. The next step is a 3x1 dilation followed by a 5x5 opening. As can be seen in Figure A.3, a 5x5 erosion succeeds in removing the noisy spot while the subsequent dilation with the same size structuring element restores the desired white clusters to their initial size.

Figure A.3 (a) Input image, (b) horizontal difference magnitude, (c) (d) binarised edge map before and after morphological processing.

Finally, a dilation in the horizontal direction using a 7x1 structuring element connects text pixels into text lines. In order to compensate for any damage to text regions during the previous processing, the text box size is adjusted by growing it by 5 pixels in all four directions. Ideally the adjustment should be proportional to the bounding box dimensions. Geometrical constraints are imposed on bounding boxes and those failing to satisfy minimum area and width criteria are discarded. In Figure A.4, the cropped text region identified from Figure A.3 is presented. Another, enlarged, cropped text image is shown in Figure A.5 illustrates the intensity variations across both character as well as background pixels.

Figure A.4 Cropped text image.

Figure A.5 Intensity variations across character and background pixesls.

A.4.2 Segmentation of characters

The purpose of the segmentation stage is to separate the character pixels from the background pixels and to form an image that contains only black character pixels on a white background, which is a suitable input for the recognition stage. Segmentation by local thresholding is performed based on the assumptions that (i) characters have high contrast to the background and (ii) characters are monochromatic regions. Some segmentation results are shown in Figure A.6.

Pre-processing of cropped image

In order to meet the high-resolution requirement imposed by OCR, the cropped image is bi-linearly interpolated by a factor of 4. This ratio is chosen so as to ensure that the smallest size font that occurs in video, such as movie subtitles, is enlarged sufficiently to constitute a suitable input to the OCR stage. This decision is based on a comparison of the movie subtitles font size in a test video and the suggested character size supplied with the OCR package we use. A last pre-binarisation step involves filtering using a 3x3 median filter in order to remove noise spikes.

Binarisation of cropped image

Separation of character pixels from non-character or background pixels is based on local thresholding using a modified Niblack algorithm as presented in [92,93]. The binarisation decision is made using a rectangular 5x5 window that is shifted across the image using the mean and standard deviation ([92,93] use variance) of grey levels in a window.

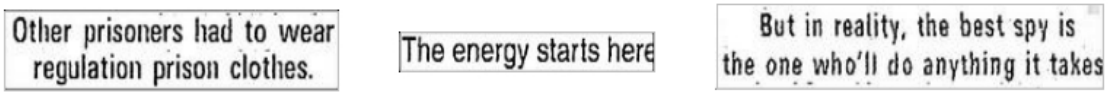


Figure A.6 Some text segmentation results.

Only those pixels that exhibit a high local contrast relative to the maximum contrast in the image and the contrast of the window are retained. The following equation is used for the calculation of the threshold value [92]:

$$T = (1 - a)m + aM + a\frac{s}{R}(m - M) \tag{12}$$

where m is the mean grey level value in a window, s is the standard deviation of grey values in the window, M is the minimum grey level value for the whole image, R is maximum standard deviation for all windows. It is suggested in [92,93] that the parameter a be set to 0.5. However, as character strokes of segmented text appeared to be too thin and fragmented with parameter a set to 0.5, different values were investigated and 0.1 was determined to be the most suitable for our purposes. Our experiments also showed that using maximum standard deviation rather than maximum variance as used in [92,93], resulted in more accurate segmentation. Following the smoothing step after the first segmentation, the second segmentation is performed as we noticed that it improves the quality of the segmented characters.

A.4.3 Recognition

Recognition is performed using freely available optical character recognition software known as Clara OCR [14]. This OCR package does not have in-built fonts and thus requires training. A considerable amount of effort has to be put into training so as to build a sufficiently large database of character patterns in order to enable Clara OCR to deal successfully with a variety of sizes, fonts and the varying degree of character fragmentation that occurs in the segmentation process. It is clear from our experiments that overall recognition results critically depend on the quality of the input provided by the segmentation stage. Any fragmentation or damage to the characters due to the presence of noise in video is likely to considerably disrupt the character recognition process.

A.5 Experimental evaluation

A.5.1 Dataset

In order to evaluate the performance of the system, testing was carried out on two MPEG-1 encoded CIF video sequences from our own video database. The first video sequence contained 1000 frames with 26 frames containing the appearance of artificial text. The second sequence contained 30000 frames, 795 of which contained artificial text. The ground truth used for evaluation was created by manually transcribing the sequences. Figure A.7 shows examples of the video frames used.

Figure A.7 Examples of video frames from our database.

A.5.2 Results of text detection

Text detection performance was evaluated manually against the ground truth by determining the percentage of characters in a frame that have been successfully located and enclosed by a bounding box. Detection recall was defined as the ratio between the number of characters enclosed and the total number of characters that appear in a frame. For each new appearance of text on screen, the best detection result was manually chosen. Automating this process using temporal information will be the basis for our future work in this area. Analysis of the

accuracy of detection within a frame over the entire test corpus showed the following. The best-candidate detection recall for the first sequence was 95%, and 83 % for the second sequence, giving an average overall detection recall of 83.2%.

Table A.1 Text detection results.

	# frames	Recall (best candidate)
Sequence 1	1000	95%
Sequence 2	30000	83%
Overall	31000	83.20%

A.5.3 Results of recognition

Since the main focus of our work is on segmentation and not OCR, and given the significant effort required to train the OCR package using segmentation results, we have only evaluated recognition performance for sequence 1. Recognition performance was evaluated through comparison of the OCR recognition results with the manually generated ground-truth. Character-based recognition recall ranged from 81-84% while the recognition precision was within the range 66-74%.

A.6 Conclusions

In this appendix we present a method to detect, localise and segment artificial text from video. The evaluation of the method on 31000 video frames showed moderately good detection recall. However, currently the evaluation is based upon manual selection of the best detection results for the appearance of a given piece of text. Further research is required in order to utilise temporal information so as to automate this process, e.g. by accumulating segmentation results over a number of frames. Further work on training the OCR package for recognition using segmentation results across different sequences is also required. Furthermore, from the Figure A.8 we can draw a conclusion that multiple-bounding-boxes approach, where a line of text is tightly bound by its own bounding box, as opposed to our one-bounding-box per frame, would be more effective and would both ease the the task of the segmentation module and

facilitate the improved character segmentation.

Figure A.8 (a) Input image, (b) horizontal difference magnitude, (c) and (d) binarised edge map before and after morphological processing, (e) cropped text image, and (f) segmented text.

Bibliography

[1] W.H. Adams, G. Iyengar, C-Y. Lin, M.R. Naphade, C. Net, H.J. Nock, J.R. Smith, "Semantic Indexing of Multimedia Content using Visual, Audio and Text cues", *EURASIP Journal on Applied Signal Processing,* pg 170-185, 2003.

[2] S. Arivazhagan and L. Ganesan, "Texture classification using wavelet transform", *Journal of Pattern Recogn. Letters, Vol. 24, 9-10*, 2003, pg 1513-1521, Elsevier Science, New York, NY, USA.

[3] A. Barla, F. Odone and A. Veri, "Old fashioned state-of-the-art image classification", *Proceedings of 12th International Conference on Image Analysis ans Processing (ICIAP'03)*, Los Alamitas, CA, USA.

[4] H. Bässmann and P.W. Besslich, *Ad Oculus, Digital Image Processing, Student Version 2.0*, ITP, London, 1995.

[5] N. Boujemaa, S. Boughorbel and C. Vertan, "Soft Colour Signatures for Image Retrieval by Content", Eusflat'2001.

[6] M. Boutell, C. Brown, and J. Luo, "Review of the State of the Art in Semantic Scene Classification", Technical Report, University of Rochester, December 2002.

[7] M. Boutell and L. Luo, "Incorporating Temporal Context with Content for Classifying Image Collections", *Proceedings of International Conference on Pattern Recognition (ICPR)*, p 947-950, Cambridge, UK, 2004.

[8] M. Boutell and J. Luo, "Beyond Pixels: Exploiting Camera Metadata for Photo Classification, *Journal of Pattern Recognition 38*, pg 935-946, June 2005.

[9] S. Brandt, "Use of Shape Features in Content-Based Image Retrieval", Master's Thesis, August 1999.

[10] K. Broom, *Gestalt Theory of Visual Perception*, http://www.users.totalise.co.uk/~kbroom/Lectures/gestalt.htm

[11] C.J.C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition", *Data Mining and Knowledge Discovery*, 2, pg 121-167, 1998.

[12] A. Canuto, G. Howells and M. Fairhurst, "The use of confidence measures to enhance combination strategies in multi-network neuro-fuzzy systems", *Journal on Connection Science*, Volume 12, Issue 3 & 4 December 2000 , pg 315 – 331.

[13] Canny edge detector, http://www.cee.hw.ac.uk/hipr/html/canny.html

[14] Clara OCR Advanced User's Manual and Tutorial, available at: http://www.claraocr.org

[15] M. Cord, P.H. Gosselin, S. Philipp-Foliguet, "Stochastic exploration and active learning for image retrieval", *In Image and Vision Computing*, N25, pg 14-23, 2006.

[16] K. Coyle, “Metadata: Data With a Purpose”, http://www.kcoyle.net/meta_purpose.html

[17] N. Dalal, and B. Triggs, “Histograms of Oriented Gradients for Human Detection”, I*EEE Conf. on Computer Vision and Pattern Recognition II*, pp. 886-893, 2005.

[18] A. Dorado and E. Izquiredo, “Exploiting Problem Domain Knowledge for Accurate Building Image Classification”, *Proceedings of CIVR 04*, Dublin, Ireland, July 2004.

[19] R.O. Duda, P.E. Hart and D.G. Stork, *Pattern Classification*, 2nd edition, A Wiley-Interscience Publication, 2001, USA.

[20] T. Ebrahimi, “Image and Video Analysis: Trends and Challenges”, Position paper at WIAMIS'04, Lisbon, Portugal, 2004.

[21] EXIF 2.2, http;//www.jeita.or.jp/english/index.htm

[22] M.A.T. Figueiredo and A.K. Jain, “Unsupervised Learning of Finite Mixture Models”, I*EEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.24, No. 3, March 2004

[23] G. D. Forney, “The Viterbi algorithm”, *Proceedings of the IEEE 61(3),* pg 268–278, March 1973.

[24] D. L. Hall and J. Llinas, “An Introduction to multisensor data fusion”, *Proc. IEEE*, vol 85, pg 6-23, January 1997.

[25] M. Hauta-Kasarim, J. Parkkinen, T. Jaaskelainen, and R. Lenz, “Generalized Cooccurrence Matrix for Multispectral Texture Analysis”, *Proceedings, 13th International Conference on Pattern Recognition*, Vienna, Austria, August 25-29, 1996, Vol 2., pg 785-789.

[26] Y. Rui, T.S. Huang and S-F. Chang, “Image Retrieval: Past, Present and Future”, *Procedings of International Symposium on Multimedia Information Processing*, 1997.

[27] Q. Iqbal and J.K. Aggarwal, “Applying Perceptual Grouping to Content-based Image Retrieval: Building Images”, *Proceedings of the IEEE Int Conf on Computer Vision and Pattern Recognition*, Fort Collins, USA, June 1999.

[28] Q. Iqbal and J.K. Aggarwal, “Lower-level and Higher-level Approaches to Content-based Image Retrieval”, *Proceedings of the IEEE South West Symposium on Image Analysis and Interpretation*, Austin, Texas, USA, pg 197-201, April 2000.

[29] Q. Iqbal and J.K. Aggarwal, “Combining Structure, Colour and Texture for Image Retrieval: A Performance Evaluation”, *Proc of Int. Conf. On Pattern Recognition (ICPR),* Quebec, Canada, 2002

[30] B. Jahne, *Digital Image Processing, Concepts, Algorithms and Scientific Applications*, 4th edition, Springer-Verlag Berlin Heidelberg, 1997.

[31] B. Jahne, H. Haussecker, *Computer Vision and Applications, Guide for Students and Practitioners*, Academic Press, A Harcourt Science and Technology Company, San Diego, USA, 2000.

[32] B. Jahne, *Digital Image Processing*, 6th revised and extended edition, Springer-Verlag Berlin Heidelberg, 2005.

[33] A.K. Jain,, "Statistical Pattern Recognition: A Review", *IEEE Transactions on Patten Analysis and Machine Intelligence*, November, 1999.

[34] A.K. Jain and A. Vailaya, "Image retrieval using colur and shape", *Pattern Recognition, Vol. 29, No. 8*, pg 1233-1244, 1996.

[35] F. Jelinek and R.L. Mercer, "Interpolated estimation of Markov source parameters from sparse data", *Pattern Recognition in Practice* (Amsterdam, May 21-23 1980), E. S. Gelsema and L. N. Kanal, Eds., North Holland, pg. 381-397, 1980.

[36] T. Joachims, SVMlight, http://svmlight.joachims.org/

[37] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features", *Proceedings of 10th Eur. Conf. on Machine Learning,* pp. 137–142, 1998.

[38] J. Kittler, M. Hatef, and R. P. W. Duin, "Combining classifiers", *Intl. Pattern Recognition*, pg 897--901, 1996.

[39] S. Kullback and R. A. Leibler, "On information and sufficiency", *Annals of Mathematical Statistics 22(1): pg 79-86,* 1951.

[40] H. Li, D. Doermann, and O. Kia, "Automatic text detection and tracking in digital video", *IEEE Transactions on Image Processing*, 9(1), pg 147-156, January 2000.

[41] H. Lieberman and H. Liu. "Adaptive Linking between Text and Photos Using Common Sense Reasoning", *Adaptive Hypermedia and Adaptive Web-Based Systems*, Springer-Verlag, Berlin. Pg 2-11, 2002.

[42] R. Lienhart and W. Effelsberg, "Automatic text segmentation and text recognition for video indexing", *ACM/ Springer Multimedia Systems, vol.8,* pg 69-81, January 2000.

[43] R. Lienhart, "Automatic Text Recognition for Video Indexing", *Proc. ACM Multimedia 96*, Boston, USA, pg 11-20, November 1996.

[44] J.H. Lim, P. Mulhem and Q. Tian, "Event-Based Home Photo Retrieval", *IEEE MultiMedia,* vol. 1, pg 33-36, 2003.

[45] W-H. Lin, R. Jin and A. Hauptmann, "Meta-classification of Multimedia Classifiers", *International Workshop in Knowledge Discovery in Multimedia and Complex Data*, Taipei, Taiwan, May 2002.

[46] W-H. Lin and A. Hauptmann, "News Video Classification Using SVM-based Multimodal Classifiers and Combination Strategies", *In ACM Multimedia '02*, Juan-les-Pins, France, 1-6 December 2002.

[47] T. Lindeberg, "Scale-space theory: A basic tool for analysing structures at different scales", *Journal of Applied Statistics*, 21, 2, pg 224-270, 1994

[48] D.G. Lowe, "Object recognition from local scale-invariant features", *Proceedings of ICCV '99*, Volume 2, pg 1150, Washington, USA, 1999.

[49] J. Luo, A.E. Savakis and A. Singhal, "A Bayesian network-based framework for semantic image understanding", *Journal of Pattern Recognition 38*, p 919-934, June 2005

[50] J. Luo and M. Boutell, "Natural Scene Classification Using Overcomplete ICA", *Pattern Recognition*, February 2005.

[51] J. Luo and A. Savakis, "Indoor vs Outdoor Classification of Consumer Photographs using Low-level and Semantic Features", *Proc. of IEEE Int. Conf. On Image Processing, ICIP 2001*, Thessalonki, Greece, October 2001.

[52] M.D. Levine, *Vision in Man and Machine*, McGraw-Hill Book Company, Series in Electrical Engineering, USA, 1985.

[53] J. Lyons-Weiler, S. Patel and S. Bhattacharya, "A Classification-Based Machine Learning Approach for the Analysis of Genome-Wide Expression Data".

[54] J. Malobabić, N. O'Connor, N. Murphy, S. Marlow, "Automatic Detection and Extraction of Artificial Text in Video", *Proceedings of the 3rd International Workshop on Image Analysis for Multimedia Interactive Services*, WIAMIS 2004, Lisbon, Portugal, April 2004.

[55] J. Malobabić, H. Le Borgne, N. Murphy, N. O'Connor, "Detecting large buildings in natural images", *Proceedings of Int. Workshop on Content-Based Multimedia Indexing*, CBMI 2005, Riga, Latvia, June 2005.

[56] D. Marr, *Vision,* W. H. Freeman, New York NY., pg 101, Figure 3-1, 1982.

[57] K. McDonald and A.F. Smeaton, "A Comparison of Score, Rank and Probability-based Fusion Methods for Video Shot Retrieval", *Proceedings of the Fourth International Conference on Content-based Image and Video Retrieval* (CIVR), Singapore, Singapore, 2005.

[58] J. Mao and A.K. Jain, "Texture classification and segmentation using multi-resolution simultaneous autoregressive models", *Journal of Pattern Recognition,* Vol. 25, Number 2, 1992, pg. 73—188, Elsevier Science Inc., New York, NY, USA.

[59] MediAssist Project, http://www.nsd.ie/htm/comm_rad/cr.php3

[60] J. Meeus, *Astronomical Algorithms,* Willmann-Bell Inc., 1998, Richmond, Virginia, USA.

[61] G. Merle, F. Dubouloz-Monnet, P.Lambert and A.C Grillet, "Global and multi-scale image analysis using power spectra", *Journal of Measurement Science Technology*. 16 No 3, pg 805-812, March 2005.

[62] Merriam-Webster OnLine Dictionary, http://www.m-w.com

[63] A. Miene, Th. Hermes and G. Ioannidis, "Extracting Textual Inserts from Digital Videos", *Proceedings of the 6th Int. Conference on Document Analysis and Recognition* (IDCAR'01), Seattle, USA, pg1079-1083, September 2001.

[64] A. Mojsilović and B. Rogowitz, "Capturing Image Semantics with Low-Level Descriptors", *Proceedings of IEEE Int. Conf. On Image Processing, ICIP 2001*, Thessalonki, Greece, October 2001.

[65] P. Ndjiki-Nya, O. Novychny and T. Wiegand, "Video Content Analysis using MPEg-7 Descriptors", *Proceedings of 1st European Conference on Visual Media Production (CVPM)*, pg 15-16, March 2004.

[66] N. O'Hare, "Indoor/Outdoor Classification Using EXIF Metadata", Internal report, Dublin City University, May 2005.

[67] T. Ojala, M. Aittola and E. Matinmikko, "Empirical Evaluation of MPEG-7 XM Colour Descriptors in Content-Based Retrieval of Semantic Image Categories", *Proceedings of 16th Int. Conference on Pattern Recognition, ICPR'02, Vol. 2*. pg 21021, Los Alamitos, USA, 2002.

[68] S. Paek, C.L. Sable, V. Hatzivassiloglou, A. Jaimes, S-F Chang and K.R. McKeown, "Integration of Visual and Text-based Approaches for the Content Labelling and Classification of Photographs", *Proceedings of 2000 ACM CIKM International Conference on Information and Knowledge Management*, New York, USA, ACM Press, 2000.

[69] K. Pelckmans, J. De Brabanter, J.A.K. Suykens and B. De Moor, "Handling missing values in support vecor machine classifiers", *Neural Networks*, Vol. 18 , Issue 5-6, June 2005.

[70] Matti Pietikäinen, Topi Mäenpää and Jaakko Viertola, "Colour Texture Classification with Colour Histograms and Local Binary Patterns", *Texture02*, pg 109-112, 2002.

[71] I. Pitas, *Digital Image Processing Algorithms*, University Press, Cambridge, 1993.

[72] E. Rome, "Simulating Perceptual Clustering by Gestalt Principles", *Proceedings of 25th Workshop of the Austrian Association for Pattern Recognition*, OAGM/AAPR 2001, Berchtesgaden, Germany, 7-8 June 2001.

[73] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval", *Information Processing and Management*, 24(5), pg 513-523, 1988.

[74] A. Savakis, S.P. Etz, A.C. Loui, "Evaluation of image appeal in consumer photography", *Proc. SPIE Human Vision and Electronic Imaging,* San Hose, California, USA, 2000.

[75] N. Sebe, M.S. Lew, X. Zhou, T.S. Huang and E.M. Bakker, "The State of the Art in Image and Video Retrieval", *Proc. of CIVR'03*, Urbana-Champaign, IL, USA, July 24-25, 2003.

[76] N. Serrano, A.E. Savakis and J. Luo, "Improved Scene Classification using Efficient Low-level Features and Semantic Cues", *Pattern Recognition*, 37, 2004.

[77] http://en.wikipedia.org/wiki/User:Aramgutang/Gallery

[78] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta and R. Jain, "Content-based Image Retrieval at the End of Early Years", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22 (12).

[79] J.R. Smith, S. Srinivasan, A. Amir, S. Basu, G. Iyengar, C-Y. Lin, M. Naphade, D. Ponceleon and B. Tseng, "Integrating features, models , and semantics for TREC video retrieval", in: *NIST Special Publication 500-200: The Tenth Text REtrieval Conference (TREC 2001)*, pg 240-249, Gaithersburg, Maryland, US, 2001.

[80] C.G.M. Snoek, M. Worring and A.W.M. Smeulders, "Early versus Late Fusion in Semantic Video Analysis", *Proceedings of the 13th annual ACM International Conference on Multimedia*, November 06-11, 2005, Hilton, Singapore.

[81] M. Sonka, V. Hlavač, and R. Boyle, *Image Processing, Analysis and Machine Vision*, 2nd edition, Brooks/Cole Publishing Company, 1999.

[82] M. Szummer and R.W. Picard, "Indoor/outdoor Image Classification", *IEEE Intl Workshop on Content-based Access to Image and Video Databases*, Bombay, India, January 1998.

[83] H. Tamura, S. Mori and T. Yamawaki, "Textural features corresponding to visual perception", *IEEE Trans. Systems, Man and Cybernetics, Vol. SMC-8, No.6*, pg 46-473, June 1987

[84] A.G. Taylor, *The Organization of Information,* Westport: Libraries Unlimited 2004.

[85] K. Toyama, R. Logan, A. Roseway and P. Anandan, "Geographic Location Tags on Digital Images", *Proceedings of the 11th ACM International Conference on Multimedia*, Berkeley, CA, USA, 2003.

[86] M. Turk and A. Pentland, "Eigen faces for recognition", *Journal of Cognitive Neuroscience*, vol. 3, pp. 71—86, 1991.

[87] A. Vailaya, M. Figueiredo, A. Jain and HJ. Zhang, "Content-Based Hierarchical Classification of Vacation Images", *Proceedings of IEEE International Conference on Image Processing*, October 1999.

[88] A. Vailaya, A. Jain and HJ Zhang, "On Image Classification: City Images vs. Landscapes", *Journal of Pattern Recognition*, 1998.

[89] P. K. Varshney, "Multisensor data fusion", *Electronics & Communications Engineering Journal*, December 1997.

[90] K. Tieu and P.Viola, "Sparse High-Dimensional Representations and Large Margin Classifiers for Image Retrieval", www.ai.mit.edu/research/abstracts/abstracts2002/computer-vision/35tieu.pdf, 1997.

[91] P. Wernicke and R. Lienhart, "On the Segmentation of Text in Videos", *IEEE International Conference on Multimedia and Expo (ICME2000)*, Vol.3, pg 1511-1514, July 2000.

[92] C. Wolf, J.M. Jolion and F. Chassaing, "Text Localization, Enhancement and Binarization in Multimedia Documents", *Proceedings of the Int. Conference on Pattern Recognition (ICPR) 2002, vol.4*, IEEE Computer Society,Quebec City, Canada, pg 1037-1040, August 2002.

[93] C. Wolf and J.M. Jolion, "Extraction and Recognition of Artificial Text in Multimedia Documents", *Technical Report RVF-RR-2002.01,* Available: http://rvf.insa-lyon.fr/~wolf/papers/tr-rfv-2002, February 2002.

[94] V. Wu, R. Manmatha, E. M. Riseman, "Finding Text In Images", *Proceedings of the 2nd ACM International Conf. on Digital Libraries '97*, Philadelphia, USA, 1997.

[95] A. Yavlinsky, M. J. Pickering, D. Heesch and S. Ruger, "A Comparative Study of Evidence Combination Strategies", *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), Montreal, Quebec, Canada, 2004.