

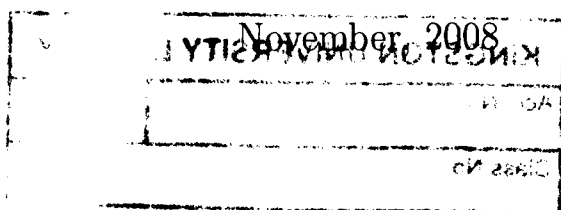
An Investigation into the Generation, Encoding and Retrieval of CCTV-derived Knowledge

James Alexander Grove Annesley

Submitted in partial fulfilment
of the requirements of Kingston University
for the degree of Doctor of Philosophy.

Collaborating Partners:

Faraday Imaging Partnership, EPSRC, Overview Ltd.





IMAGING SERVICES NORTH

Boston Spa, Wetherby

West Yorkshire, LS23 7BQ

www.bl.uk

BEST COPY AVAILABLE.

VARIABLE PRINT QUALITY

Abstract

Modern video surveillance systems generate diverse forms of data and to facilitate the effective exchange of these data a methodical approach is required. This thesis proposes the Video Surveillance Content Description Interface (VSCDI), a component of ISO/IEC 23000-10 – Information technology – Multimedia application format (MPEG-A) – Part 10: Video surveillance application format. The interface is designed to describe content associated with and generated by a surveillance system. In particular, a set of descriptors are included for: content-based image retrieval; user-defined Classification Schemes to impose any required description ontology; and to provide consistent descriptions across multiple sources. The VSCDI is evaluated using comparisons with other meta-data frameworks and in terms of the performance of its colour descriptor components. Two new data sets are created of pedestrians in indoor environments with multiple camera views for re-identification experiments. The experiments use a novel application of colour constancy for cross-camera comparisons. Two evaluation measures are used: the *Average Normalised Mean Retrieval Rate* (ANMRR) for ranked estimates; and the Information Gain metric for probabilistic estimates. Techniques are investigated for using more than one descriptor both to provide the estimate and to represent a person whose image is split into Top and Bottom clothing components. The re-identification of pedestrians is discussed in the context of providing both a coherent description of the overall scene activity and within an embedded system.

Acknowledgements

The support of Overview Ltd., the EPSRC (Engineering and Physical Sciences Research Council) and Imaging Faraday Partnership is gratefully acknowledged, without whom I would not have been able to complete this research.

I would like to thank Dr. James Orwell, my Director of Studies, for the excellent supervision and direction he gave me throughout the project. I would also like to thank my other supervisors Prof. Graeme Jones, Dr. Paolo Remagnino and my industrial supervisors, Dr. Paul Jones, Dr. Alex Starling, who helped with the Faraday Writing Competition, Mr. Ben Henrickson, who was interested in colour-based retrieval and Mr. David Watkins. Thanks to those at MPEG: Mr. Gero Bäse; Mr. Houari Sabirin; Dr. Kate Grant; and Mr. Jim Aldridge, who provided me with valuable experience and collaboration. Thanks to those at Kingston University who have assisted me in reaching my goals: Dr. Valerie Leung, who provided expertise in information theory and image processing; Mr. Alberto Colombo, who provided expertise in camera calibration; Mr. Justin Cobb, for his programming support; and Mr. John-Paul Renno, who processed the GENERICK data set to segment the moving objects. He also provided some invaluable image processing “tips of the trade”. Thanks also to Dr. Gordon Hunter, who provided useful explanations regarding Bayes’ Theorem. Finally, thanks goes to my father for his extensive proof reading endeavours.

My love and gratitude goes to my wife who provided enduring love and support. A big thank you goes to my parents and parents-in-law for their support and finally, a special message goes to my children.

List of Acronyms

Acronym — Meaning

AAF	Advanced Authoring Format
AC-3	Arc Consistency Algorithm #3
AF	Application format
ANMRR	Average Normalised Mean Retrieval Rate
ANPR	Automatic numberplate recognition
API	Application Programming Interface
ARTS	Association for Retail Technology Standards
AVC	Advanced Video Coding
CARETAKER	Content Analysis and Retrieval Technologies to Apply Knowledge Extraction to Massive Recording
CASE	Collaborative Awards in Science and Engineering
CAVIAR	Context Aware Vision using Image-based Active Recognition
CBIR	Content-Based Image Retrieval
CBR	Content-Based Retrieval
CCTV	Closed-circuit television
CDP	Core Description Profile
CIF	Common Interchange Format
COM	Common Object Model
CORBA	Common Object Request Broker Architecture
CPU	Central processing unit
CVML	Computer Vision Markup Language
DAML+OIL	DARPA Agent Markup Language + Ontology Inference Language
DARPA	Defense Advanced Research Projects Agency
DAVP	Detailed Audiovisual Profile
DCT	Discrete Cosine Transform
DDL	Data Definition Language
DEPA	Distributed Enhanced Processing Architecture
DID	Digital Item Declaration
DIG	Digital Imaging Group
DoD/IC	Department of Commerce Information and Communications
DPA	Data Protection Act
DRM	Digital Rights Management

DS	Description Scheme
DSP	Digital Signal Processor
DVD	Digital Versatile Disc / Digital Video Disc
EBU	European Broadcasting Union
EPG	Electronic Programme Guide
EPSRC	Engineering and Physical Sciences Research Council
ETISEO	Evaluation du Traitement et de l'Interpretation de Sequences Video
ETSI	European Telecommunications Standards Institute
EU	European Union
Exif	Exchangeable Image File Format
FCD	Final Committee Draft
GENERICK	Generation Encoding and Retrieval of CCTV-derived Knowledge
GMM	Gaussian Mixture Model
GOF	Group of Frames
GPS	Global Positioning System
GT	Ground Truth
GUO	Global Unique Object
HTTP	Hypertext Transport Protocol
IA3	International Imaging Industry Association
IEC	International Electrotechnical Commission
IETF	Internet Engineering Task Force
iLids	Imagery Library for Intelligent Detection Systems
IPMP	Intellectual Property Rights Management
ISO	International Organization for Standardization
IST	Information Society Technologies
ITU	International Telecommunication Union
JEITA	Japan Electronics and Information Technology Industries Association
JPEG	Joint Photographic Experts Group
KB	Kilobyte
KL	Kullback-Leibler
KLV	Key-Length-Value
LA	Licensing Authority

LLO	Low Level Object
MDS	Multimedia Description Scheme
METS	Metadata Encoding and Transmission Standard
MISB	Motion Imagery Standards Board
MISP	Motion Imagery Standard Profile
MM	Multimedia
MOD	Ministry of Defence
MP2	MPEG-1 Audio Layer 2
MP3	MPEG-1 Audio Layer 3
MPEG	Motion Picture Experts Group
MPEG-7	Multimedia Content Description Interface
MPQF	MPEG Query Format
MXF	Material eXchange Format
NAL	Network Abstraction Layer
NATO	North Atlantic Treaty Organization
NIST	National Institute of Standards and Technologies
NRF	National Retail Federation
OCR	Optical Character Recognition
OWL	Web Ontology Language
PCA	Principle Components Analysis
PCM	Pulse-code Modulation
PETS	Performance Evaluation of Tracking and Surveillance
PGC	Program Chains
PPM	Perspective Projection Matrix
PTZ	Pan-tilt-zoom
QBA	Query by Attribute
QBE	Query by Example
QBS	Query by Symbol
QCIF	Quarter CIF
RDF	Resource Description Framework
REL	Rights Expression Language
RFID	Radio-frequency Identification
ROI	Region of Interest
RSS	Really Simple Syndication
SDK	Software Development Kit

SERKET	Security Keeps Threats away
SMP	Simple Metadata Profile
SMPTE	Society of Motion Picture and Television Engineers
SQL	Structured Query Language
STANAG	Standardisation Agreements
TCM	Town Centre Management
TIFF	Tagged Image File Format
TRECVID	Text Retrieval Conference (TREC) Video Retrieval Evaluation
TS	Transport Stream
TV	Television
UAV	Unmanned Aerial Vehicle
UDP	User Description Profile
URI	Uniform Resource Identifier
UTC	Coordinated Universal Time
UUID	Universally Unique Identifier
VEML	Video Event Markup Language
VERL	Video Event Representation Language
VHS	Video Home System
ViDE	Video Development Initiative
ViPER-GT	Video Performance Evaluation Resource – Ground Truth
ViPER-PE	Video Performance Evaluation Resource – Performance Evaluation
VISCA	Video System Control Architecture
VISP	Video Imagery Standard Profile
VITC	Vertical Interval Time Code
VSAF	Video surveillance application format
VSCDI	Visual Surveillance Content Description Interface
W3C	World Wide Web Consortium
XM	Experimentation Model
XML	Extensible Markup Language
XMP	Extensible Meta-data Platform

List of Figures

3.1	CCTV-derived knowledge: Generation, Encoding and Retrieval. . .	43
3.2	The top-level structure of the ISO Base Media File Format. The ‘meta’ boxes can contain XML meta-data. The ‘mdat’ boxes contain the binary data.	76
4.1	Example data showing the same person, from the two different cameras, each with two views. From left to right: A^{in} , B^{in} , B^{out} and A^{out}	87
4.2	The 15 participants: manually segmented (top); and automatically segmented (bottom).	90
4.3	Example of automatic (left) and manual splitting (right) to produce Top and Bottom data.	91
4.4	Retrieval Rate for Experiment I (same camera, different direction of movement) with automatically segmented foreground data. The names of the colour descriptors are stylised for brevity: Whole is the complete motion mask, Top and Bottom are the ‘top’ and ‘bottom’ mask portions respectively. Sum, Product, Min and Max are calculated from the ANMRR values from the Top and Bottom mask regions.	94
4.5	Retrieval Rate for Experiment I (same camera, different direction of movement) with manual segmentation with automatic split into Top and Bottom clothing (top) and manual segmentation with manual segmentation of Top and Bottom clothing (bottom). The most effective descriptor in this scenario is the ColorStructure DS (top). Compare with the results for automatic segmentation. . . .	96

- 4.6 Retrieval Rate for Experiment II (different camera, same direction of motion) (top) and Experiment III (different camera, different direction of motion) (bottom). Both are automatically segmented and preprocessed for colour constancy. Data without colour constancy is provided for comparison. 97
- 4.7 Calculating the height descriptor: the sample image with calibration guides (top); the calibration ratio between pixels and centimetres (px/cm) shown with calibration points (blue blobs) and the interpolated data as a line (middle); the manually created occlusion mask (inverted for clarity) (bottom-left); examples of the discarded observations (bottom-middle and bottom-right). 100
- 4.8 Left: Example query subject. Right: Six sub-images comprising object data set. 101
- 4.9 The frequency distributions for True & False matches (top) as a function of DominantColor DS (left), ScalableColor DS (right). The probability distributions for these give the probability densities conditioned on True and False matches in Gaussian form (2nd row). The Bayesian probability functions are shown with an equal prior (3rd row) and one example of an unequal prior (bottom). . . 104
- 4.10 ANMRR results: single camera (top); dual camera (bottom). . . . 106
- 4.11 The Information Gain metric on individual features with the validity mask (top). On combined features using a GMM (Gaussian Mixture Model) (middle). The ANMRR metric is computed from the ranked results of the Information Gain metric (bottom). . . . 108
- 4.12 The different feature combinations (left) are listed to demonstrate the correlation between the Information Gain metric (middle) and ANMRR (right). 109
- 4.13 A portion of the GENERICK CAVIAR single camera data set is displayed (top). A closeup of one individual is shown. The second image (bottom) demonstrates the variability between images of an individual. For the same camera data set, the number of images per individual ranges from three to 71. Three is the minimum number for viable ANMRR calculations. 112

4.14	The MPEG-7 colour descriptors including the simple Mean and Random features. The results are from the same camera data (above), the two camera data with Gray World processing (middle) and without (bottom).	114
4.15	The Bayesian probability functions of DominantColor DS (left), ScalableColor DS (right) for True & False matches on CAVIAR data with an equal prior.	115
4.16	The Information Gain metric results on CAVIAR data. The results are shown for MPEG-7 Colour Descriptors and Co-occurrence texture singly (top) and combined (bottom).	116
5.1	The scope of the proposed meta-data is technical (left) and observation (right). User-defined labels and identifiers for equipment and individuals (top).	126
5.2	A Classification Scheme illustrating the three different possible relationships between taxa. The relationships are either <i>type of</i> , <i>part of</i> , or <i>is a</i>	128
5.3	A multi-sensor CCTV system process flowchart.	130
5.4	The camera sensor produces meta-data about the Low Level Objects (LLOs).	131
5.5	The structure of the VSCDI, from the perspective of identity preservation, showing the technical information about the camera and media (top), the LLO (middle) and GUO (bottom) observations.	133
5.6	The File Level (left) and Track Level meta-data (right). Bold denotes the element is required to be included in the description and * denotes multiple descriptions of this type are possible.	140
5.7	An example File Level XML document containing the majority of elements available. The document describes itself, relations and a Classification Scheme.	142
5.8	An example Track Level XML document containing the majority of elements available. The document describes itself, relations, camera calibration information, the video and its frames.	146
6.1	The scene is surveyed by multiple cameras. Relationships are built between the Low Level Objects (LLOs) and the Global Unique Objects (GUOs).	151

6.2	The Overview Ltd. DVIEO [®] video server.	161
6.3	The DVIEO [®] client.	162
6.4	The Query by Attribute colour search interface.	165
6.5	The Overview data set showing moving people (top), corresponding motion masks (middle) and an example of errors in the motion detection (bottom).	167
6.6	The results for the success rate of retrievals on each of the 15 colours are shown (top). The units on the x-axis relate to Table 6.1. The average results for the 15 participants are plotted (bottom). In addition, both graphs plot the <i>agreement</i> data.	170

List of Tables

2.1	CCTV operator target identification code of practice (Norris and Armstrong [NA99]).	35
3.1	Contingency table for Performance and Recall, where N is the number of data set elements (reproduced from [Rij79]).	46
4.1	Example of retrieval scenarios and the Normalised Modified Retrieval Rate (NMRR). Three ‘true’ items can be retrieved; all the rest are ‘false alarms’.	93
4.2	Descriptor combination experiments with Experiment I automatically segmented data, combined with the Min (minimum) operator. The results suggest an improvement over a single descriptor.	98
5.1	List of functional requirements for the meta-data.	123
6.1	The 15 query colours for the colour constancy experiment. Each row comprises a colour label and values in h,s,v , r,g,b and hexadecimal.	169

Contents

1	Introduction	16
1.1	Problem statement	16
1.2	Contribution to knowledge	17
1.3	Projected benefits	18
1.4	Problem scope	20
1.4.1	Material within the scope of the thesis	20
1.4.2	Material outside the scope of the thesis	20
1.5	Thesis structure	22
2	Analysis of surveillance operations	24
2.1	Introduction	24
2.2	CCTV's origins	25
2.3	Analysis of CCTV applications	28
2.3.1	Transport surveillance	29
2.3.2	Office and retail surveillance	31
2.3.3	Military surveillance	32
2.4	CCTV: the interested parties	33
2.4.1	End-users	34
2.4.2	Operators	36
2.4.3	System integrators	37
2.4.4	Manufacturers	38
2.5	Summary	39
3	Review of supporting technology	41
3.1	Introduction	41
3.2	Content-based image retrieval	42
3.3	Performance evaluation	45

3.3.1	Precision and Recall	46
3.3.2	Receiver Operating Curves	46
3.3.3	ANMRR metric	46
3.3.4	Information Gain metric	48
3.4	Features for use in video surveillance	50
3.4.1	Co-occurrence texture	51
3.4.2	MPEG-7 Colour Description Schemes	52
3.5	Segmentation	56
3.5.1	Compression technology	58
3.6	Multiple camera correspondence	59
3.6.1	Colour calibration	59
3.7	Ontologies	61
3.8	Analysis of key video surveillance applications	62
3.8.1	Visualisation	62
3.8.2	Automatic numberplate recognition	63
3.8.3	Search and retrieval	63
3.8.4	Pattern discovery	64
3.8.5	Automatic alarms	64
3.8.6	Access control and transparency	65
3.9	Existing video surveillance meta-data	65
3.10	Related meta-data standards	68
3.11	MPEG standards	71
3.11.1	MPEG-7 – a summary	72
3.11.2	MPEG-21	75
3.11.3	MPEG-4: (Part 12) – ISO Base Media File Format	75
3.11.4	MPEG-A – MPEG multimedia application format	76
3.12	Media file formats	77
3.12.1	Music player application format	79
3.12.2	Photo player application format	79
3.12.3	Digital Versatile Disc / Digital Video Disc (DVD)	80
3.13	Summary	81
4	Analysis of colour-based meta-data	85
4.1	Introduction	85
4.2	GENERICK data set	86
4.2.1	Limitations of the data set	86

- 4.3 Applying the MPEG-7 Colour Descriptors 87
 - 4.3.1 Experimental design 89
 - 4.3.2 Splitting and combining descriptors 91
 - 4.3.3 Evaluation of retrieval accuracy 92
 - 4.3.4 Results 93
- 4.4 Fusing multiple features 98
 - 4.4.1 Height feature 99
 - 4.4.2 Experimental design and methodology 99
 - 4.4.3 Results 105
- 4.5 CAVIAR data comparison 109
 - 4.5.1 Building the data set 110
 - 4.5.2 Experimental procedure 111
 - 4.5.3 Results 113
- 4.6 Conclusion 117
- 5 The Content Description Interface 120**
 - 5.1 Introduction 120
 - 5.2 User requirements 121
 - 5.2.1 Functional 121
 - 5.2.2 Non-functional 122
 - 5.3 Architecture of the VSCDI 125
 - 5.3.1 Technical meta-data 126
 - 5.3.2 Observation meta-data 127
 - 5.3.3 Classification Schemes 127
 - 5.3.4 Identity preservation meta-data 129
 - 5.3.5 Profiles and levels 132
 - 5.3.6 Structural details 135
 - 5.3.7 File Level meta-data 139
 - 5.3.8 Track Level meta-data 142
 - 5.4 Conclusion 147
- 6 Evaluation of the proposed solution 149**
 - 6.1 Global identifiers for pedestrians 150
 - 6.2 Application of the VSCDI 152
 - 6.2.1 Video surveillance application format: overview 153
 - 6.2.2 CARETAKER evaluation 156

6.2.3	Grand Challenge evaluation	158
6.3	Considerations for an embedded system	160
6.3.1	Embedded digital video server	160
6.3.2	Meta-data format	161
6.3.3	Developing colour search capabilities	163
6.3.4	Evaluating the colour search tool	166
6.3.5	Results	168
6.4	Conclusion	171
7	Conclusions and future work	174
7.1	Summary of achievements	174
7.2	Discussion	175
7.3	Future work	179
7.3.1	Content features	179
7.3.2	Identity preservation	181
7.3.3	Schema extensions	181
A	Video surveillance application format – MPEG-7 profile	183
A.1	Table of contents and Annex B (Schema omitted)	183
B	CCTV technology	199
C	Writing competition entry	201
	Bibliography	205

Chapter 1

Introduction

1.1 Problem statement

The purpose of this thesis is to investigate the most appropriate elements and techniques for the design of a video surveillance content description interface. Work is required to standardise digital video and meta-data whose elements must be able to describe key components of the information contained within, and associated with, video surveillance data.

Part of the process of building such a content description interface involves the definition of the requirements for the meta-data descriptions. A careful analysis of these requirements needs to be undertaken in order to define the scope of these descriptions. In addition to supplying cataloguing information, the aim is for the meta-data to allow some enquiry into the events documented within the media stream.

The components required by a useable meta-data content description interface are investigated with special focus on two specific problems. First comes an analysis examining the capabilities of current colour meta-data descriptors which includes experiments to test their performance in terms of retrieval accuracy in

a search and retrieval application. Second comes the consideration of automatic preservation of the identity of both animate and inanimate objects within a video stream. The problem involves the accommodation of uncertainty in the description of identification and a standard way to present it.

In the surveillance domain standardisation of knowledge description is becoming more important, but a widely used content description interface is lacking. Such an interface must consider the meta-data for the above problems and present them using a syntax, grammar and format. The presentation must be scalable to allow a generality or specificity in the description, as appropriate to the application.

1.2 Contribution to knowledge

The standardisation of knowledge derived from the generation, encoding and retrieval components of CCTV (closed-circuit television) will be useful in a surveillance application. The main contribution this work brings is a new content description interface for CCTV meta-data. The interface comprises two main components. The first is the description of general surveillance application knowledge and the second concerns the presentation of signals processed in the application.

The ‘knowledge description’ components include the following novel elements: the use of a standard Classification Scheme [ISO04a] to describe knowledge about the surveillance application; a probabilistic methodology achieving identity preservation over multiple cameras; the selection of standard meta-data tools to describe the technical and observation information derived from a civilian visual surveillance system.

Three significant components are developed for the signal processing meta-data. First two new data sets are built for training and test validation purposes. Second a novel approach to colour constancy is developed. Third is the application of a framework incorporating uncertainty, used to identify surveyed objects.

1.3 Projected benefits

CCTV is a visualisation tool with greatest potential for use in forensic analysis. The main problem with current systems is a lack of quality control, particularly in the video image and timing information. A useful piece of forensic data is likely to consist of clues to an event's time and place. Forensic scientists value a standard format since a basic level of quality can be enforced. Time details are not often readily available but a meta-data standard could impose the requirement that time-stamps and identifiers must be attached to CCTV video. A reliable time-stamp will also help the data captured by visual surveillance systems to be used as evidence in court. Without the availability of reliable timing cues it is possible for this evidence to be discredited.

One feature of automatic processing is the capability for systems to produce colour feature descriptions from image information. These descriptions are stored as meta-data and can provide colour signatures about objects of interest in a surveillance scene. One problem with surveillance systems arises from the huge quantity of video data generated and the consequential time required for forensic analysis. These colour descriptions are applicable to image search and retrieval and provide an opportunity to speed up this forensic analysis. Standard colour descriptions further enhance this capability, allowing common matching tools to be used and an opportunity for systems to produce and exchange standards compliant meta-data.

Incorporating a notion that the outputs of automated processes can be uncertain increases the robustness of the surveillance system, providing a mechanism to combine the output of different automatic processes to improve the description of the semantics. A single sensor and automatic process is unlikely to be able to reach a concrete semantic conclusion on its own without extra contextual information. Indeed, the output of the process is likely to be fallible, *e.g.* the extraction of moving objects from a surveillance video stream is currently unreliable. When combined with information from other sensors, *e.g.* RFID (Radio-frequency Identification) gate sensors and time information, the chances of maintaining a coherent scene description are improved. This coherency implies the system can track individuals surveyed by multiple sensors more effectively, *e.g.* an individual tracked travelling on a public transport network. Standard meta-data tools are chosen to annotate a measure of uncertainty in the scene description. These same tools have application in other areas, *e.g.* manual annotation or content management.

A user-defined Classification Scheme provides a standard yet customisable scheme allowing the user to choose the characteristics of their application and domain. Since the traditional visual surveillance application is bespoke such flexibility will be welcome. This scene description is readable by standard decoders and contains the details important to the organisation. This customisation can be applied to many descriptions, *e.g.* the names of cameras, security guards, and objects of interest from a surveyed scene, *i.e.* the scene semantics.

1.4 Problem scope

1.4.1 Material within the scope of the thesis

The content described can be divided into two parts. The division is between *technical* and *observation* meta-data. The *technical* information contains equipment descriptions vital to the *observation*. These include time-stamps, equipment identifiers, video coding schemes, equipment names, and equipment settings. Typical camera information includes equipment, location, preset position, control protocol and calibration data. The *observation* information “encodes” semantics useful to a CCTV application. This could be the expression of colour features, the detection of an event or the gathering of data associated with an event.

For the work on colour descriptors, an array of colour feature meta-data is analysed so that a recommendation can be made for those most appropriate for video surveillance. This analysis is extended with a technique incorporating the uncertainty in the output of the feature meta-data and can be used to improve the robustness of automatic scene analysis, with the ability to combine the data from multiple outputs and sensors. Standard meta-data tools are chosen to describe this information and are evaluated, including the work on Classification Schemes which provides the user with a set of standard meta-data tools with which to define a personalised system description.

1.4.2 Material outside the scope of the thesis

Certain topics important to CCTV making *use* of meta-data are not covered by the thesis. Meta-data to handle data-protection, data-integrity, authentication and audit trails are not investigated. The protection of sensitive data against inappropriate use is likely to be based upon digital rights management (DRM)

systems [ISO04e] which lie outside the scope of the thesis.

The details of where and how CCTV cameras are best sited presents an important issue when installing a CCTV system. The thesis considers how such decisions can be described using a Classification Scheme, but does not consider the processes involved in such decisions as, camera viewing angles, appropriate camera types, *etc.* [ARG07].

There are problems related to the potential proximity of the meta-data. A surveillance scene description with a frame-by-frame analysis can become huge especially if it contains a great deal of scene information, *e.g.* regions, colour features and textual annotations. These issues have already been addressed with various binary XML (Extensible Markup Language) formats [ISO05h] and are not investigated.

Network communication is important in the exchange of information. Inter-system communication architectures and applications lie beyond the scope of this work. For example, complex implementations such as CORBA (Common Object Request Broker Architecture) [OMG], as implemented in the European project ADVISOR (Annotated Digital Video for Intelligent Surveillance and Optimised Retrieval) [Ren03, VLS04], or simple Internet delivery tools, *e.g.* RSS (Really Simple Syndication) feeds [RSS]. Standard interfaces for querying databases are not considered. Two are in development: MPEG (Motion Picture Experts Group) are developing the MPEG Query Format (MPQF) (previously known as the MPEG-7 Query Format) [GTD⁺07]; and JPEG (Joint Photographic Experts Group) are developing ‘JPSearch’ [DAE07].

The majority of surveillance cameras are assumed to have limited optical distortion, despite omnidirectional or ‘fish-eye’ cameras being applicable to surveillance. Additionally, no attempt is made to produce a standard communication

protocol for remote control between cameras and control rooms. For video conferences, the standard [ITU94] has potential use in CCTV but is not considered.

1.5 Thesis structure

Chapter 2 provides an overview of the CCTV landscape relevant to the topics introduced in this Chapter. CCTV as a concept is introduced together with a brief history of its technology. The current applications for CCTV are reviewed with a focus on security, transport, retail and the military. The perspectives of the parties interested in CCTV are analysed with respect to a standard content description interface.

Chapter 3 reviews meta-data description technology. Content-based image retrieval is reviewed first in Section 3.2. This review continues with a discussion on performance evaluation in Section 3.3 and features for use in a CBIR application in Section 3.4. A brief overview of segmentation technology is given in Section 3.5 including a review of compression technology. This is followed by an analysis of the technology required by systems which process data from multiple cameras (Section 3.6). Colour-based solutions are emphasised throughout the discussion on these topics. A brief discussion is provided on current uses for ontologies in surveillance (Section 3.7).

The main CCTV applications using meta-data are reviewed in Section 3.8. Various current research and commercial CCTV systems using meta-data are reviewed in Section 3.9. Section 3.10 reviews general image and video meta-data standards, with a review of MPEG (Motion Picture Experts Group) technology in Section 3.11. A review of file formats in Section 3.12 examines the packaging of media and meta-data together and performs a comparative study.

Expanding upon the search and retrieval application, Chapter 4 performs an

analysis of standard MPEG-7 colour features [ISO02c]. The experiments are conducted on bespoke and standard data sets for pedestrian re-identification. Three performance evaluations are completed: the first uses a ranking metric; the second uses a metric derived from Information Theory; and the third evaluates both experimental frameworks on standard video data.

Chapter 5 proposes a surveillance content description interface. The types of content which the meta-data must describe are introduced. A standard description for Classification Schemes for use in video surveillance is discussed. The concept of describing uncertainty in the scene is also discussed, based upon the outputs of the probabilistic framework, described in Chapter 4. An analysis of the requirements for meta-data is divided into two, *technical* and *observation*, at *file level* and *track level*. The MPEG-7 meta-data standard [ISO02a] is proposed to meet the syntactical, grammatical and format requirements and a restricted subset is devised.

Chapter 6 discusses three different evaluation studies performed to verify and validate the proposed content description interface. First is a qualitative assessment of a system designed to preserve the identity of pedestrians as they walk around a surveyed scene. Second is an evaluation of the components of the proposed interface, as used in three visual surveillance projects: the MPEG-A (Part 10) Video surveillance application format reference software [AS08], the CARE-TAKER [IST] surveillance scheme and the Grand Challenge visual surveillance competition [UKG08a]. Third is a description of the implementation of an colour-based image retrieval system developed on prototype CCTV hardware.

The conclusions are presented in Chapter 7. Section 7.1 reviews the main accomplishments of this work. Section 7.2 provides an overall conclusion of the previous Chapters. Section 7.3 reviews the areas where this work could be continued in the future.

Chapter 2

Analysis of surveillance operations

2.1 Introduction

Surveillance is defined as ‘close observation, in order to learn something about the observed’ [SS06], with application in espionage or criminal investigation and used either overtly or covertly. The history and development of surveillance, in terms of CCTV (closed-circuit television), is given here as an introduction to this thesis and includes a review of its main applications, together with their predicted future developments. Particular attention is paid to meta-data use in a CCTV system which describes the knowledge about the system and its scenario. The meta-data must satisfy the needs of a community comprising end-users, operators, system integrators and manufacturers, and a review of these parties with their priorities provides a basis from which to define the scope of the proposed technology.

2.2 CCTV's origins

The following two paragraphs are based upon text by Norris and Armstrong [NA99]. Sekula [Sek86] indicated that large scale surveillance activities were first used in the U.K. in the late 19th century in order to deal with convicts re-offending. This was required after the increase in the number of criminals who remained in Britain, having not been transported. At this time a criminal could re-offend but remain unknown to police because neighbouring districts could not share information effectively. As a consequence Alphonse Bertillion developed a bi-modal statistical technique which could identify a re-offender. He used eleven biometric measurements based upon the nine from the French statistician Queletet's 'average man' and two photographs. He calculated that similarity between individuals was one in four million and arranged the biometric data in clusters of similarity. This method was developed into today's biometrics system of fingerprints, facial photographs and identification numbers.

Criminal surveillance was not, however, the only motivation for increased surveillance. In modern times, general civilian surveillance is required by the state. In the U.K. after the Industrial Revolution the advent of railways and motor vehicles enabled people to travel far afield thus requiring a national knowledge bank of information, *e.g.* passports, qualifications, criminal records and vehicle registrations. Previously this knowledge was maintained at a local level in the individual's community.

CCTV systems have been in use since the 1940's as remote monitoring devices for hazardous industrial conditions, *e.g.* rocket launches [Abr87]. Video recording became viable in the 1960's and consumer-based CCTV systems were introduced in the 1970's. Today we can have systems with thousands of cameras. The video is fed to a central operation room where the activities of vehicles and pedestrians

can be monitored online and police can be directed to incidents. To save costs it is common practice to *multiplex* the video feeds from multiple cameras to a single media stream. The video is mostly analogue and transmitted via cable or a microwave link over far distances. A form of early *encoded* meta-data is the time-stamp, which is written onto the video picture, still a common technique used today. Another form of simple meta-data is the cassette label, or file identifier, which provides an index key. Second generation systems digitise the data onto hard disk drives, offering improvements over video cassettes in terms of cost, quality, reproduction, flexibility and data retrieval. The negative aspects are an increased variation in video coding formats with consequential problems in exchanging information. Digital systems make use of meta-data within databases: time-stamps; image and video indexing; motion information; events and alarms. The motion information is generated simply from pixel differences between neighbouring frames. Third generation systems may deploy ‘edge-based’ and ‘back-end’ processing, *i.e.* processing at the camera using a DSP (Digital Signal Processor), and processing at the CPU (central processing unit), respectively. In general, a DSP processes images at a low-level and the CPU is a metaphor for higher-level processing such as required by a user-interface. The meta-data features of CCTV equipment is increasing, as automatic processing becomes more powerful. For example, a camera developed by Sony based upon the DEPA (Distributed Enhanced Processing Architecture) platform streams events containing object motion meta-data [Pro06c] and stores the video locally for forensic analysis. Appendix B provides more information on CCTV cameras.

The following two paragraphs are again based upon text by Norris and Armstrong [NA99]. The history of CCTV in the U.K. began in the 1950’s, when police forces made limited use of cameras to monitor the public. In the 1960’s CCTV systems were deployed in banks, when Photo-Scan introduced the first

commercial system [Goo04]. The London Underground introduced CCTV for crime prevention in 1975, and since then high-street stores and shopping centres have made increasing use of CCTV. A U.K. initiative called Town Centre Management (TCM) allows retailers, the police and local governments to share CCTV resources addressing problems of expense, duplication and accessibility to the video footage. A shared control room where operators watch banks of monitors is the main product of this arrangement. The police have priority over the operators and can demand control of the cameras when necessary.

Another factor that has increased CCTV usage in the U.K. is the 1993 bomb in Bishopgate, London. For future protection, the City of London Authorities created a 'Traffic and Environment Zone', known as the 'Ring of Steel', which is the origin of the London Congestion Charge. All entry points are monitored by a network of CCTV cameras. Automatic numberplate recognition (ANPR) systems check the identity of all vehicles entering and leaving the zone. Also launched at this time and subsequently re-launched in 2007 was the 'CameraWatch' scheme [UKG07b]. Originally the scheme allowed access to a shared CCTV infrastructure and provided advice on new system installations. Additionally the Closed Circuit Television Challenge Competition in 1996/7 funded the use of CCTV in schools, town centres and council estates. Furthermore in the course of high-profile court cases in the 1980's and 1990's CCTV images were broadcast on national television. Various studies have analysed the actual impact which CCTV has had on crime. One study investigates the demographic effects of CCTV [Til93], while another investigates the effectiveness of CCTV as a tool in criminal apprehension [BHN01].

2.3 Analysis of CCTV applications

Some specialised applications are outlined below with each discussion concentrating upon the themes of ‘crime prevention’, ‘forensics’ and ‘customer analysis’. If these categories are relevant to the scenario the online and forensic processing requirements are considered.

Today CCTV is used in various application domains and the main scenario is crime prevention through visualising remote activity for real-time and forensic analysis. Systems installed in public places monitor pedestrians and traffic where the data is manually processed [UKG06]. In general, public systems will have less budgetary restrictions than private systems and are more likely to make use of cutting edge technology, *e.g.* military or police forces. Private systems, on the other hand, usually comprise cost-effective monitoring devices with limited processing power.

Presently the CCTV industry has problems where too few operators monitor too many cameras, and with systems that are non-standard making sharing video data difficult. Increasing the use of automated processes can help this situation and one application is the generation of alarms which can alert operators to anomalous activity. Research in visual surveillance is focused around automatically extracting objects from the surveillance video [TKBM99, GROJ08, HTWM04], automatic behaviour detection [CBT03], content-based image retrieval (CBIR) [BPB03], neural networks monitor exclusion zones [Pro08b] and complete surveillance systems [HBC⁺04]. The techniques used to deliver surveillance data are being improved and the state-of-the-art of IP and wireless technology can be used [AWW05].

Automated processes make use of meta-data and complex processes generate

abstract information which must be handled in a structured way. A simple process, such as the output of a movement detection process, could be described as a number or a binary yes or no. A more complex process, such as an *Alert*, might require an annotation comprising an *Event* and various description fields, *e.g.* a ‘Suspicious’ Alert with a *Person* Object causing a *Running* Event. The Object and the Event are coupled and have a corresponding sequence of video frames over the time, *e.g.* trajectory information. Colour information could be added and requires a compact yet powerful description tool. Increasing the processing power of devices allows the generation and storage of this video annotation. Complex processes require a sophisticated meta-data encoding scheme, *e.g.* XML.

2.3.1 Transport surveillance

Public and private traffic surveillance systems are deployed by police, local governments, government agencies and companies. Passive cameras, which only record video and do not receive telemetry or transmit video, are installed on buses and trains purely for forensic analysis. High-volume transport termini, *e.g.* railway stations and airports, are vulnerable to sabotage and have sophisticated camera networks [Whe02]. The conjoining of CCTV systems improves security, *e.g.* the CCTV system at Arsenal Football Club in the U.K. is connected to the local transport hubs [Pro06b] and the flow of pedestrians between transport termini and the stadium are monitored. Again, in the football domain, CCTV is a commonly used tool used to address hooliganism. Controlled access-points, *e.g.* passport control and security screening, deploy more intense surveillance measures. At passport control a camera records an image of a passenger’s face for later forensic analysis, if necessary. X-ray machines inspect luggage, iris scanners allow fast-track passport control. Millimetre-wave X-ray cameras are used

to screen people [BBC03]. These are semi-autonomous signal processing systems designed to detect the presence of danger. An operator then makes a decision on the reality of the threat's potential. For forensic analysis U.K. airport car parks record the registration mark of vehicles and the driver's image on entry to and exit from the system. Car parks use CCTV as a security service which is provided to customers for the safety of themselves and their vehicles [Til93].

In the U.K. police use an in-car CCTV system called ProVIDA, made by JAI Ltd., with both front and rear-facing cameras [Pro08d]. Incidents can be recorded and the video footage can be used as evidence in the event of litigation. Automated online vehicle identity checking is carried out by cameras equipped with ANPR (Automatic numberplate recognition) from within police vehicles. These automated processes use meta-data to associate image data with the ANPR numberplate information.

In the U.S.A. Motorola have installed digital IP (Internet Protocol) systems in patrol vehicles and these work in a similar way to the ProVIDA system. They also provide wireless multi-point communication for the transmission of live video data between mobile units [AWW05], operators and commanders. This type of system is expensive in terms of design, equipment and operating costs, due to the complexity of the systems deployed within each vehicle.

ANPR is not possible on many standard CCTV cameras for both technical and economic reasons [Pro06a], so the Transport Authorities aim to make use of computer vision technology for the purpose of vehicle classification [GMMP02, ROJ02] and anomalous event detection [ROJ02], *e.g.* lane contravention. The system meta-data would need to describe the type of event perhaps with reference to an ontology describing vehicle types and vehicle behaviour.

CCTV applications in underground railway networks [CBT03] and airports

[BTF⁺05] can benefit from real-time alerts. Behaviour recognition in both individuals and crowds [ABF06] could also be applied. The system meta-data describes the types of behaviour, the objects and their movement. Perhaps an ontology is referenced, describing the application, the scene and how the objects interrelate. Private car manufacturers are also making increased use of CCTV systems for blind-spot viewing and pedestrian avoidance systems [GM07].

The specification and capabilities of a standard format is limited by its intended application and budget. It is therefore difficult to design a format which is suitable for all domains.

2.3.2 Office and retail surveillance

Retailers use CCTV to help reduce revenue loss in their merchandising operations. Clerical error, misplaced stock, shoplifting, employee theft, theft by supplier, returns fraud, tag switching and *sweet-hearting* – where a customer is permitted to steal by an employee – are all channels of revenue loss [SBH⁺07]. Office surveillance, although uncommon in the U.K., is widespread in the U.S.A. where CCTV cameras are used by employers to survey employees and provide operations-based surveillance. For example, a camera-based room occupancy detector is on the market manufactured by ObjectVideo Inc. [Pro08e].

The cameras are placed in strategic places, *e.g.* above a cash register, and record activity in order to detect anomalous events. The store's point-of-sale system can be linked to the CCTV system allowing the invoice number to be associated with video of the transaction. Store-detectives watch for certain behaviour patterns in the shoppers, *e.g.* 'concealment' behaviour, and use an audio-link to communication with the operator.

Casinos make use of a high density of specialised cameras [Wis06]. These

cameras require a high-frame rate and are positioned to capture many aspects of a gambler's behaviour. In this way card cheats can be foiled even if their quick movements can deceive the croupiers, so the operator must be especially attentive. The casino may have a database of known cheaters which can be relayed to the doormen.

The CCTV market is developing through the isolation of different elements within the system into components. Retailers are interested in maximising the return on their investment and the addition of automated market analysis components is desirable, *e.g.* customer behaviour patterns [SBH⁺07]. Components exist which produce alarms based upon events and in future could include behaviour detection or face recognition. These modules form part of a chain of analyses where the different processes are split into core competencies, *e.g.* object tracking algorithms generate object and track meta-data which feed into systems generating object colour and shapes meta-data for re-identification [BPB03]. Higher-level surveillance and business processes analyse this output and provide forensic and market analysis data, respectively.

2.3.3 Military surveillance

The military make extensive use of surveillance in their activities. Specialist military applications for CCTV technology exist in many spheres and the online security web site Army Technology provides many examples [Pro08c]. These include aerial surveillance, battle field surveillance, bomb inspection tools, night-vision systems, diagnostic tools and weapon aiming assistance. Sensitive property, *e.g.* government buildings or military bases also use CCTV to help secure perimeter fences.

The military have demanding performance requirements for their technology

and their standard specifications are of particular interest. Standards are used to enforce a detailed specification and remove implementation errors or ambiguity, and promote the exchange of information. These criteria are especially important in a theatre of war. Without such specifications serious problems could result, *e.g.* if different timing information is used, equipment could malfunction. NATO (North Atlantic Treaty Organization) specifies the STANAG (Standardisation Agreement) to define the scope of an application and address the implementation issues.

A considerable component of these specifications is the maintenance of compatibility with legacy systems, *e.g.* digital systems communicating with analogue systems. It is not possible to simply replace aging legacy equipment due to budget constraints. For example, STANAG 4609 : Digital Motion Imagery Standard [NAT07b] implementation guide [NAT07a], specifies how legacy ‘closed-caption’ meta-data is converted to the SMPTE (Society of Motion Picture and Television Engineers) KLV (Key-Length-Value) [SMP01c] meta-data format. This reflects the value of the image data where the time-stamp is not superimposed on the images. In the future, analogue systems are likely to disappear but support for legacy specifications will always be an important factor.

2.4 CCTV: the interested parties

Complementary to the application domains described above, CCTV is associated with various independent parties. More commonly known as *stake-holders*, these parties have an interest in its technology. The following text introduces each party with a statement about their interests and how these relate to the development of standards in CCTV, in particular meta-data standardisation.

2.4.1 End-users

End-users, otherwise known as *consumers*, are the people who use CCTV systems as a tool in their work. Above all, they want a system that fulfils their expectations within the boundaries of a budget and satisfies the requirements particular to an application. The largest single cost of operating a CCTV system arises from employing the operators [NA99]. Presently, the operators form part of a less than optimal manual process. Currently the systems installed are bespoke solutions with associated peculiarities, *e.g.* the Rome underground taxonomy for camera labelling is based upon a non-consecutive station numbering scheme. Compliance with government regulations is a priority for the end-user and changing a system to comply with new regulations could be expensive. Published recommended industry guidelines exist, and include the code of practice for public CCTV systems [UKG08b], and operation room specifications [WD98].

Modules are becoming available carrying out some automatic processing although these have proprietary interfaces where individual companies specialise in particular areas, as shown in Section 2.3.2. The end-user would like to increase the value of their system and installing a module which provides new functionality can achieve this. While the best module can be selected it may not be compatible with the existing system. A standard format between software video analysis components will overcome compatibility problems. A standard meta-data interface supporting user-defined labelling and Classification Schemes will allow esoteric naming conventions. This will facilitate compliance with government regulations and data access in the event of an audit, inspection or crime.

1.	Given the sheer volume of candidates for targeted surveillance, the operators utilise their already existing understanding of who is most likely to commit crime or be troublesome to provide potential candidates for targeted surveillance.
2.	Certain behaviour unquestionably warrants surveillance because it is criminal or disorderly. However, there is a range of other actions may not be criminal but operators treat as indicative of potential or recently occurring criminality.
3.	Certain people are immediately worthy of surveillance because they are known by operators to have engaged in criminal or troublesome behaviour in the past.
4.	Operators must learn to treat locales as territories of normal appearances and against background variation can be noticed. This involves utilising temporal and spatial variations of activities within a locale to judge what is both 'out of place' and 'out of time'.
5.	For operators the normal ecology of an area is also a 'normative ecology' and thus people who don't belong are treated as 'other' and subject to treatment as such.
6.	There is an expectation that just as operators treat territories as a set of normal appearances, so others are expected to treat them as such. And thus if a person appears lost, disoriented, or in other ways at unease with the locale, this will seem suspicious.
7.	Operators learn to pick out those who treat the presence of the cameras as other than normal.

Table 2.1: CCTV operator target identification code of practice (Norris and Armstrong [NA99]).

2.4.2 Operators

Operators define the real-time monitoring capabilities of a CCTV system within the CCTV Control Room. The operator overlooks a bank of monitors outputting the video from many cameras and in order to achieve adequate supervision, an operator can only work continuously for a short period of time. From managerial perspective their work programme is described as the seven point plan [NA99] shown in Table 2.1. Although not official it provides a starting point for analysis of the skills required by an operator. The first and third items can be interpreted as requiring experience or training. The use of a visual database of suspects and known offenders could be useful here. Items two, four, five, six and seven are similar, describing how the anomalous individual is recognised. The word ‘normal’ has a subjective meaning and may be problematic under cross-examination in a court of law. The operators also communicate with the police, store-detectives and members of the public. The latter can demand to see the footage of a CCTV system, if they are likely to have been captured on it. In this situation, the operator is required to mask other personally identifiable data, before allowing external viewing.

The addition of automated processes to detect anomalous behaviour could increase the time period an operator can work at full capacity. Instead of dividing up attention over many screens, time could be devoted to a single screen, where an alert has attracted the operator’s attention. The details of the alert could be contained within a meta-data structure and carefully designed meta-data could provide a clearer definition of what caused the alert. The language available for the objects and semantics described by the Operator could be defined within a Classification Scheme, similar to the police lexicon. The police are trained to use a specific lexicon in the course of their work and removing ambiguity from their

dialogue is important a) for appropriate communication with the public and b) to avoid problems during police investigations or litigation. A Classification Scheme can be autonomously processed which could remove ambiguity within software, *e.g.* database fields and user-interface labels. Tightly defined terms are important criterion for computer processes where computers can only understand what they are programmed. An operator must use carefully chosen or tightly controlled terms for successful processing, *e.g.* using the search term “find me all cars between 10am-11am” would return nothing if the system expects “vehicles” instead of “cars”. Furthermore, the use of meta-data can assist in protecting personal privacy by easing the process of masking out sensitive information. In this case, the meta-data would play a role in manual and automatic methods to render the content anonymous, *e.g.* by pixel blurring. The company Eptascope produces cameras that produce a stream of MPEG-7 [ISO02a] meta-data alongside the video [Sac07] to increase the range of functions, *e.g.* the description of people’s whereabouts to enable automatic pixel masking methods. In this example, it is the presentation that is altered, and not the original video data. A mechanism would be required to ensure that this situation is not abused, *e.g.* restricting access to the video.

2.4.3 System integrators

System integrators install a CCTV system based upon the requirements of the end-user. The best solution depends upon cost and performance and, if appropriate, how well the new system integrates with the old.

In order to accommodate the end-user’s requirements, the application, its surroundings and the end-user’s characteristics are investigated. The equipment is chosen, which could be a PTZ (pan-tilt-zoom) camera or a fixed camera. Fixed

cameras are effective for constant scene monitoring as they cannot be moved out of position. A casino requires high frame rate cameras and these in turn need high-performance data capture systems. The characteristics of the building may require additional lighting. The system is built from preferred components and suppliers. The result is proprietary with potentially limited ability to exchange information with different systems and components. Presently, the police often have to remove the video storage system from the site in order to process it.

A shared meta-data format can give the system integrator more scope when choosing the components for the system. Indeed this standardisation requirement will increase as products make more use of meta-data. A degree of compatibility between systems will make replacing obsolete hardware easier, if a clear upgrade path is defined and upgrades are carried out frequently since specifications change quickly in the digital domain. Compatibility will also benefit the system integrator since they can reuse existing knowledge.

2.4.4 Manufacturers

Manufacturers produce the components of a CCTV system. These include cameras, video recorders, monitors, cabling, operator suites, and software. The products have a development cycle and the manufacturer will have a particular business model. The system integrator is likely to be linked with a particular manufacturer. To achieve competitiveness manufacturers continually improve their products, react to market forces, change their business practices and rely upon feedback from their customers. External catalysts will also affect change, *e.g.* governmental legislation.

Compatibility problems arise from the different techniques used in equipment design. Digital video recorders may use different multiplexing techniques and

coding schemes. Incidentally, serial full-frame video multiplexing will exacerbate the generation of digital artifacts if inter-frame video compression is used.

Adherence to standards can improve the success of a commercial product by broadening its market and appeal. The most desirable situation for a company is that its product becomes the standard, *e.g.* Microsoft Windows. The next best solution is for the company to incorporate its patented intellectual property within a standard. Governments and customers view standards compliancy as a quality attribute and the adoption of standard techniques can simplify the development process, *e.g.* MPEG video coding schemes are widely used.

2.5 Summary

CCTV is a remote visualisation tool with its roots in both industrial inspection and surveillance. For surveillance CCTV is used to provide a deterrent, real-time reaction and forensic analysis. Well-defined automated computer vision processes have become wide-spread, *e.g.* ANPR. The latest computer processing research has significant potential to improve the capabilities of current systems. Other than security, the main applications are transport, retail and the military. In the civilian domain, the research shows the retail industry has the most potential for using the state-of-the-art in computer processing.

Meta-data is a significant component of existing analogue visualisation systems. As systems become more sophisticated automated processes within them will generate more complex meta-data likely to describe the video content. The use and potential for a standard content description interface will simplify the process of developing meta-data components used within these algorithms. The interface should be loosely coupled to the content it describes.

CCTV system end-users need compatible and customisable meta-data within

their systems. Increased use of meta-data can help operators to define a clearer role in their capacity as the system interpreter. System integrators need standard meta-data to ease system installation and manufacturers can benefit by adhering to standards.

Chapter 3

Review of supporting technology

3.1 Introduction

This Chapter reviews the specifics of current meta-data implementation, standardisation, deployment and research. It is important to analyse how meta-data can be and is used in a CCTV (closed-circuit television) system. First a review of content-based image retrieval (CBIR) examines the technology required to generate, retrieve and evaluate content-based image signatures. Second the important topics of segmentation, multiple camera correspondence and ontologies are reviewed.

The roles meta-data plays in a surveillance system are divided into six applications. The existing meta-data used in CCTV industry and research is investigated. An overview is provided of various mature image and video meta-data standards outside the surveillance domain. An introduction is provided to the standards body MPEG (Motion Picture Experts Group) which produces suites of standards dedicated to audio and video coding. This Chapter also reviews a selection of file formats which store media and concomitant meta-data.

3.2 Content-based image retrieval

Content-based retrieval (CBR) is an established technology with applications in many specific domains since the 1970's, *e.g.* entertainment, natural history and sport [HSD73]. CBR is the technique of indexing images for retrieval based upon signal content, rather than external attributes, *e.g.* date, location or title. The descriptions are created and stored alongside the original signal, to facilitate rapid and effective querying. These descriptions are known as *features*. In the image and video context the signal content is the image data which is commonly a raster scan of eight bit values in a colour space, *e.g.* r, g, b or y, u, v . The features can be encoded from colours, shapes and textures, *etc.*

The first query type is simply an alternative numerical description of the data, *e.g.* Fourier coefficients, or (h, s, v) values. This type of content-based query is *by example* (QBE), in which stored data signals are compared with a query data signal in order to retrieve objects similar to the query subject [Eid00]. For instance, the query image attributes are extracted and compared against the stored set of images. In this type of query there are no constraints on the attribute type of structure: they are 'hidden variables' that are private to any given CBIR system. The diagram depicted in Figure 3.1 demonstrates this retrieval scenario in terms of the GENERICK (Generation, Encoding and Retrieval of CCTV-derived Knowledge) process. GENERICK is described as follows: the Generation process captures video from a sensor which enters the system from the top; the Encoding process encodes the video to a coding scheme and stores and indexes it. This process also produces meta-data descriptions, which include signatures for content-based retrieval; the Retrieval process involves an operator who submits a query, perhaps an image, and based upon the characteristics of the query, a ranked list of similar videos is returned via a user-interface.

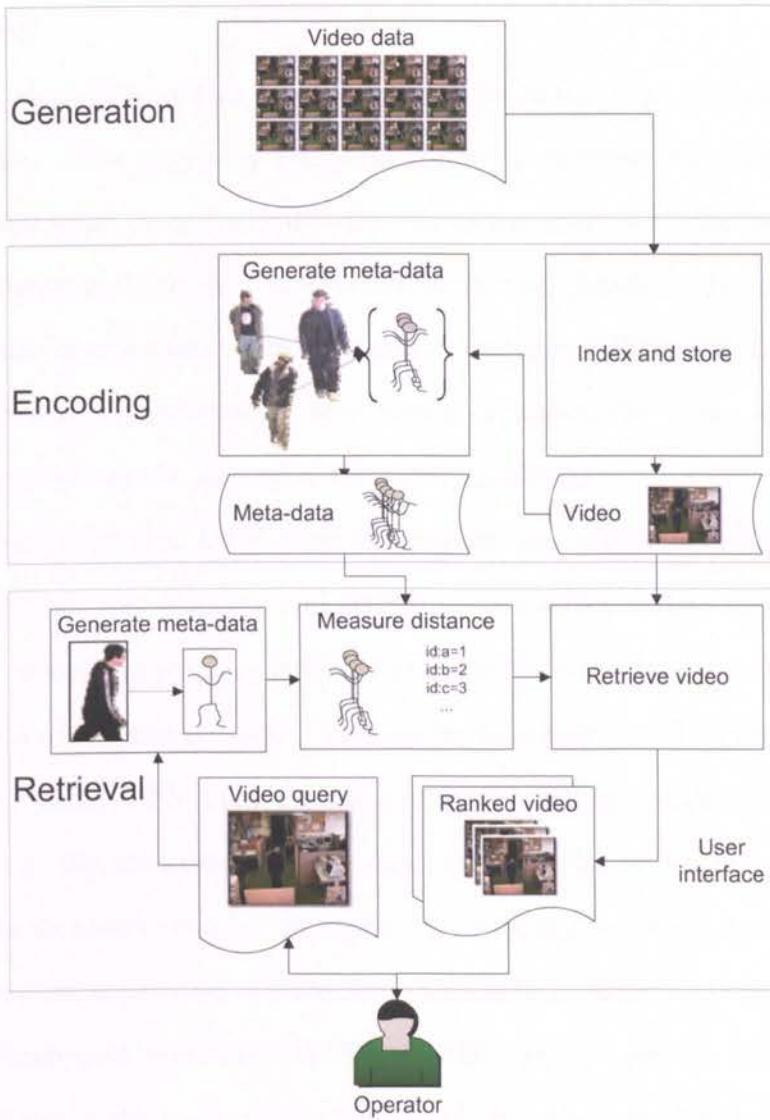


Figure 3.1: CCTV-derived knowledge: Generation, Encoding and Retrieval.

Secondly, a variant of the QBE is a query consisting entirely of *attributes i.e. query by attribute* (QBA). An instance of this type of query could be: ‘Please retrieve images of white shoes or red hats’, where red is a colour selected from a colour palette. A feature is then generated from the attribute ‘white’ and the attribute ‘red’.

Thirdly, descriptions can carry semantic meaning, *e.g.* wooden texture or bicycle shape. This query *by semantic* (QBS) is without any data signal and requires a semantic layer to bridge the ‘semantic gap’, *i.e.* the lack of a clear connection between high- and low-level information [Bim99]. The process turns pixel data into descriptions useful to human operators [MWLG02, DCI04]. Other methods include a preprocessing step which organises the video sequences into classes dependent on the nature of the scene [MSS02].

It is conjectured that QBE does not require any public ontology for its successful execution: any features and structures manufactured in order to retrieve other observations of the query subject can be private constructions of the process that makes and compares them. This seems to apply to all examples of QBE. On the other hand, QBS must use some explicit ontology for defining a semantic condition, *e.g.* the predicate ‘wears a cap’, and its relation to the source data.

A further technique can be employed by informing the system about how well it has performed, supplying outside information to improve a subsequent search, known as Relevance Feedback [WP04, DD03]. In the process of developing a CBR algorithm, a data set will be first used to train and then to test retrieval performance. A division is usually made between training and test data due to noted performance differentials when the two share data [RH99].

3.3 Performance evaluation

When a search and retrieval command is executed a list of answers will be returned. These results can be taken from the researcher's perspective: to assess the performance of the technology; or from an end-users perspective: to reduce the amount of manual searching required. A decision as to whether the returned images are correct can be made in the following ways:

- Binary – The first or top of the list is returned as either a correct match or not. This could be the closest of n items and a threshold is defined to give the binary decision. This is suitable for applications where uncertainty is unacceptable.
- Ranking – A list of the n closest items is returned. This can allow the operator to choose and assess the results. Its use is beneficial when incorrect images are likely to be returned, *i.e.* *false positives*. There is no estimate of whether the images returned are valid. Imagine an example where a query subject is wearing white clothes and compared with a data set containing 99 people wearing black and one wearing grey. A distance metric using a white query image would rank the grey colour highest.
- Probability – an indication of the likelihood that a successful match has been achieved. In general a ranking metric has shortcomings because it only provides a relative performance measure. Ranked matches contain no guarantee that a data set and hence result contains an instance of the searched individual, *i.e.* the search space could be within an *open-world*. If the ranking includes a likelihood of the correct match the CBIR can be applied in an Open World. The likelihood measure is derived from previously observed data.

	Relevant	Non-relevant	
Retrieved	$A \cap B$	$A \cap \bar{B}$	B
Not Retrieved	$A \cap \bar{B}$	$\bar{A} \cap \bar{B}$	\bar{B}
	A	A	N

Table 3.1: Contingency table for Performance and Recall, where N is the number of data set elements (reproduced from [Rij79]).

3.3.1 Precision and Recall

For the researcher it is necessary to evaluate the performance of the system or algorithm objectively. The technique of *Precision and Recall* (PR) is a commonly used in information retrieval [Rij79]. The *Precision* is the measure of how many of the articles returned are relevant and *Recall* measures the proportion of relevant articles returned. Based upon the contingency table, shown in Table 3.1, suitable criteria need to be devised for the experiment.

3.3.2 Receiver Operating Curves

Receiver Operating Curves (ROC) are widely used [PF97] and plot true positives against false positives in order to evaluate binary decision algorithms. ROC curves have similar characteristics to PR [DG06] and are used to evaluate segmentation and tracking algorithms, which require object-based rather than pixel-based evaluations [LMRMJ08]. It is shown, *idem*, that there can be a difficulty in specifying true negatives making the use of ROC problematic.

3.3.3 ANMRR metric

The *Average Mean Normalised Retrieval Rate* (ANMRR) [MSS02, WP04] is a widely used [NNRM⁺00] ranking metric combining the Precision and Recall metrics. The purpose of the metric is to allow an evaluation of different descriptors unbiased with respect to different sample and Ground Truth (GT) sizes, where

GT means the number of true matches in the data set. The metric correlates well with perceptual judgment about the retrieval success rate. Scores are based upon the position of results and not their distance value. The position of each correctly retrieved datum is counted and penalties are issued if any of the items come after a threshold, K . The same penalty applies to all items after K , *i.e.* the procedure penalises low-ranking GT items no matter how low-ranking. The size of the GT set determines the rank at which the threshold is placed. The rule of thumb is for K to be set at twice the number of correct items in the data set. The steps to calculate ANMRR [MSS02] are as follows:

First the rank is generated:

$$\mathbf{Rank}(k) = \begin{cases} \mathbf{Rank}(k) & \text{if } \mathbf{Rank}(k) \leq K(z) \\ 1.25 \cdot k(z) & \text{if } \mathbf{Rank}(k) > K(z) \end{cases} \quad (3.1)$$

Where k is:

$$K(q) = \min\{4 \cdot NG(z), 2 \cdot \max[NG(z), \forall z]\} \quad (3.2)$$

For a particular query z the Average Rank from Equation 3.1 is generated:

$$\mathbf{AVR}(z) = \frac{1}{NG(z)} \sum_{k=1}^{NG(z)} \mathbf{Rank} \cdot (k) \quad (3.3)$$

To deal with varying GT size, we generate the Modified Retrieval Rank:

$$\mathbf{MRR}(z) = \mathbf{AVR}(z) - 0.5 \cdot [1 + NG(z)] \quad (3.4)$$

Each retrieval operation z is assigned an NMRR, the Normalised Modified Retrieval Rate:

$$\mathbf{NMRR}(z) = \frac{\mathbf{MRR}(z)}{1.25 \cdot K(z) - 0.5 \cdot [1 + n_1(z)]} \quad (3.5)$$

Where K = relevant rank mark, n_1 = number of correct data elements and z = query. This is averaged over all operations in the set to obtain the ANMRR.

$$\text{ANMRR} = \frac{1}{NZ} \sum_{z=1}^{NZ} \text{NMRR}(z) \quad (3.6)$$

3.3.4 Information Gain metric

The Information Gain metric addresses the problems of ranking metrics in that it can work in an Open World. It uses the probability of correct retrieval to give a general and transferable evaluation measure. The measure also has the ability to combine descriptors with a probabilistic framework.

In Information Theory [CT91], *entropy* is the measure of the amount of information contained within a message source and is used to calculate the amount of bandwidth required to successfully transmit a message. Claude Shannon showed the maximum entropy H of a message source is equal to $\log_2 n$, if the message length is one unit. 2^H is then the average probability of ‘guessing’ the next message in a sequence.

Generally speaking prior information influences subsequent decisions, and in this case the decision is about the likelihood of a correct match. This likelihood is a measure of the difference in uncertainty between randomly picking an individual from the data set and using *a priori* information. A reduction in uncertainty is equivalent to a gain in information and known as ‘Information Gain’ [JB02].

In order to measure the Information Gain, the identity of the query subject can be written as a discrete random variable Z that can assume values between z_1 and z_n . An n -dimensional vector \mathbf{x} describes the match measure x between the query and each element in a data set using an arbitrary feature. This has a corresponding continuous random variable \mathbf{X} . The information gained through

observation of the descriptor(s) can be written [CT91] as:

$$I(\mathbf{X}, Z) = \sum_i \int_{\mathbf{x}} p(\mathbf{x}, z_i) \log \frac{p(\mathbf{x}, z_i)}{p(\mathbf{x})p(z_i)} d\mathbf{x} \quad (3.7)$$

$$\approx E \left[\log \frac{p(\mathbf{x}, z_i)}{p(\mathbf{x})p(z_i)} \right] \quad (3.8)$$

where $E[\cdot]$ is the expectation operator ranging over the expected joint input of \mathbf{x} and z_i .

This expression can be evaluated using the two p.d.f.s $p(x_i|y_i)$, where $y_i = 1$ if $Z = z_i$, *i.e.* a correct match, and $y_i = 0$ otherwise, *i.e.* an incorrect match. For any single feature they are one-dimensional p.d.f.s that can be estimated from training sets of true and false matches. For combinations of features it is more appropriate to use a parametric (Gaussian) distribution to estimate the covariance of their match measures. The two terms in Equation 3.7 can be approximately expressed in terms of these p.d.f.s:

$$p(\mathbf{x}) = p(x_1, \dots, x_n) \quad (3.9)$$

$$\approx \prod_i p(x_i) \quad (3.10)$$

$$= \prod_i (p(x_i, 1) + p(x_i, 0)) \quad (3.11)$$

$$= \prod_i \left(\frac{1}{n} p(x_i|1) + \frac{n-1}{n} p(x_i|0) \right) \quad (3.12)$$

provided that n is not too small, and similarly:

$$p(\mathbf{x}, z_i) \approx \frac{1}{n} p(x_i|1) \prod_{j \neq i} \frac{n-1}{n} p(x_j|0) \quad (3.13)$$

3.4 Features for use in video surveillance

A key task for an automated visual surveillance system is the construction of a description of observed scene activity. The content-based image retrieval (CBIR) techniques, introduced in Section 3.2, rely on feature vectors, otherwise known as descriptors, which describe the image content. These features can comprise faces, colours, textures, shape, gait [YWHT04], speed of motion, *etc.* This Section discusses the specification of texture and colour visual features and their matching processes. For a more general review of features for image retrieval see Eidenberger [Eid00].

A feature is a model which describes some characteristic. A simple feature vector could be a colour histogram, or an ellipse, *i.e.* the co-variance of an object's width and height or colour range. A feature could be calculated from an image, sub-image, or image sequence. The performance parameters of a feature are generally bound by its accuracy, compactness, speed of generation, speed of comparison and robustness. Low-level descriptors will invariably be generated automatically.

In order to compare two descriptors a scalar distance between them is normally defined [Bim99]. This match measure can be based upon a Euclidean distance, but the procedure is specific to each feature. Once the distances have been measured a process is required to compare the results, described in Section 3.3. Features can be combined in a single classifier to improve its results. The features can be combined using simple operators, *e.g.* sum, difference and product, or by joining estimates of the probability of a true match, especially if the distances are not directly comparable [ALC⁺06]

3.4.1 Co-occurrence texture

Introduced by Haralick *et al.* [HSD73] and reviewed by Randen [RH99], the Co-occurrence texture feature describes image texture as a statistical measure of tone variation assuming that texture is dependent on tone. The procedure works by accumulating pixel-pair grey-level intensities in different orientations and displacements. The technique is as follows: the image is quantized to reduce size and any sparseness in the data. A co-occurrence matrix, P , is then calculated and the neighbouring pixel-pairs are taken, at a certain displacement ($d = n$), in four orientations; 0° , 45° , 90° and 135° , or in all directions, *i.e.* eight-way.

The co-occurrence matrices are compared against one another using Bhattacharyya's distance and Kullback-Leibler (KL) divergence. These measures provide robust measurements of the distance between two p.d.f.s [Kai67]. They measure the mean and the variance of both the model and the observed data and are used as the match measures for the co-occurrence feature vectors, as shown:

$$d_{Bhat} = \sqrt{1 - \sum_{ij} \sqrt{p_{ij}^1 p_{ij}^2}} \quad (3.14)$$

$$d_{KL} = \sum_{ij} p_{ij}^1 \log_2 \frac{p_{ij}^1}{p_{ij}^2} \quad (3.15)$$

Where p_{ij}^1 is the i^{th} row and j^{th} column of the matrix P^1 . These measures are evaluated in Section 4.4.

Haralick *et al.* [HSD73] also suggest 14 possible methods to summarise the statistics from each co-occurrence matrix. The method using entropy is shown:

$$H = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} p_{ij} \log p_{ij} \quad (3.16)$$

The difference between two statistics is simply their absolute measure.

3.4.2 MPEG-7 Colour Description Schemes

Standards for feature definitions are rare but MPEG-7 [ISO02a] and SQL (Structured Query Language) are exceptions. SQL is a database query language and the extension SQL/MM (Multimedia) includes a histogram descriptor. MPEG-7 is a meta-data description scheme, with applicability to multimedia in general. MPEG-7: (Part 3) Visual [ISO02c] specifies colour Description Schemes (DS) [Cie01, MSS02] to describe colour within an image. They comprise Dominant-Color DS for colour clustering, the histogram-based ScalableColor DS, a histogram with a structuring element in ColorStructure DS, ColorLayout DS which separates the image into a grid and processes each cell with a DCT (discrete cosine transform) and GoFGoPColor (Group of Frames Group of Pictures) DS which is an aggregation of histograms. A benefit of the ScalableColor DS descriptor, against more general histograms, is that features with differing quantization levels can be compared.

MPEG-7 descriptors have been used in surveillance research. In particular, the DominantColor DS and ContourShape DS have been combined in a surveillance retrieval task [BPB03], and several visual descriptors have been combined using a neural-network for surveillance [HKK04]. These papers use the Query by Example (QBE) technique. Once the feature is generated, an additional process is required to measure the distance between the query and the data set.

The following descriptions of the MPEG-7 Colour Descriptors, as paraphrased from Manjunath *et al.* [MSS02], provide an overview of the technology and the processes required to generate and retrieve them. These *informative* methods form only suggested implementations, developed as part of the MPEG-7 Core Experiments, which the MPEG-7 Reference Software (XM) [ISO03c] implements.

DominantColor

This descriptor represents colours in an image region as:

$$F = \{(c_i, p_i, v_i), s\}, \quad (i = 1, 2, \dots, N) \quad (3.17)$$

The image data is clustered in the l, u, v colour space. Up to eight colour clusters can be found. The descriptor specifies the number of clusters, the spatial homogeneity in the image, and for each cluster: the colour; variance and percentage values. The colour values have a resolution of five bits per channel, as do the percentage and spatial homogeneity components. Spatial homogeneity and variance are optional and increase performance and computational requirements. A Connected Components algorithm joins neighbourhoods of the same dominant colours and produces a global spatial homogeneity component.

The extraction algorithm is given as:

$$D_i = \sum_n h(n) \| \mathbf{x}(n) - c_i \|^2, \quad \mathbf{x}(n) \in C_i \quad (3.18)$$

Matching involves searching the data set for similar distributions of colours. The output is not a distance measure but a similarity measure. To calculate the dissimilarity:

$$D^2(F_1, F_2) = \sum_{i=1}^{N_1} P_{1i}^2 + \sum_{j=1}^{N_2} P_{2j}^2 - \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} 2a_{1i,2j}P_{1i}P_{2j} \quad (3.19)$$

Between two colours the similarity coefficient is $a_{k,l}$ as given by,

$$a_{k,l} = \begin{cases} 1 - d_{k,l}/d_{max} & d_{k,l} \leq T_d \\ 0 & d_{k,l} > T_d \end{cases} \quad (3.20)$$

This can be a two-pass process where colours can either be searched for individually and then combined, or where the complete descriptions are compared. The matching process calculates the Euclidean distances between the clusters. The spatial homogeneity is calculated with a weighted difference calculation and the variance is calculated using a mixture of Gaussians measure.

It is recommended that the clustering is performed in the l,u,v colour space. l,u,v is a perceptual colour space similar to the human visual system. It separates illumination from the colour components and allows colours to be compared together effectively unlike r,g,b which is machine optimised. The colours have coordinates on a continuous spherical x,y,z plane.

Lindbloom provides conversion equations used for r,g,b and l,u,v conversions [Lin03]. The choice of reference white (Hoffman [Hof06]) affects the colour calibration of the system. For a system in the U.K. the reference white for PAL is a good choice since U.K. surveillance cameras are PAL; a useful alternative is s,R,G,B [SA96], the standard for white on the World Wide Web.

ScalableColor

This descriptor is a Haar transformed colour histogram defined in the h,s,v colour space. The high- and low-pass coefficients from the Haar transform are accumulated. This descriptor can specify varying bin quantization levels, between 16 and 256 and it is possible to specify how many bits to use for each bin, depending on the required compactness. A low number of histogram bins with a low number of bits gives a fast descriptor suitable for indexing and quick queries, but with low descriptive power. The colour channels are quantized with a bias towards the hue. Since a portion of the high-pass coefficients are redundant these are heavily compressed.

For feature matching the total distance between the histogram bins is measured. The principal advantage of this descriptor is the properties the Haar scaling process which allows descriptors to be matched against another with differing quantization and numbers of coefficients.

GoFGoPColor

The GoFGoPColor (Group of Frames Group of Pictures) DS works with a group of frames, *i.e.* video, and can be summarised as a series of aggregated ScalableColor DS's. The resultant GoFGoPColor descriptor has the same characteristics as a ScalableColor descriptor but is aggregated from an average, median or intersection over all the ScalableColor descriptors generated from frames in an image sequence. The matching technique is the same as for the ScalableColor DS.

ColorStructure

This descriptor is a histogram with a structuring element. The spatial structuring element is used to compile the colour histogram. Hence the spatial structure in which the different colours appear is incorporated into the representation. A 4×4 kernel is passed over the image and the colour channel bins are incremented if a colour is present. The descriptor uses the hue, max, min, difference (H, M, M, D) colour space [MSS02], which is quantized unevenly. Different quantization levels are available as with the ScalableColour DS, with the highest quantization levels giving the best results.

For feature matching the query and data set descriptors are equalised. This is more complex than histogram equalisation because colour quantization affects this descriptor. Unlike the other descriptors the similarity matching process is explicitly defined in the standard and involves bin unification and quantization stages.

ColorLayout

This descriptor is emphasised as a quick resolution independent descriptor that is suitable for indexing or sketch-based retrieval [Bim99] and video segment identification on low-powered devices. Extraction is performed by a discrete cosine transform (DCT) in Y, Cb, Cr otherwise known as y, u, v colour space. The input image is partitioned into blocks where each block is changed to the mean of its colour components. A DCT transform is applied to the blocks in a zigzag scan and weighting gives binary marks. These descriptors can be matched by using a distance measure derived from the combined coefficients produced over the three colour channels in y, u, v colour space.

3.5 Segmentation

The segmentation of objects from the background is an important primary activity for automated surveillance processes. When performed effectively segmentation produces high quality data for further processing. Segmentation can be done manually which results in high quality foreground objects, but this is impractical for surveillance video. Automatic segmentation of moving objects from a static background is possible due to the different pixel values generated of the moving object. In theory the background pixels remain static but in practice they vary due to noise. This noise has a multitude of sources which include fluctuations in lighting and data transmission conditions.

The state-of-the-art computer vision techniques in segmentation and tracking continues to develop. The phases required in a tracking algorithm are well known, Jones *et al.* discuss the object detection, data association and update phases [GROJ08]. Various methods are proposed for the maintenance of background models [TKBM99]. If the background is modelled as a Gaussian small

fluctuations in noise are compensated for so long as the lighting conditions remains static. A model of the background is usually an accumulated running average of background data. This technique can suffer from *lock-in* and *lock-out* of the segmented pixels: when an object becomes assimilated into the background or, respectively, remains present even when the real object has gone.

A single Gaussian is inadequate for outside surveillance and Stauffer and Grimson [SG99] introduced the well-known mixture of Gaussians method, which models different background lighting conditions and switches to most appropriate model according to the current conditions. The basic Stauffer and Grimson method is very sensitive to illumination changes and shadow. Methods to remove illumination sensitivity are focused around colour chromaticity, as with Ellis and Xu [XE01] and Renno *et al.* have been able suppress shadows [JRJ04]. Problems include occlusions of the foreground objects and backgrounds with similar colour to the foreground objects. These can be partially solved by tracking them over several images and camera calibration. The problems of segmentation and tracking remain a key problem.

The benefits of the advanced motion-based automatic segmentation methods are often outweighed by the disadvantages. Due to the requirements of complex software, demands upon hardware and the low reliability of these methods mean most surveillance systems use simple frame *differencing*. This provides a basic but usable level of segmentation, based upon the difference of pixel values between frames. If the difference is above a threshold, then that pixel is segmented. This is useful for basic event detection and the calculation can be based upon the motion blocks used in calculating the motion-vectors used in image compression.

3.5.1 Compression technology

Compression technology is important due to the large amount of data naturally contained within a visual signal. The bandwidth available for transmission can be limited and the aim is to reduce bandwidth requirements but maintain an acceptable quality level. The compression can be *lossless* or *lossy*, where the original uncompressed signal can and cannot, respectively, be regenerated. Lossy techniques yield higher compression levels than lossless and often are used with video. A video signal can contain a lot of repeating information which is therefore redundant. The problem of efficient signal communication was addressed by morse-code and telegraph communication but more significantly by Claude Shannon and the development of Information Theory [Pie80].

The basic techniques employed in image and video compression are as follows [Ric03]: the first stage of compression converts r, g, b colour-space image data to the y, u, v (luma and chrominance) colour-space; the y channel or intensity channel remains full frame, while the (u, v) channels are reduced to two half or quarter size frames, *e.g.* 4:2:2, and 4:2:0, respectively; a DCT (Discrete Cosine Transform) is performed on the y, u, v image at macro-block level; the resulting bitstream is quantized; run-length encoded; and then entropy encoded.

The r, g, b colour space is used by machines for the display of images and videos while y, u, v exploits human insensitivity to colour variation. A macro-block is the cell in a $n \times m$ grid dividing up the image. Intra-frame, or *Key-Frame* encoding compresses a frame where inter-frame, *I-Frame* encoding compresses data between frames. Only changing pixels calculated with *motion-vectors* are compressed in I-Frames.

3.6 Multiple camera correspondence

Multiple camera surveillance systems are deployed to survey a wider area than possible with a single camera and displayed in a control room. Various methods have been suggested on how to keep track of objects between cameras. One method calibrates the cameras into a single world coordinate system, but this is only adequate of cameras overlooking the same location [JRSS03]. Khan and Shah's use 'Field of view lines' [KS03] and Black, Ellis and Makris [BEM04] used statistical information to automatically learn the common *entry* and *exit* points of pedestrians. Porikli and Divakara [PA03] use colour models to maintain the object correspondence. These systems pass meta-data about the objects from one sensor to another. The meta-data contains information about the location, appearance, speed and direction of the tracked object. This information allows the correct camera to track and calculate any correspondence between previously tracked objects.

3.6.1 Colour calibration

Colour variation of a tracked object under different lighting or in a different camera will have an adverse effect on the retrieval performance, as previously noted ([KS03]). The following paragraph is paraphrased [CH98]. Cameras digitise analogue signals to produce three colour channels, r, g, b . These channels are calibrated so each gives zero volts for the colour black and maximum for white. All the channels should be calibrated so they perform at these extremes in unison. The correct voltage required to achieve white will be sensitive to the illumination levels of the scene and calibration will be required when the illumination levels change. Secondly, an increase in voltage from black to white will not be linear and a gamma correction function is used to linearise this function. Calibration

for both the black white levels and the gamma function can be achieved by using colour charts. The charts are based on the colours from standard colour spaces, *e.g.* International Commission on Illumination (CIE) $I^*u^*v^*$ or CIELUV [CIE07] and one example is the Macbeth ColorChecker TM[XR]. Manually re-calibrating the camera's colour response each time there is an illumination change is not practical for a CCTV camera. These cameras, therefore, deploy an automatic gain control and a pre-defined gamma function.

Automatic calibration techniques give good overall performance except when precise measurements of colour are required. These factors must be given due consideration in a CBIR application where colour features are used. In this case these functions will produce outputs which vary even if the cameras are the same model. Various techniques have therefore been proposed which automatically calibrate the colours and cameras. These methods use existing observations to achieve a constancy of colour [FHMO00], and include the colour histogram transformation [Por03] or the Gray World technique [LM71, Buc80, BFC02].

A simple form of colour calibration is achieved with the *gray world* assumption which is one of the earliest developed which attenuates any discrepancies between the camera colour gamuts. It is based upon the assumption that the spatial average of surface reflectance in a scene is achromatic. Since the light reflected from an achromatic surface is changed equally at all wavelengths it follows that the average of the light leaving the scene is the colour of the unknown illuminant. The implied diagonal transform is simply the ratio of the average gray of the image illuminated under the canonical to that of the unknown.

$$\begin{aligned} \Lambda_r = G_r^c / G_r^u & \quad \Lambda_g = G_g^c / G_g^u & \quad \Lambda_b = G_b^c / G_b^u \\ \begin{bmatrix} R' \\ G' \\ B' \end{bmatrix} & = \begin{bmatrix} \Lambda_r & 0 & 0 \\ 0 & \Lambda_g & 0 \\ 0 & 0 & \Lambda_b \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \end{aligned} \quad (3.21)$$

Where G^c and G^u are the canonical and unknown illuminants respectively.

3.7 Ontologies

Once objects are segmented and defined logic can be applied to evaluate their behaviour. The first step is to assign labels to objects from a taxonomy, *e.g.* car or human. The next step is to deduce the object's behaviour.

In the surveillance context Cupillard *et al.* [CBT03] use a Bayesian system of attaching probabilities to *Atomic* events, so *Events* can be built. An Atomic event is defined as a single action, *e.g.* *not moving*. An Event is a composition of Atomic events, *e.g.* *not moving for a long period of time* indicates a *loitering* Event. The Events are defined within an ontology, where the relations between Events can be defined.

The *Video Event Representation Language* (VERL) [NHB05], uses an ontology to describe processes defined out of the atomic elements comprising *Objects*, *States* and *Events*. These descriptions can be linked to some specific evidence in the video content. The corresponding *Video Event Markup Language* (VEML) is used to express an instant of the event defined in VERL.

VEML is expressed in the W3C's (World Wide Web Consortium) Web Ontology Language (OWL) [MvH04] designed to describe the Semantic Web. OWL uses XML (Extensible Markup Language) and RDF (Resource Descriptor Framework) [Hay04], developed from the DARPA (Defense Advanced Research Projects Agency) DAML+OIL (DARPA Agent Markup Language)+(Ontology Inference Language) [MFHS02]. RDF uses URIs (Uniform Resource Identifiers) [BLFM05] to define persistent entities, with a purpose similar to the *term* in a Classification Scheme.

3.8 Analysis of key video surveillance applications

An overview of the different visual surveillance uses for meta-data is presented through the following *categories* of user requirements:

- Visualisation
- Automatic numberplate recognition (ANPR)
- Search and retrieval
- Pattern discovery
- Automatic alarms
- Access control and transparency

These categories are neither exhaustive nor future-proof, but serve only to mark out the expected uses of any meta-data for the visual surveillance domain. They are discussed in more detail below.

3.8.1 Visualisation

The act of observing the scene is important for review and evidential purposes. As described in Section 2.2, manual or automatic meta-data annotation [GRJ02] can help with this task by storing information on format and the media's contents. Such meta-data can help with the task of systems and humans being able to share data with each other and a standard meta-data type will increase this potential.

3.8.2 Automatic numberplate recognition

The advent of reliable automatic numberplate recognition (ANPR) has expanded the use of cameras for traffic monitoring. This specialised system relies upon OCR (Optical Character Recognition) [WP98] and image-processing in conjunction with a vehicle registration database. As described in Section 2.3.1, ANPR has many applications, including general activity monitoring for police enforcement, *average speed* cameras, controlling traffic. Examples include the London Congestion Charge, electronic tolling, marketing and planning. Such dedicated ANPR systems require high-performance camera equipment, dedicated processing systems, and specialised meta-data descriptions. ANPR systems are considered to lie outside the scope of this work which is concerned with generic visual information processing from standard CCTV cameras.

3.8.3 Search and retrieval

A common use for meta-data is for fast and accurate retrieval of particular segments of content. The most elementary form of query is based on information about the *production* of the data, *e.g.* the time, date and specifics of the camera from which the data was generated. A more sophisticated type of query is couched in terms of the properties of what is being observed: queries of this type are known as *content*-based queries. This topic is described in Sections 3.2 and 3.4.

‘Production’ and ‘content’ based queries are not clearly distinguished. There are cases which one might consider to belong to either category, *e.g.* ‘select all video where the camera is panning rapidly’ or ‘select all video where there is rain on the lens’. The method for servicing this query may possibly include the analysis of content, depending on the scope of the available production information,

and the availability of processing power, *i.e.* both human and machine.

3.8.4 Pattern discovery

The above examples require the retrieval of one or more specific segments of video. Another type of requirement relates to the retrieval of patterns of data established over time, or possibly over multiple camera installations. This *data mining* recovers implicit patterns [Som01], *e.g.* customer buying habits discoverable with the advent of electronic point-of-sale systems. In a surveillance context, examples of data mining may include: analysis of historical crowd densities to enable more efficient deployment of space and staff on future occasions; and retrieval of *atypical* behaviour through a comparison with a historically-derived model of *normal* behaviour.

3.8.5 Automatic alarms

All the above are cases of *retrieval* scenarios: a human operator invokes the query and is the recipient of the results. Another type of scenario is that in which alarms are set off autonomously through automatic monitoring of events [Ren03, Kru06]. Examples of this type include: recognition of overcrowding, trespass, unattended baggage, and dangerous behaviour. These are all examples of alarms for which the basis for invocation is a discrete isolated event.

In most cases the alarm must be invoked in approximately the same time as would be taken by a human operator. Exceptionally, automatically generated suggestions about longer-term planning could also be triggered by the data mining analysis introduced above.

3.8.6 Access control and transparency

Retrieval of evidence and generation of alarms are the core activities of the surveillance system. There are additional requirements for any proposed meta-data standard, however, and these become increasingly important as visual surveillance technology permeates more domains within society. Primarily there is a requirement to facilitate compliance with the relevant legal framework [DPA98]. In particular, there is legislation on personal privacy, articles of human rights, that define the notion of ‘fair use’, as discussed in Section 2.4.1. Notwithstanding any potential commercial or pragmatic reasons for wanting to allow controlled access to data and meta-data, there is also the possibility of legislation to obligate such compatibility, on grounds of public safety.

The primary ways in which these ideas relate to meta-data requirements are through restriction and recording of access. It may be considered useful for the meta-data standard to provide a secure, transparent system allowing a full audit trail of all (meta-)data access, as described in Section 2.4.2.

3.9 Existing video surveillance meta-data

Many modern surveillance cameras provide built-in video processing capabilities which range from simple motion detection through to recognition of more complex events, *e.g.* traffic violations and left luggage detection. These devices define meta-data frameworks for use in their systems.

The IBM Smart Surveillance System (S3) [BHC⁺05] has an ‘open and extensible architecture’ for video analysis and data-management. Its role in video analysis is twofold: to encode the camera streams and send them to a database, and to analyse the camera streams for events and store the resulting meta-data in

a corresponding XML database. Its data management module provides a human-interface layer for queries, alerts, events and statistics.

The European project CARETAKER (Content Analysis and Retrieval Technologies to Apply Knowledge Extraction to Massive Recording) [IST] is intended to progress the state-of-the-art in CCTV video and audio processing. The project uses video encoded as MPEG-4: (Part 2) Visual. The automatically produced meta-data conforms to an XML Schema based upon the SERKET (Security Keeps Threats away) [ITE] schema defining the types and syntax. The schema is used for the exchange of meta-data between the collaborators. It is simple and defines the key elements *event*, *object* and *tracked object*. Time-stamps for each frame are imprinted onto the video media, so there is no requirement for the meta-data schema to handle this aspect.

Standard annotated test data can provide improvements in the science. Video has a large amount of visual information and its annotation is a laborious process. The organisation NIST (National Institute of Standards and Technologies) sponsors a conference series called TRECVID in order to drive some improvements in this area [TRE08]. This organisation achieved notable results with speech recognition research using a similar technique. In the surveillance domain various similar initiatives exist. Standardised data sets complete with definitions of the expected answers, *i.e.* *ground truth*, allow a fair comparison of algorithm performance.

Designed for this purpose, the U.K. Home Office have created a data set called the Imagery Library for Intelligent Detection Systems (iLids) [UKG07c]. Other examples include the PETS (Performance Evaluation Test Set) data set series [PET]. In general the data sets cover a particular scenario, *e.g.* left luggage, doorway surveillance, parked vehicle, and sterile zones. *Event* meta-data comprises: the Event label; a time-stamp; and bounding box information.

Similarly, the European project CAVIAR (IST-2001-37540) (Context Aware Vision using Image-based Active Recognition) developed a set standard videos sequences [Fis04]. The video was manually annotated with the expected algorithm results using CAVIAR CVML (Computer Vision Markup Language), a framework for standard algorithm description and evaluation [LF04]. Standardisation of the meta-data is necessary for tool reuse and sharing of the video annotations.

The ViPER-PE (Video Performance Evaluation Resource) *Performance Evaluation* created by the Language and Media Processing Laboratory in the USA is widely used [MMP⁺02, DM00]. ViPER-GT (Ground Truth) is designed to evaluate computer vision processing algorithms. The toolkit specifies a protocol which categorises the evaluation criteria. For example, extracting moving regions from a non-moving background is complicated by many different situations, *e.g.* cluttered background or low-lighting. It is shown that if these situations are broken down into well-defined sub-categories, algorithm performance can be properly evaluated [LMRMJ06].

ETISEO (Evaluation du Traitement et de l'Interprétation de Séquences Vidéo or Evaluation for Video Understanding) is a framework for the evaluation of computer vision algorithms [NBTV07]. ETISEO defines the following stages: algorithm characteristics under analysis; the video data sets to work with; the Ground Truth (GT) medium; the metrics for the performance evaluation; and an automated algorithm performance evaluation service via a web page, where the results can be uploaded and compared. The last stage is similar to the PETS Online Evaluation Service. The results of this initiative were, aside from an insight into performance of the different components. Examples are: metric advantages and disadvantages; how researchers tailor their algorithms to a particular data set; and the effect their different perceptions have, *e.g.* what is an object, whether it is moving or not, *etc.*.

MPEG-7 [ISO02a], has been proposed by Grant *et al.* [GLMP03] as a possible language for describing video surveillance streams. The MPEG-7 visual descriptors described in Section 3.4 are shown to be used in various video surveillance scenarios. These descriptors can be applied to people and vehicles observed in a surveillance scene and can reduce the bandwidth requirements of object and frame-based video coding [COE05].

In the United States, the National Retail Federation (NRF) has a subgroup called ARTS (Association for Retail Technology Standards). ARTS is developing a standard in collaboration with the retail industry to address the needs of industry to consolidate the data produced by multiple video analysis applications used in their CCTV systems [Mad07]. Although early in its development the standard will address issues of theft and missing items, known as ‘Loss Prevention’ in the retail industry. The standard will use XML to structure the data. Section 2.3.2 describes retail surveillance activities and meta-data in more detail.

3.10 Related meta-data standards

The importance of meta-data in digital systems is clear and there is a need for these systems to exchange meta-data via computer processes. This section reviews various standards from domains different to surveillance. A broader overview is given by the W3C (World Wide Web Consortium), as part of its investigation into how many of these standards can be used with OWL (Web Ontology Language), as part of its development of the Semantic Web [BBC⁺07]. The requirements of the underlying technology used for storage and exchange of the meta-data are not reviewed, such as databases or XML.

The television and film industry uses meta-data extensively to describe information related to the moving pictures and audio. Compatibility is vital because

information is shared between a wide variety of interested parties, ranging from end-users, cameramen, production workers, directors, editors, *etc.* The Society of Motion Picture and Television Engineers (SMPTE) is very active in this area, generating a wide variety of standards covering all areas from the definition of a time-stamp to meta-data to file formats.

Time is a fundamental attribute in need of standardisation. The 'SMPTE Time Code', [SMP99], forms the basis for subsequent standards which use time. The Vertical Interval Time Code (VITC) [SMP01a] specifies a method of encoding extra time-code information into every frame in an analogue or digital video stream. Their Metadata Dictionary [SMP01b] defines meta-data to describe: identifiers; time-stamps; detailed sensor movement; platform movement, *etc.* The KLV (Key-Length-Value) format [SMP01c] defines a mechanism with which to encode and include meta-data within a binary stream.

The US DoD/IC (Department of Commerce - Information and Communications) Motion Imagery Standards Board (MISB) have defined several relevant meta-data standards. MISP (Motion Imagery Standard Profile) [MIS07], previously VISP (Video Imagery Standard Profile) is used in NATO (North Atlantic Treaty Organization) standards. The standard describes the meta-data used in various applications including the Security Metadata Set.

In the Library domain managing electronic resources has become increasingly important, both for electronic indexing systems and electronic library items. Library catalogues use the Metadata Encoding and Transmission Standard (METS) [Car07] which is a complex overarching library standard based upon the widely used Dublin Core meta-data set. The Dublin Core Metadata Initiative [DCM08] specifies a set of simple meta-data types defining essential meta-data for cataloguing purposes, *e.g.* author, title, *etc.*, all of which are optional. Dublin Core has been extended for use in the multimedia domain [HL01]. This research work

included RDF (Resource Description Framework) [Hay04] tags for image and video descriptions. The Video Development Initiative (ViDe) [VID08] are a working group developing a variety of video related technologies, especially for video conferences. They have developed the Dublin Core Application Profile for Digital Video. This mixes Dublin Core, RDF and ViDe meta-data and concentrates on high-level textual and semantic annotations of the video. The International Imaging Industry Association (IA3) and their Digital Imaging Group (DIG) have developed the DIG35 specification [DIG]. The specification is designed to improve the compatibility and interchange of meta-data about images.

The EBU (European Broadcasting Union) have developed the P/Meta meta-data standard for digital broadcasting [EBU07a]. This is based upon Dublin Core meta-data set and is used to share EPG (Electronic Programme Guide) information between systems. Similarly, the European Telecommunications Standards Institute (ETSI) have published TV-Anytime [EBU07b]. This is a set of specifications defined by the ‘TV-Anytime Forum’, whose members comprise major companies involved in the broadcasting industry. The specification is designed as a component in the consumption of digital data, using consumer’s local storage facilities. The standard combines SMTPE and MPEG-7 meta-data components. TV-Anytime is scheduled to be published by the Internet Engineering Task Force (IETF), part of The Internet Society, as an ‘Internet standard’ .

The JEITA (Japan Electronics and Information Technology Industries Association) Exif (Exchangeable Image File Format) [JEI03] is based upon the meta-data fields within the TIFF (Tagged Image File Format) file format [Ado08] and has been adopted as the meta-data format for JPEG (Joint Photographic Experts Group) files. These are attached to a standard JPEG image [ISO94]. For example, surveillance products developed by Grandeye [Proa] use this type of mechanism.

In the corporate domain meta-data is used to manage, maintain and deliver documents. Adobe's Extensible Meta-data Platform (XMP) [Pro08a] is a meta-data standard designed to help the work-flow of media in a business context. It is designed to standardise the meta-data used in the process of information interchange, *i.e.* standardising the format of the meta-data, *e.g.* author, time-stamps, usage rights *etc.*

3.11 MPEG standards

The ISO (International Organization for Standardization) / IEC (International Electrotechnical Commission) working group MPEG (Motion Pictures Experts Group) develop standards for audio, video and multimedia in general, such as, computer graphics. Examples of the video standards are MPEG-1 [ISO98], which is used for digital camera video and MPEG-2 [ISO07a], used for digital television. Music players use MPEG-1 Audio Layer 3 (MP3) [ISO93b], MPEG-2 Audio [ISO00b] and MPEG-4 Audio [ISO04d].

MPEG is composed of *National Body* categories of participating countries, whose members comprise national companies and institutions. MPEG standards standardise the decoder and the form of encoded data or *bitstream*. The encoder is not standardised nor are the specifics of the decoding process. Reference Software is provided with MPEG standards to promote industrial adoption and used to develop bitstreams for conformance testing.

Companies and institutions invest in MPEG technologies to include their Intellectual Property within international standards. Their endeavours are rewarded through remuneration from royalties due on patented intellectual property. MPEG-LA (Licensing Authority) is a company dedicated to creating licence pools

of patent holders and also to issue the licences. Royalties are paid by manufacturers to patent holders on a per-unit basis.

The processes involved in creating a standard are as follows. A proposal for a new standard is received during a *Requirements* phase. Here the proposal is refined until it is acceptable to the members. The proposal moves to the next phase, called *Systems*, where the appropriate technologies are gathered together and the standard is drafted. During this phase, an amendment to the standard is also drafted, called *Conformance and Reference Software*, which comprises a document containing the details about testing a bitstream's conformance and a *Reference Software* implementation. The draft standard goes through transitional promotional phases, *e.g.* *Committee Draft*, *Final Committee Draft*, and so on, until the status of International Standard is reached. These promotions are dependent upon the results of ballots on which the National Bodies have voted. Any comments are addressed during the quarterly meetings.

The MPEG technologies: MPEG-1 (Part 2) [ISO93a]; MPEG-2 (Part 2) [ISO00a]; MPEG-4 (Part 2) [ISO04c]; MPEG-4 (Part 10) (AVC) (Advanced Video Coding) [ISO03a] are all video compression technologies which use the methods described in Section 3.5.1. They vary the macro-block size, bit-rates and I-Frames composition.

3.11.1 MPEG-7 – a summary

MPEG-7 [ISO02a] is the Multimedia Content Description Interface from 2001. MPEG-7 is an eleven part standard with five main parts. The main parts are Part 1: System [ISO02a], Part 2: Data Definition Language (DDL) [ISO02b], Part 3: Visual [ISO02c], Part 4: Audio [ISO02d] and Part 5: MDS (Multimedia Description Interface) [ISO03b]. In 2005 profiles of MPEG-7 were developed

in Part 11 [ISO05d] to refine the standard's scope. The profiles are the Simple Metadata Profile (SMP), the User Description Profile (UDP) and the Core Description Profile (CDP).

MPEG-7 defines an exhaustive list of data structures for many aspects of multimedia meta-data, such as: production, rights, usage, syntax and semantics [Mar04, MKP02]. Manjunath *et al.* list surveillance as one of MPEG-7's application scenarios [MSS02] (Chapter 2). The MDS provides the structure of the MPEG-7 document. All the basic description types are provided, *e.g.* time, identifiers and media descriptions. The following list taken from Manjunath *et al.* (Chapter 6) *idem* successfully categorises the main components:

- Content Organisation – collections and models
- Content Management – media, creation and production, usage
- Content description – structural and semantic aspects
- Navigation and access – summaries, views, variations
- User Interaction – user preferences, usage history
- Basic Elements – schema tools, basic data-types, links and media localisation, basic tools

As well as basic meta-data the media descriptors describe all aspects of the media to fine detail including format, decomposition of the media into segments, content-based descriptors, camera movement and calibration. Semantic descriptor types express high-level information about the media. MPEG-7 is evaluated for semantic modelling with COSMOS-7 [AA05] and for semantic retrieval [LG05]. Many different mathematical models are possible, *e.g.* probability models, relations, graphs and state models. These can be applied to either high-

or low-level semantic descriptions. Creation, summarisation, user preference, advanced text annotation and Classification Scheme definition tools also exist. The System tools allow the XML meta-data to be updated using fragments rather than re-sending the whole XML document [GLMP03].

The meta-data is expressed in XML as simple atomic elements called *Descriptions* or more complex composite types called *Description Schemes*. Every MPEG-7 document starts with the **Mpeg7** root element and there are the three possible types of child element: a media's description (**Content Entity Type**); models (**Content Abstraction Type**) and content management tools (**Content Management Type**). The DDL allows the creation of data structures not defined within XML Schema, *e.g.* matrices. It also demonstrates the preferred method of extending the standard with custom data structures.

The MPEG-7 standard provides a reference software implementation, called the Experimentation Model (XM) [ISO03c]. This contains the routines necessary to create and query all the elements defined within the standard, including routines for performing CBIR (content-based image retrieval) evaluations. The XM is enormous and has a complex architecture. Other software implementations exist which provide an API (Application Programming Interface) to manipulate the MPEG-7 XML elements [FNB07].

MPEG-7 has been extended and profiled [ISO05d] for use in audio broadcasting. The profile is called the Detailed Audiovisual Profile (DAVP), by Joanneum Research [BS06], and defines more detail in the audio descriptors. The institute is also interested in the semantic validation of the meta-data content [TBH⁺06]. Semantic validation can be a problem because the XML Schema only ensures that the XML document is valid syntactically. This problem limits the extent to which normative content can be strictly defined.

3.11.2 MPEG-21

MPEG-21 [ISO04e] is a multimedia meta-data framework comprising eighteen parts. It has also been used as a repository for outstanding meta-data standardisation tasks, perhaps not addressed by MPEG-7 [BdWH⁺03]. MPEG-21 (Part 2) – Digital Item Declarations (DID) [ISO05f] defines the objects and entities that work within the MPEG-21 framework. The MPEG-21 File Format [ISO05g] allows media to be protected and associated with meta-data. Other significant components include Intellectual Property Management and Protection (IPMP), the Rights Expression Language (REL) to express the rights given to systems and individuals when manipulating digital objects and entities, *i.e.* Digital Items, and a dictionary of terms used in the management of these rights, respectively.

3.11.3 MPEG-4: (Part 12) – ISO Base Media File Format

MPEG-4: (Part 12) [ISO05b], also known as the ISO Base Media File Format, defines a general purpose object-oriented and extensible file format for media and meta-data able to package a variety of media. It is designed for stand-alone storage or streaming and to be extendable. MPEG-TS (Transport Stream) [ISO07a] has been amended to be able to carry the format. The AVC File Format is an extended version able to carry AVC bitstreams [ISO05c]. It is also possible for the format to contain other complete ISO file instances. The file format is object-oriented and its compartmentalisation into ‘Boxes’ is shown in Figure 3.2.

The main boxes are *file*, *moov* and *trak*, which describe different semantic levels, *e.g.* the file, the movie and the tracks, respectively. Each *trak* box contains a link to the media bitstream (*mdat* box) and information about the bitstream, such as, a sample table (*sbt*) to facilitate fast-forward and rewind functionality. An optional meta-data box called *meta* can be placed at *file*, *moov* or *trak* level.

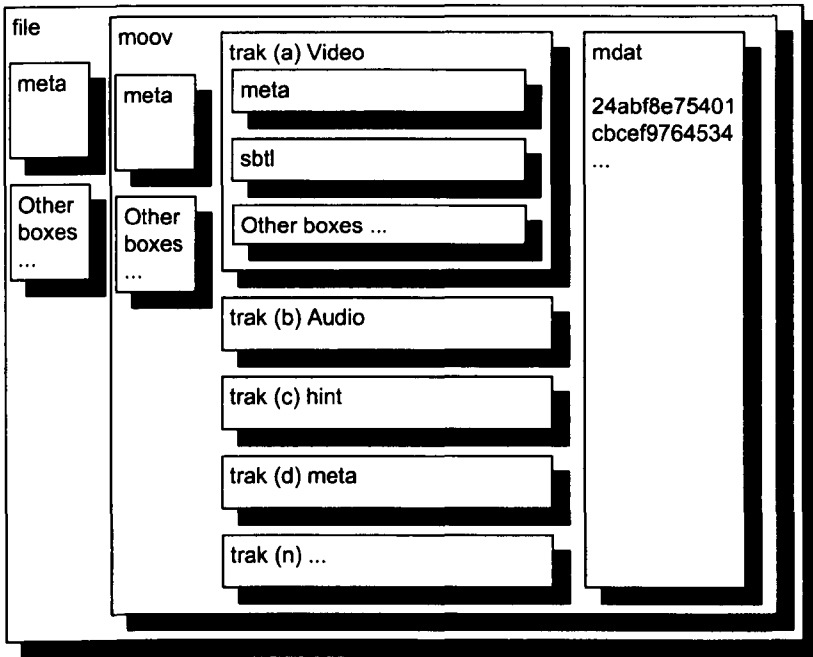


Figure 3.2: The top-level structure of the ISO Base Media File Format. The ‘meta’ boxes can contain XML meta-data. The ‘mdat’ boxes contain the binary data.

The *meta* box can contain XML or binary XML, *e.g.* MPEG-7. The ISO file has many other features which include *Timed meta-data* tracks which deliver meta-data at timed intervals [ISO05c], and *Hint* tracks to provide optional information for the stream decoder.

3.11.4 MPEG-A – MPEG multimedia application format

MPEG multimedia application formats (AF) or MPEG-A [ISO07c] offer a swift standardisation process for new file formats based upon MPEG technology. The MPEG-A formats are based upon MPEG-4: (Part 12) – ISO Base Media File Format [ISO05b] whose structure decouples the file format from the media and meta-data. At present the standard comprises 10 application formats, each designed for a different application. The standards are based upon elements from the wide-ranging MPEG standards, *e.g.* video from MPEG-4, meta-data from

MPEG-7 and content protection from MPEG-21. Major considerations concern the choice of profiles taken from the existing standards and how the ISO Base Media File Format extensions will be defined, if at all. AFs exist for storing music and photographic collections and MPEG-A: (Part 10) – Video surveillance application format (VSAF) [BR08] is designed for CCTV. A summary of MPEG-A is provided by Diepold *et al.* [DPC05], and an overview of all the AFs including those under consideration is provided in the MPEG application format overview document [SD07].

3.12 Media file formats

The media and meta-data must be packaged to allow it to be distributed and archived together in a cohesive manner. The approaches taken by industry depend upon the application. The following Section reviews various file formats in use with an emphasis on their meta-data capabilities.

The AAF (Advanced Authoring Format) [Gor00] was developed by the Advanced Media Workflow Association. Its purpose is to exchange material in an unfinished state between processes within the production industry. The format solves the problems of the preservation and the presentation of detailed and accurate information. The different parties in the process chain, *e.g.* sound engineers and video editors, need a common interface in order to optimise the work-flow. The format organises different media in a single file alongside meta-data descriptions. For access the file is not played but the media items are unpacked in their unfinished state, *i.e.* work-in-progress. There are different categories of meta-data designed for different purposes, *e.g.* structure and content *etc.* The Microsoft COM (Common Object Model) structured storage technology solves problems of storing multiple objects in a single file. The AAF Authority provides

a SDK (Software Development Kit) for the production of AAF files to encourage interest in the standard. The format is agnostic to the content, so any media or meta-data can be placed within it.

The SMPTE Material eXchange Format (MXF) [SMW07] is an extension of the AAF format designed for the exchange of the *finished* project work between members of the production team. The AAF and MXF are designed to be interchangeable with one another. Both the AAF and MXF format are in use in other application areas, *e.g.* military standards.

A NATO STANAG (Standardisation Agreement) defines a standard concerned with military activity. STANAG 4609 [NAT07b] standardises the use of surveillance and remote sensing image and video data. The specification defines a format for digital signals comprising video and meta-data, and specifies profiles and levels of complexity. The video is coded as MPEG-2 Video [ISO02c] and MPEG-4 AVC [ISO03a], and the meta-data is selected from the SMPTE Meta-data Dictionary and MISP (Section 3.10). The video and meta-data can be streamed using MPEG-TS [ISO07a], and the meta-data packaged in the SMPTE KLV format. The specification allows the video and meta-data to be stored within a file using the MXF format. Various meta-data sets are defined in the STANAG 4609 Implementation Guide [NAT07a], *e.g.* UAV Metadata Set and the Security Metadata Set. These SMPTE meta-data elements describe identifiers, time-stamps and those more suitable for particular applications, *e.g.* airborne surveillance has detailed sensor requirements and platform movement descriptions for roll, pitch and slant. Other important issues include byte-order specification and timing reconciliation information.

3.12.1 Music player application format

Meta-data has sometimes been a secondary consideration when designing file formats, *e.g.* the JPEG (Joint Photographic Experts Group) concentrated only on compressing the image bitstream with the JPEG file [ISO94]. Similar to the JPEG format used in cameras, MP3 files do not contain meta-data. As a consequence a set of meta-data tags were developed called 'ID3' [Nil99] which are appended to the bitstream. The MPEG-A: (Part 2) 'Music player application format' [ISO06c] is a format that contains both music media and meta-data. The audio format is MPEG-4 and the meta-data as MPEG-7. A new format called 'MP3onMP4' was developed to allow MPEG 'MP3' audio [ISO93b] to be compliant. The Music Player AF offers improvements on MP3 with ID3 tags in the following ways. Since MPEG-7 is an XML format meta-data processing is simplified. Additionally since an MPEG-7 XML document can be validated against an XML Schema there are no subtle differences between versions as there can be with ID3 tags, *e.g.* version 1, 2, *etc.* The format offers improved streaming capabilities when compared to MP3 with ID3. ID3 tags have been employed for carrying binary data useful for streaming purposes. The format uses the MPEG-4 File Format to provide similar mechanisms.

3.12.2 Photo player application format

The MPEG-A: (Part 2) Photo player application format [ISO07b] is similar to the Music Player AF and is designed to package image and meta-data content. The format extends the functions provided by a JPEG image with Exif meta-data. The Exif tags are converted to MPEG-7, providing benefits such as, meta-data is not restricted to 64 kilobytes (KB) in total size, and the MPEG-7 image feature descriptors can be used for CBIR (content-based image retrieval). The format

profiles MPEG-7 to reduce its scope and also allows the use of binary meta-data [ISO05h]. The format uses MPEG-4 File Format to package the images and meta-data.

3.12.3 Digital Versatile Disc / Digital Video Disc (DVD)

The DVD (Digital Versatile Disc or Digital Video Disc) [Pro04] format is the result of cooperation between the home video entertainment industry and the video equipment manufacturers. The DVD format has a clear ‘read-only’ application with a software meta-data layer for internal control of advanced playback functions. The entertainment industry required a customisable presentation layer, while the equipment manufacturers required an efficient video and audio file format and coding scheme. Furthermore, the content providers needed digital rights management for copy protection. DVD is an improvement on the VHS (Video Home System) format [IEE] especially due to its software layer. The software allows flexibility in the media content and presentation, unlike Compact Discs.

The following paragraph is paraphrased from an article on the Doom9 web site [Doo03]. The format comprises a container with IFO ‘InFOrmation’, BUP ‘BackUP’ and VOB ‘Video OBjects’ files. The VOB files comprise multiplexed video, audio and subtitle tracks. The video is compressed using the MPEG-2 technology. The audio is either uncompressed pulse-code modulation (PCM), or most commonly Dolby Digital AC-3 (Arc Consistency Algorithm #3), but MP2 (MPEG-2 Audio Layer 2) is available for use. VOBs contain bitstream meta-data, *e.g.* encryption key, identifier, width, height, aspect-ratio, frame rate and a bitstream reference (cell). A Cell is the smallest component of the DVD format. A VOB can contain up to nine different audio streams, 32 different subtitle streams.

A movie can comprise many ‘titles’, known as Program Chains (PGCs). These PGCs can contain a series of VOB files. The IFO file contains chapter, audio or subtitle information for use by the player. A BUP file is a backup of the IFO file.

The DVD format provides limited scope for complex meta-data interaction. The meta-data is read-only, set at creation time and protected. The format is not designed to perform content-based retrieval, *e.g.* a search to retrieve all instances of a particular actor. For comparison, object-based MPEG-4 bitstreams [ISO04b] can produce a presentation similar to a DVD and include MPEG-7 meta-data.

3.13 Summary

This Chapter reviews the key technologies relevant to surveillance meta-data description. CBIR (content-based image retrieval) technology, in Section 3.2, is concerned with retrieving images from a database using signal content. Three different query paradigms are described with increasing complexity. The simplest query is *by example* (QBE). A Query by Attribute, as defined by this document, requires an intermediary process to generate QBE data and a Query by Symbol requires an ontology to describe the scene. For CBIR development a data set is required for testing and metrics are required in the evaluation process (Section 3.3). To evaluate the feature’s performance three different types of evaluation methods are reviewed: a binary yes or no, a ranking metric and a probability metric. Section 3.4 reviews a Co-occurrence texture feature and standard MPEG-7 colour descriptors. Algorithms are specified for the encoding and matching phases for these standard CBIR features.

Segmentation is essential for quality features for use in a CBIR system and the techniques are discussed in Section 3.5. The Stauffer and Grimson algorithm is the basis of most approaches, although industry use simpler methods. Multiple

camera systems are described in Section 3.6 and require calibration in order to allow tracking and CBIR applications. This can be achieved by colour or other methods. A brief discussion of Ontologies is given in Section 3.7 and how they are used in surveillance.

Section 3.8 reviews several key applications for the use of meta-data in visual surveillance. Unlike the other applications, many of which are under development, ANPR (automatic numberplate recognition) is an example of a commonly used signal-processing technique for CCTV, which makes use of specific and well defined meta-data. Meta-data is also essential search and retrieval and is a component in automated alarms. Standard interfaces are required for CBIR systems to perform inter-system searches. For pattern discovery the meta-data generated by low-level detection algorithms, *e.g.* for movement or behaviour, helps generate statistics for business decisions. Standard meta-data can help in information transparency and availability.

Section 3.9 provides an analysis of existing advanced meta-data use in the surveillance. IBM sell the Smart Surveillance System which is an example of proprietary technology making use of complex meta-data. The NRF (National Retail Federation) ARTS (Association for Retail Technology Standards) standard is the exception to the norm of proprietary systems and is biased towards *retail* video surveillance.

Standards are mainly used in the research domain for performance evaluation descriptions, *e.g.* ViPER (Video Performance Evaluation Resource). ViPER-GT (Ground Truth) is widely used in the surveillance research community for authoring Ground Truth descriptions. A ViPER-GT description is oriented to the objects rather than frames. A meta-data format can be frame- or object-based, with associated benefits and disadvantages. The schema is developed by the research community and restricted to performance evaluation without the

tight control of an international standard. This means errors and omissions are more likely to be present. The Section also shows that automatically processed scene meta-data is usually described with the following terms, *e.g.* agents, objects and events. It follows, therefore, that a new content description standard should both draw from this lexicon and define types with expected characteristics.

Section 3.10 reviews existing meta-data standards that lie outside the surveillance domain. These are applied in the Production and Library industries. SMPTE (Society of Motion Picture and Television Engineers) has produced many meta-data standards and has defined essential meta-data types, *e.g.* time-codes and KLV (Key-Length-Value), with wide use. The SMPTE Metadata Dictionary is intended to be deployed within a user-defined structure, *e.g.* XML Schema. The other standards focus on cataloguing meta-data, *e.g.* Dublin Core Application Profile for Digital Video or are application specific, *e.g.* TV-Anytime. Adobe produce a standard for information flow within the corporate environment and it is tightly coupled to their application and framework.

MPEG (Motion Picture Experts Group) demonstrates expertise in a wide array of multimedia standards. The MPEG video coding tools are briefly reviewed as necessary background since these are commonly used in surveillance. MPEG-7 provides a professional quality standard designed to encompass all video annotation requirements and demonstrates itself as the most comprehensive multimedia meta-data standard available with a very broad application focus and huge size. It has applicability to surveillance and includes a complete suite of meta-data for application in CBIR. It may be possible to convert ViPER-GT documents into MPEG-7's frame perspective. MPEG-21 complements MPEG-7 with a suite of standards for implementing frameworks where Digital Items can be managed for consumption. MPEG-21 provides many tools with potential applicability to

surveillance but not the fine grained content description as in MPEG-7. The surveillance industry is known to be conservative and MPEG-21, at present, appears not to be used.

The file format review discusses solutions to the problem of packaging media and meta-data. A file format specifies the structure for media and meta-data storage and can produce static files, files for media streaming and provide structure when the media has a transient role within a system. Significant file formats from SMPTE and NATO (North Atlantic Treaty Organization) are reviewed. The AAF (Advanced Authoring Format) and MXF (Material eXchange Format) file formats are potentially inappropriate for surveillance applications due to their flexibility in allowing any information to be included. A comparison is given between three further file formats, in particular the MPEG-A family. The discussion shows improved meta-data support is a feature of recent formats.

Chapter 4

Analysis of colour-based meta-data

4.1 Introduction

Video content analysis underpins the key objectives of the visual surveillance programme. This Chapter focuses on evaluating standard colour feature descriptions for use in a CBIR (content-based image retrieval) system where the scenario is the retrieval of information about the identity of human subjects, observed with medium- to far-view image sequence data from multiple cameras. In such a scenario face recognition is not presently reliable and the subject can only be identified by other features, *e.g.* clothing, hence a short timescale is assumed. Nevertheless there are two key applications for this technology: one facilitates the retrieval of information used by the operator to understand the past or present whereabouts of a specific individual; the other gives a multi-camera system the automatic ability to develop a coherent scene description.

The Chapter is composed of three main Sections based around two different types of metrics. The methodology presents results of Query by Example (QBE)

experiments using the colour features from MPEG-7 (Part 3) Visual [ISO02c] and additional non-standard descriptors. A data set is developed for the purpose, called GENERICK. Various methods are incorporated to improve the results which consist of the application of colour constancy and different methods for combining the descriptors. The first Section uses simple operators to generate an aggregated ranking based upon pedestrian data split between upper and lower body components. The second Section uses a combination technique based upon measuring the amount of information each descriptor contains and combining them in a probabilistic framework. The third Section provides a further evaluation of the experimental framework using standard data. A new specialised data set is produced from the CAVIAR [Fis04] data set.

4.2 GENERICK data set

The GENERICK data set, constructed for this project, contains people observed on multiple occasions from multiple cameras. Video sequences showing 47 people entering and leaving a room are assembled. Each person is filmed by two cameras A and B , which both observe the two movements in and out , resulting in four image sequences for person i : A_i^{in} , A_i^{out} , B_i^{in} , B_i^{out} , where A and B refer to the sequences captured from the two cameras (Figure 4.1). The data set is designed to allow multiple camera analysis, tracking, and search and retrieval applications.

4.2.1 Limitations of the data set

The data set is used with various assumptions in mind:

- The subjects are limited to undergraduate students from a single computer science lecture at Kingston University.



Figure 4.1: Example data showing the same person, from the two different cameras, each with two views. From left to right: A^{in} , B^{in} , B^{out} and A^{out} .

- The scene is indoors and lighting is constant, with a slow and steady decrease in the amount of light let in from the windows.
- The cameras are the same type and do not move, aside from small variations in position caused by the manual starting and stopping operations.
- The subject matter comprises single moving objects, occasionally occluded by static office furniture; there are no moving occlusions.
- There are at times other subjects in the video but they are seated and do not substantially interfere with the observations of the subjects.

4.3 Applying the MPEG-7 Colour Descriptors

This Section presents results showing how effective standard MPEG-7 descriptors are in Visual Surveillance retrieval problems. The surveillance problems involve the retrieval of video data containing a given person, specified by an example image captured at a different time, or by a different camera. The MPEG-7 reference software [ISO03c] is used to produce the MPEG-7 colour descriptor

distances and again to perform the QBE tests.

The MPEG-7 Description Schemes (DS) used are DominantColor, ColorLayout, ScalableColor and ColorStructure, described in Section 3.4.2. Notably MPEG-7 specifies differing levels of quantization for the ScalableColor and the ColorStructure. Two settings are used, *e.g.* 32 and 256 histogram bins, to represent a minimum and maximum range of complexity. This results in a total of six separate descriptor results.

Additionally a ‘Mean’ colour meta-data descriptor is created, which describes the average of all image pixels, over three separate colour channels in r, g, b space. The matching stage used a Euclidean distance to measure the distance between an input query against the database items, combining the differences between each channel. Finally, a random classifier is included as part of the verification of the experimental and evaluation method; for each rank position it selects any of the query set with a uniform probability (without replacement).

The foreground regions are segmented from the background using both manual and automatic processes¹ as described in Section 3.5. The automatic motion detection uses a per-pixel multi-modal background model in the h, l, s colour space with shadow suppression [JRJ04]. The model comprises one two-dimensional Gaussian which represents the background colour and a one to one-dimensional Gaussian representing the intensity [SG99].

In the context of static surveillance camera output, in order to achieve colour constancy, the methods described in Section 3.6.1 are considered. The Gray World assumption is selected due to the colour-oriented focus of this study and is applied in a novel way, described as follows. Assuming a common illuminant across both cameras and a relatively constant foreground appearance, the Gray

¹The author gratefully acknowledges the work of J-P Renno in producing the automatically segmented data.

World algorithm should be sufficient for colour correction over common foreground areas. Not only is the foreground the area of interest but it also negates the problem of significantly different and constant background colours between cameras. The foreground regions form the most dynamic part of the scene and contain the data we are interested in, *i.e.* the pedestrian. Gilbert and Bowden [GB06] also make this observation in a parallel development.

4.3.1 Experimental design

Presented here are three experiments designed to test differing CBIR scenarios of increasing difficulty. Experiment I presents retrieval results using the same camera, but different direction of movement, *e.g.* comparing A^{in} with A^{out} . Experiment II compares the same movement with different cameras, *e.g.* comparing A^{in} with B^{in} . Experiment III investigates retrieval performance using sequences from a different camera, with a different direction of movement, *e.g.* comparing A^{in} with B^{out} . For Experiments II and III, the retrieval performance is evaluated with and without a Gray World method for achieving colour constancy.

15 pedestrians are randomly selected for the tests. The moving regions, *i.e.* the pedestrians, are both manually and automatically segmented from the background, illustrated in Figure 4.2. Each pedestrian has nine separate same person images and a set of three queries. Three same person images and one query image are used for each of the three experiments. The images taken for the same person data are chosen accordingly, *e.g.* side-view or front-view to satisfy the different camera and pose requirements. The query data is taken randomly from a predetermined range of usable images created entirely from people leaving the room from the front-facing camera.



Figure 4.2: The 15 participants: manually segmented (top); and automatically segmented (bottom).



Figure 4.3: Example of automatic (left) and manual splitting (right) to produce Top and Bottom data.

4.3.2 Splitting and combining descriptors

The two methods described here require the similarity output from the above descriptors to be combined to generate a joint ranking. The first exploits the frequent separation of the clothed appearance of pedestrians, which allows the foreground mask of each person to be split in two, giving separate top and bottom meta-data. The experiments use both automatically and manually split images (Figure 4.3). The automatic process splits the foreground region half-way down the extracted foreground blob. The manual process separated the two outer layers of top and bottom items of clothing. Although these top and bottom data are calculated separately the intention is to combine them to give a retrieval process that jointly uses both top-half and bottom-half meta-data called a ‘Spatial Combination’. The combination process maintains an explicit distinction between the two to exploit the assumption that pedestrians stay the same way up. This method may also have applications where people are occluded, *e.g.* from the waist down. The second case is for different colour descriptors which may have complementary characteristics and if combined appropriately could improve the overall retrieval rate. This is called a ‘Descriptor Combination’.

For both Spatial and Descriptor Combinations there are several methods by

which the individual results are combined. In either case a potential difficulty is the incompatibility of outputs from the two Descriptors which may have completely different units and scales of output. Although there are solutions to this problem, *e.g.* converting each into a Mahalanobis distance, this is not without complications and so the rank output from each Descriptor is selected as the most appropriate input to combine into a joint descriptor. As a result, four different operators are tested to combine ranks: *sum*, *product*, *minimum* and *maximum*. Only subjects appearing in both ranks, while limited to 20 entries, are operated upon. Subjects not in both lists are assigned a rank value larger than the ANMRR k value (Section 3.3.3). The results after a *sum* operation have more effect on the higher ranks. The *product* operator acts in a similar way, and the output has the same effect as the mean rank of the two. The *minimum* operator selects the best (lowest) rank from the two, while the *maximum* selects the worst (highest) rank. The resulting ranked list is sorted in descending order. The result is a total of seven different configurations: the whole region can be submitted, or Top only, or Bottom only; or combinations of the Top and Bottom descriptors using the four different operators.

4.3.3 Evaluation of retrieval accuracy

The *Average Normalised Modified Retrieval Rate* (ANMRR) metric, described in Section 3.3.3, is used to evaluate the performance of the colour features. The metric ranks the results according to a score associated with the closeness of the match. The ANMRR metric is invariant to data set and *ground truth* sizes, both of which can affect the ranking results. The term Ground Truth (GT) refers to correctly identifiable subjects within a data set, *e.g.* the number of different instances of person x in a data set of size n . The result is that two algorithms

Rank (1-K)	GT Retrieval (n)	NMRR (0-1)
1.2.3.x.x.x	3 in top 6	0
1.2.x.x.x.x	2 of 3 in top 6	~0.25
1.x.x.x.x.2	2 of 3 in top 6	~0.5
x.x.x.x.1,2	2 of 3 in top 6	~0.75
x.x.x.x.x.x	0 in top 6	1

Table 4.1: Example of retrieval scenarios and the Normalised Modified Retrieval Rate (NMRR). Three ‘true’ items can be retrieved; all the rest are ‘false alarms’.

might deliver a similar performance despite giving different retrieval results.

Six different MPEG-7 features are evaluated in each scenario alongside two simple control features: the Mean and Random features. These are evaluated using the ANMRR: where a value of 0.0 indicates perfect retrieval and 1.0 corresponds to no retrieval at all. Table 4.1 shows how to equate the ANMRR values to different retrieval success rates.

4.3.4 Results

Experiment I is first conducted with a *manual* segmentation of each person in the image. This virtually eliminates contamination from the background. The results for automatic segmentation and splitting are shown in Figure 4.4.

For all configurations of data the random classifier gives a result of roughly 0.9 providing one useful validation point for the experimental procedure and a point of reference by which the other methods may be judged. This value indicates the computer has one chance in 15 of randomly identifying the individual. A second reference point is provided by the simple Mean descriptor: here (Figure 4.5) (top), the Top and Whole configurations provide a retrieval rate of about 0.54. Unsurprisingly it is less reliable to retrieve an individual’s identity using only their bottom half (0.68). All four Combination methods, however, when performed on Top and Bottom retrieval significantly improve the performance

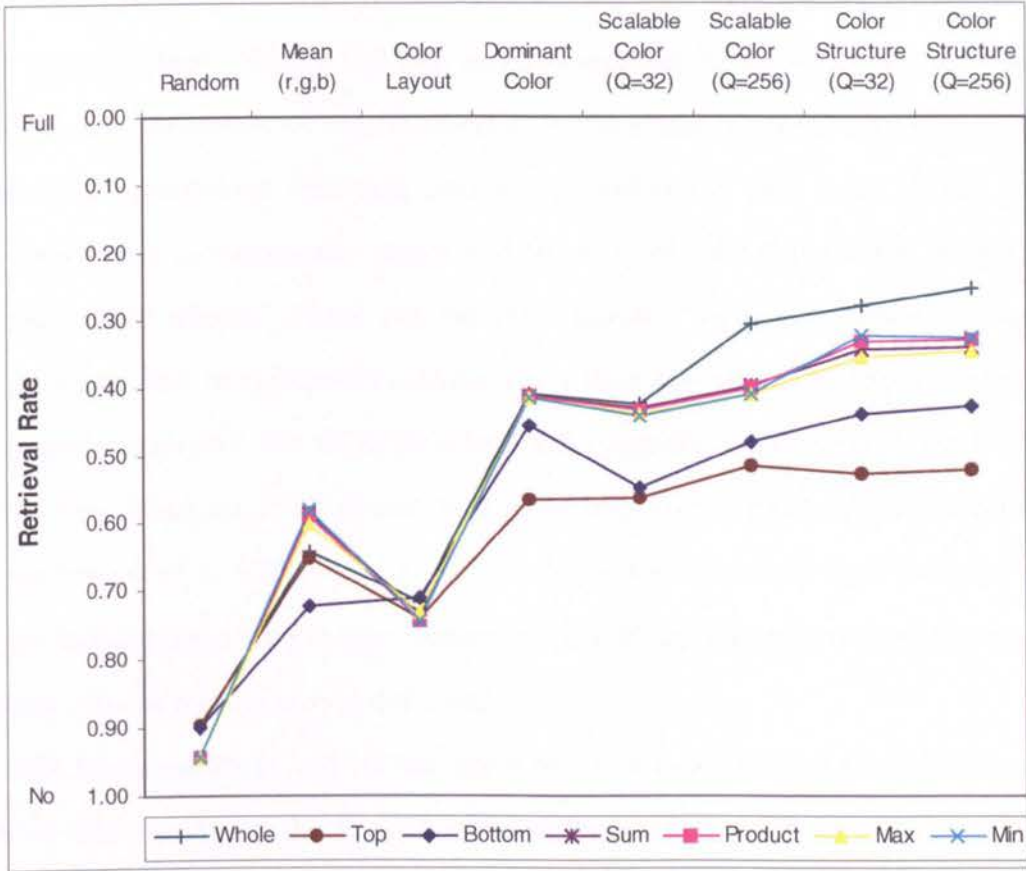


Figure 4.4: Retrieval Rate for Experiment I (same camera, different direction of movement) with automatically segmented foreground data. The names of the colour descriptors are stylised for brevity: Whole is the complete motion mask, Top and Bottom are the ‘top’ and ‘bottom’ mask portions respectively. Sum, Product, Min and Max are calculated from the ANMRR values from the Top and Bottom mask regions.

of the mean classifier, achieving 0.45. When the Top and Bottom halves are manually segmented and separated (Figure 4.5) (bottom) the Mean classifier gives nearly 0.20.

The best retrieval performance is displayed by the ColorStructure DS. Experiment I demonstrates an ANMRR of around 0.21 with manual segmentation (Figure 4.5) (top). Where the Top and Bottom halves are manually segmented (Figure 4.5) (bottom), an improvement over the whole is noted, with ColorStructure DS 256 combined with max operator providing the best result (0.15).

Processing automatically segmented foreground data using these techniques yields similar results. These regions will exhibit a higher incidence of missing components and background contamination than the manually segmented data. The performance of the ColorStructure DS degrades moderately, from 0.21 to 0.27. The ColorLayout DS shows most sensitivity to noisy data: its retrieval rate drops from 0.58 to 0.73. Using the proposed system the same high retrieval rate is not maintained if the source camera for the query image is different from the source camera for the stored data set.

For Experiments II and III the query image is captured from a different camera to that providing the images selected for retrieval. These experiments are conducted with and without the Gray World preprocessing to improve the colour constancy, shown in Figures 4.6 (top) and 4.6 (bottom). All segmentation is fully automated. The best results are obtained using the ScalableColor DS and the DominantColor DS with an ANMRR of approximately 0.4. Splitting the subject into Top and Bottom and combining these in different ways did not significantly improve the performance. If the Gray World preprocessing is not used then the performance drops significantly: see Figure 4.6 (bottom) (dotted line). The best performing descriptor without Gray World is DominantColor DS giving an ANMRR of 0.5.

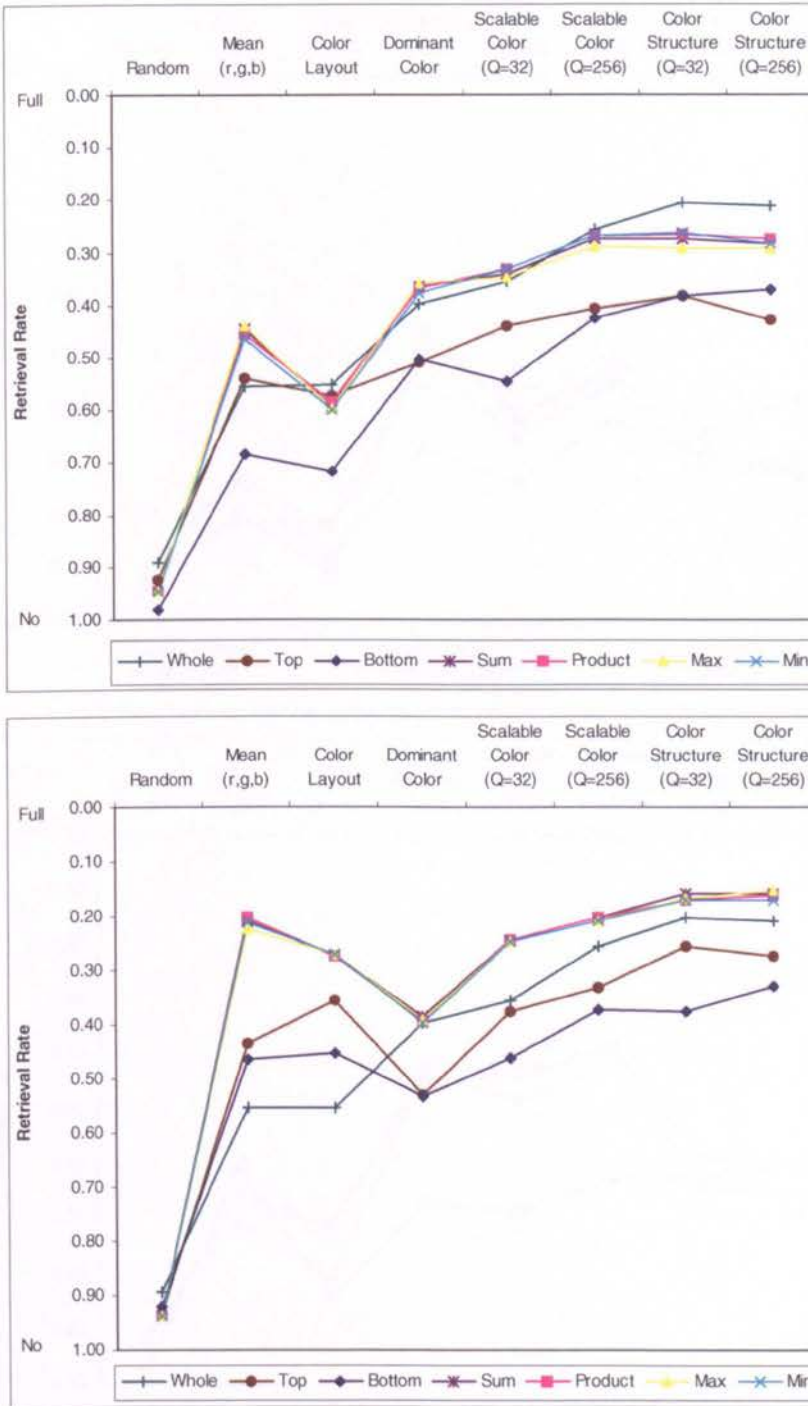


Figure 4.5: Retrieval Rate for Experiment I (same camera, different direction of movement) with manual segmentation with automatic split into Top and Bottom clothing (top) and manual segmentation with manual segmentation of Top and Bottom clothing (bottom). The most effective descriptor in this scenario is the ColorStructure DS (top). Compare with the results for automatic segmentation.

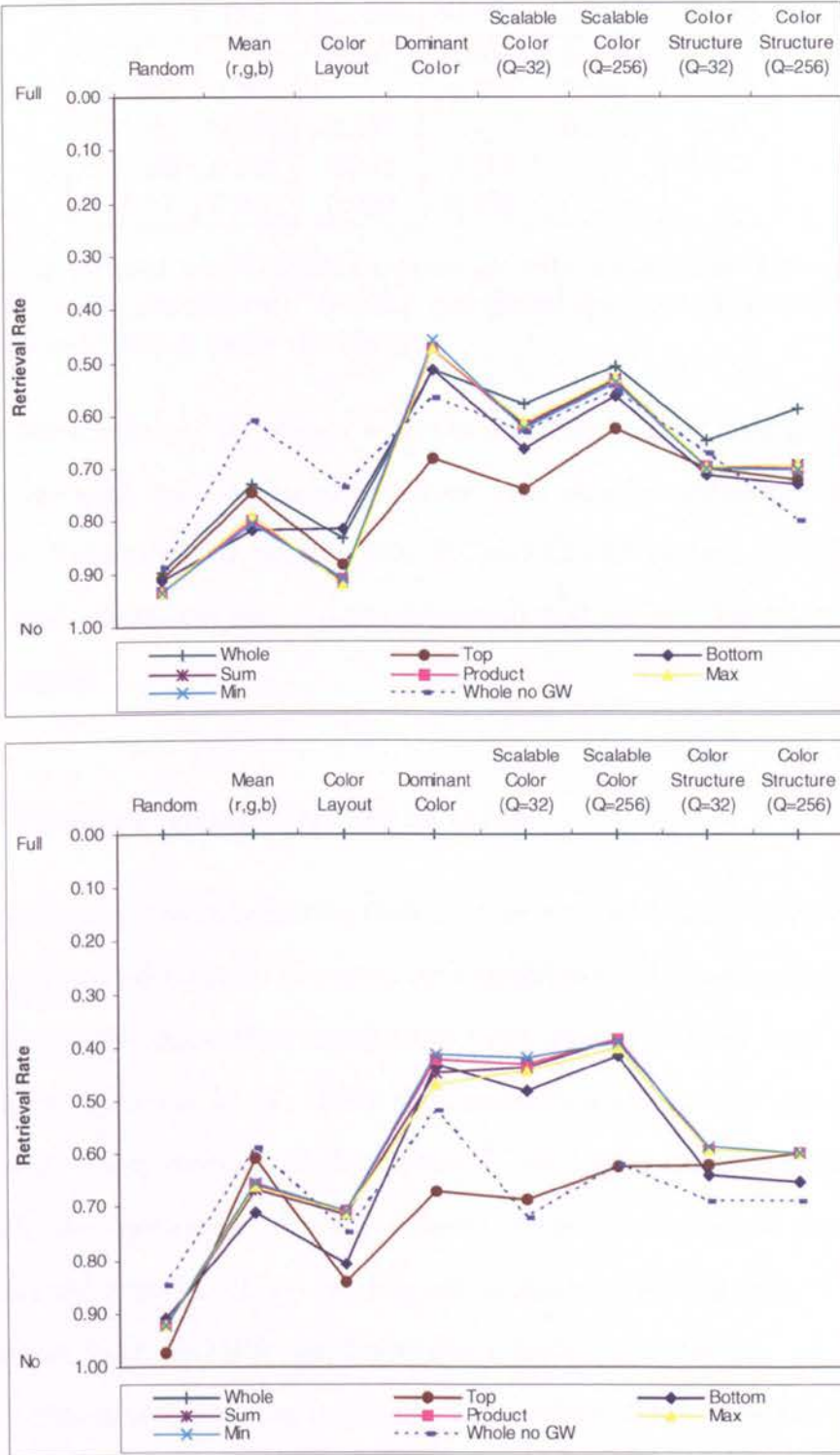


Figure 4.6: Retrieval Rate for Experiment II (different camera, same direction of motion) (top) and Experiment III (different camera, different direction of motion) (bottom). Both are automatically segmented and preprocessed for colour constancy. Data without colour constancy is provided for comparison.

-	DC	SC 256	SC 32	CS 256	CS 32
DC	-	0.307	0.369	0.272	0.244
SC 256	0.305	-	0.408	0.248	0.290
SC 32	0.347	0.357	-	0.292	0.331
CS 256	0.278	0.252	0.317	-	0.268
CS 32	0.256	0.292	0.335	0.258	-

Table 4.2: Descriptor combination experiments with Experiment I automatically segmented data, combined with the Min (minimum) operator. The results suggest an improvement over a single descriptor.

Small improvements are shown with the combined colour descriptors. Table 4.2 plots retrieval rates obtained in Experiment one for automatic segmentation policy. Combining DominantColor DS and ColorStructure DS with a Min operator gives a retrieval rate of 0.244 compared to 0.253 or 0.424 when these are used separately.

4.4 Fusing multiple features

In this experiment the Information Gain metric, reviewed in Section 3.3.4, measures the gain in information provided by a descriptor. Without using any prior knowledge the likelihood that a particular individual is selected from the GENERICK data set is one in 47. Using this metric it is shown that this likelihood is increased. Combinations of descriptors is used to increase this gain in information. An assessment of the scalability of the descriptors is also made by performing the experiments on random sub-samples of the data set. The results are shown for both ANMRR and Information Gain calculations.

Two extra descriptors are included in the experiments: the Co-occurrence texture feature, reviewed in Section 3.4.1, and a spatial feature based upon a person's height. These additional descriptors are able to be combined with the MPEG-7 colour descriptors to reduce uncertainty without colour correlation.

4.4.1 Height feature

The spatial descriptor is the height of a person calculated² using a camera calibrated ratio of pixels-to-metres px/m from the y -position in the image. This ratio is calculated from some static features within the image, in this case, office partition panels, at 1.22m tall. No roll angle, zero skew, unitary aspect are assumed. Perspective distortion is calculated using the top of the image as a constant c and the height y directly proportional to it where the vanishing point is outside the image. The top image of Figure 4.7 illustrates the calibration data. Linear interpolation provides the mapping, and is known as the ppm (Perspective Projection Matrix). A person's height is calculated by $h_m = h_{px}/ppm(y)$.

Due to constraints in the data, it is conjectured that there is a high probability that a person will often be partially occluded. In order to detect this occlusion, all the static occluding regions within the image are masked (Figure 4.7 (bottom)). When extracting the height there should be a gap between the y position of the bottom of the person mask and the occlusion mask or the images' extreme edge. Segmentation quality is important in this respect. When a pedestrian is excluded as a result of this test, all other related results for this pedestrian and the other descriptors are also excluded, *i.e.* from the joint statistical model.

4.4.2 Experimental design and methodology

The data set described in Section 4.2 is processed to provide six images for every pedestrian. This gave a total of 282 (6×47) images. Subsets of the 47 pedestrians are randomly sampled containing between five and 45 subjects in increments of five. As described in Section 4.3, an adaptive background subtraction algorithm is used and the meta-data is generated with the MPEG-7 Reference Software.

²The author gratefully acknowledges the contribution of Mr. A. Colombo in calculating the height feature.



Front Camera Calibration

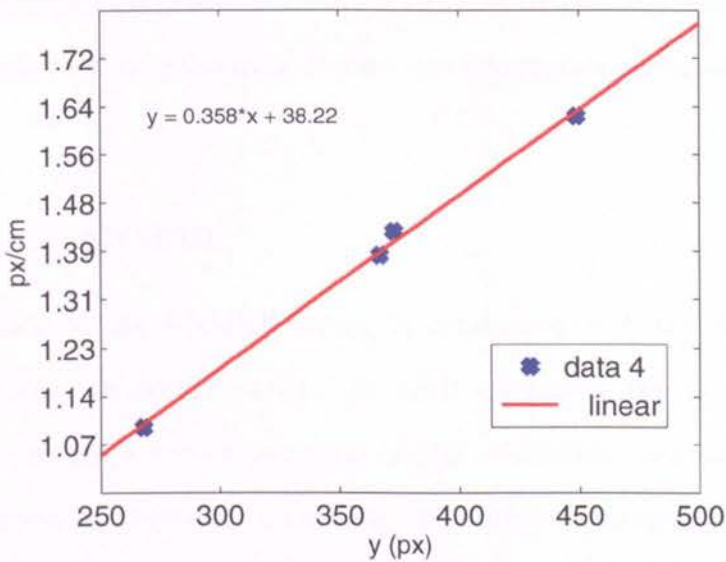


Figure 4.7: Calculating the height descriptor: the sample image with calibration guides (top); the calibration ratio between pixels and centimetres (px/cm) shown with calibration points (blue blobs) and the interpolated data as a line (middle); the manually created occlusion mask (inverted for clarity) (bottom-left); examples of the discarded observations (bottom-middle and bottom-right).



Figure 4.8: Left: Example query subject. Right: Six sub-images comprising object data set.

The Gray World processing and Mean and Random features are included. The automatic segmentation rules ensured six distinct images are generated of a minimum size and that neighbouring frames are adequately separated, as shown in Figure 4.8.

Measuring the ANMRR

The calculation of the ANMRR values is conducted with the same method as previously shown in Section 4.3.1. At each increment the samples are taken 30 times to provide a robust estimate of the evaluation measures for a typical retrieval scenario by effectively increasing the number of samples. This calculation is performed in two modes: for same camera data; and for different camera data.

Measuring the Information Gain

At each increment, the experiment is design so exactly one of the n elements in the data set has the same identity as the query subject. The data set is split into training and testing sets using a 'soft-partition'. The training set is 50% of the data taken at random and the test set is the second half of the data. This means it is possible for some samples to be in both training and test sets. For the generation of the ANMRR ranking, an additional experiment is carried out

where the query data is compared with three elements allowing the computation of ANMRR values.

Matrices are produced for the MPEG-7 colour descriptors (Section 3.4.2) selected in Section 4.3. The Co-occurrence texture feature³, described in Section 3.4.1, is generated using an eight-way or all-direction pixel relationship with a distance of one pixel apart. The choice of the all-direction pixel relationship is suitable for this data set since the texture orientation is unknown for each pedestrian. These experiments use the entropy measure (Section 3.4.1) to summarise the matrix statistics.

Deriving the p.d.f.s For the Information Gain metric, described in Section 3.3.4, it is necessary to calculate the probability that a retrieved image is correct. Matrices are generated containing the distances between all pairs of people for every descriptor. This results in a ‘confusion matrix’ from which both triangular halves are duplicated. The leading diagonal is ignored, since these zero values are the result of ‘like with like’ comparisons. Each matrix is split between *true* (Correct) and *false* (Incorrect) matches. True matches are considered to be values generated from observations of the same subject at different times and False matches are values from observations of different subjects. The accumulation of frequency histograms for both True and False matches can be used to estimate the probability that a given value of a match measure is True (or False), provided that the following is valid: the number of samples is sufficiently large, the statistics are stationary⁴, and an estimate is available of the prior probability that a measurement is caused by a True match, before to the actual observation of the match value.

³The author gratefully acknowledges the contribution of Dr. V. Leung in calculating the texture features.

⁴Stationary means all the samples within the data set have the same statistical properties.

The frequency histograms from Figure 4.9 (top-left and top-right), are accumulated from the DominantColor DS (Description Scheme) (left) and the ScalableColor DS (right). Parameterised versions of these histograms are shown in the Figure (2nd row) and use a Gaussian approximation taking the mean and variance. The functions show the probability that a *given* measurement or observation taking a specific value belongs to each distribution, *i.e.* the probability density functions (p.d.f.s).

Bayes' Theorem is used to generate the posterior probability that any given distance is likely to be from either the True or False distribution. The conditional probabilities are the distances between images in both the True and False p.d.f.s. The priors used have unit probability mass. Figure 4.9 shows (3rd row) an equal prior for the True and False outcomes and an unequal prior (bottom row), namely 0.1 for the True distribution and 0.9 for the False. It can be observed from these figures that the DominantColor DS provides a poorer separation between true and false matches than does the ScalableColor DS.

To estimate the posterior probability that two observations refer to the same individual, the prior probability needs to be provided alongside the match measure calculated from the two observation descriptors. Conceivably, this prior could be set at a constant value appropriate to the number of people under surveillance in a given temporal window. This assumes that an observation is *a priori* equally likely to be any of those people. More realistically, the value of the prior can accommodate as many of the additional cues available to the system, notwithstanding the actual value of the measurement, *e.g.* bias towards or against the most recently observed people, the patterns of behaviour previously observed, *etc.*

Furthermore multiple descriptors can be used in a joint estimate of the probability of a True match. In this joint probability space the density of available

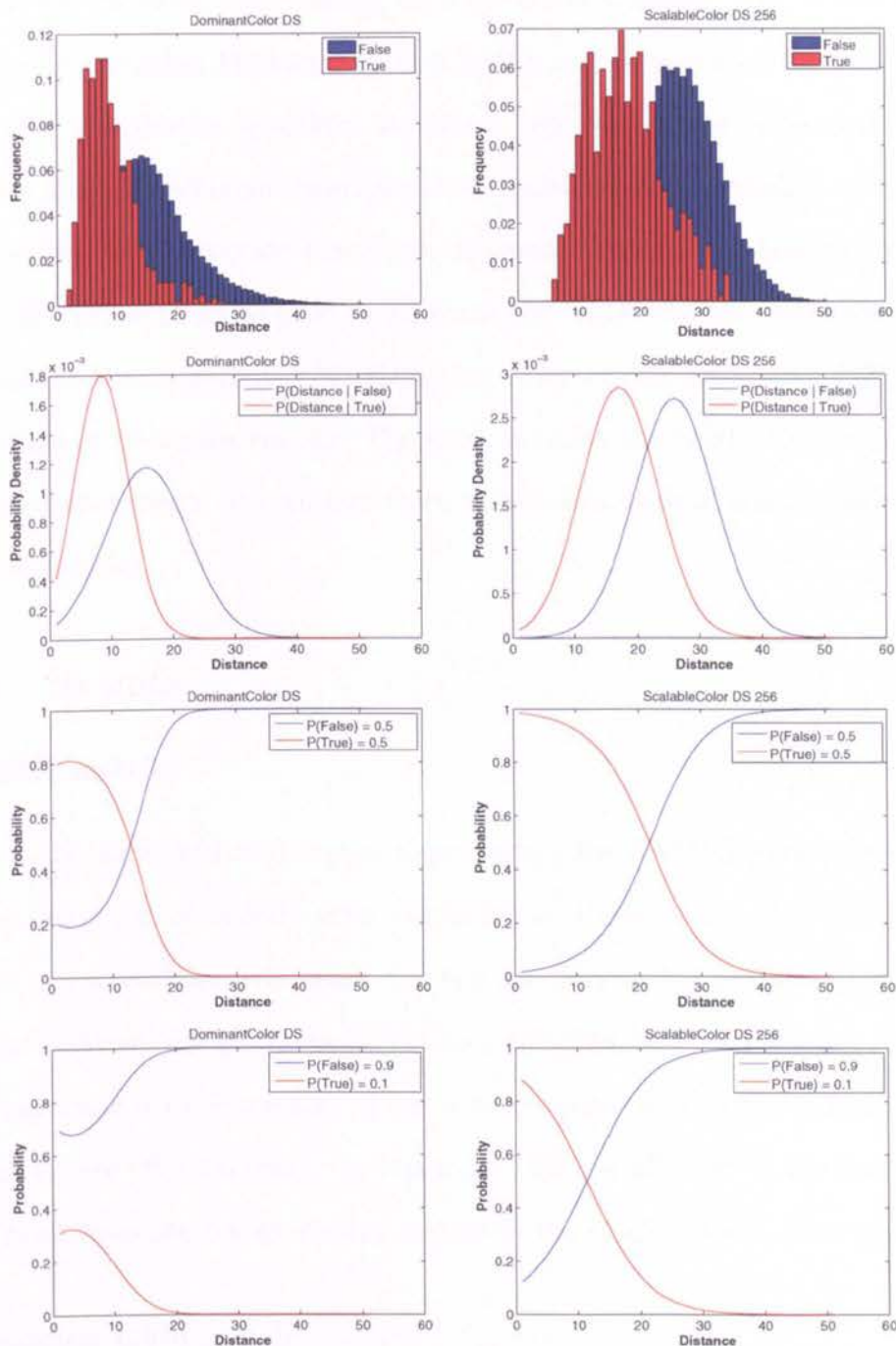


Figure 4.9: The frequency distributions for True & False matches (top) as a function of DominantColor DS (left), ScalableColor DS (right). The probability distributions for these give the probability densities conditioned on True and False matches in Gaussian form (2nd row). The Bayesian probability functions are shown with an equal prior (3rd row) and one example of an unequal prior (bottom).

samples will decrease significantly, so a parametric model, *e.g.* a multivariate Gaussian or Gaussian Mixture Model (GMM), will be necessary to estimate the probability distribution function. Despite a high correlation between the match measures from the different descriptors, the multivariate approach is attractive in its capability to incorporate cues from disparate sources. In these experiments, Netlab [SW06] software is used to generate the GMM models, estimated by the Expectation-Maximisation (EM) algorithm. The models estimated, $I(\mathbf{X}, Z)$, are the average of 10 measurements. The mask used for the height feature is applied in all the experiments to maintain feature vector dimension compatibility in the GMM estimation.

4.4.3 Results

ANMRR metric

For both the single and dual camera experiments, the ANMRR shows the retrieval rate declining monotonically with increasing n (Figure 4.10), *i.e.* the weighted number of correct elements inside the top six matches. It is interesting to note that the performance of the ScalableColor DS (256 bins) is the same for single and dual camera experiments. This is in contrast to the performance of the ColorStructure DS (256 bins) which is some 15% less effective in the dual camera experiment than the corresponding results in the single camera experiment.

Information Gain metric

For the single descriptors the most information is provided by the ScalableColor DS 256 (Figure 4.11) (top)⁵. Assuming there is an equal likelihood for each person to be found and the initial uncertainty is $\log(45) \approx 3.8$ nats, or 5.7 bits.

⁵The author gratefully acknowledges the contributions of Dr. V. Leung in calculating the Information Gain measures, depicted in Figures 4.11 and 4.12.

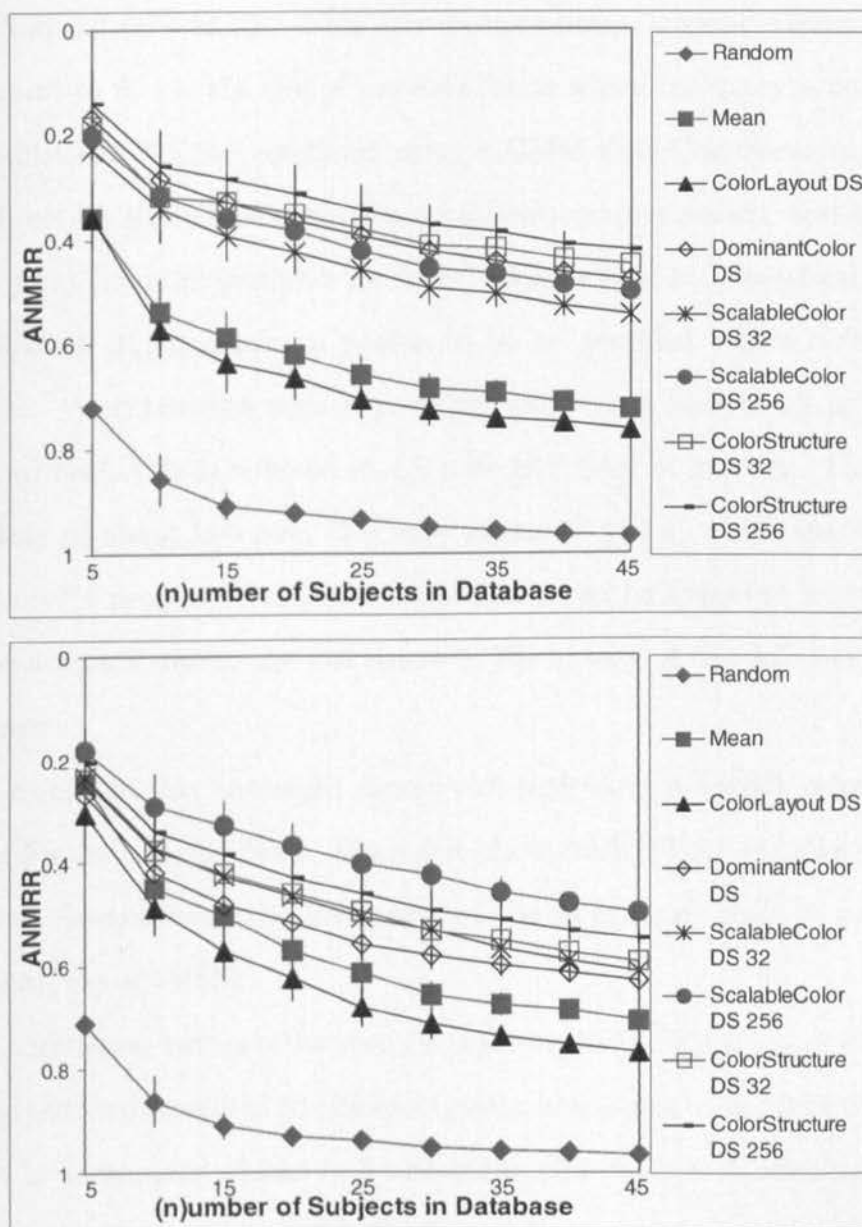


Figure 4.10: ANMRR results: single camera (top); dual camera (bottom).

For a data set size of 25 and upwards, a single colour descriptor can remove approximately one bit of entropy. An initial uncertainty of 3.5 nats (choosing one from 35 people) is reduced to an uncertainty of 2.5 nats, *i.e.* choosing one person from 12 ($e^{2.5} \approx 12$). This gain in information appears relatively stable with respect to n , *i.e.* the size of the data set to which the query is compared.

ScalableColor DS 256 combined using a GMM with ColorStructure DS 256, an all direction Kullback-Leibler Co-occurrence texture matrix and the height (Figure 4.11) (middle) provides the most information. In a practical example, an application might require a person to be re-identified within a data set of 47 people. With the information provided about each individual, as shown in this experiment, this is reduced to 1.9 nats (2.7 bits) of entropy. The residual uncertainty of about two nats (2.9 bits) means $e^2 \approx 7.4$, where the data set is reduced to 7.4 people. Real-world performance can be expected to reduce this posterior estimate due to the differences in the quality of the data extracted for each person.

The results for the combined descriptors with their ANMRR values is illustrated in Figure 4.11 (bottom). The result of the ANMRR metric, show in Figure 4.10 (top), demonstrates the Information Gain metric can produce results consistent with the ANMRR.

The correlation between the metrics is illustrated by Figure 4.12 which compares the retrieval results of 10 different combined features. The biggest difference is shown in a swapping of first to fourth place. The Spearman correlation coefficient gives a high correlation of 0.8788.

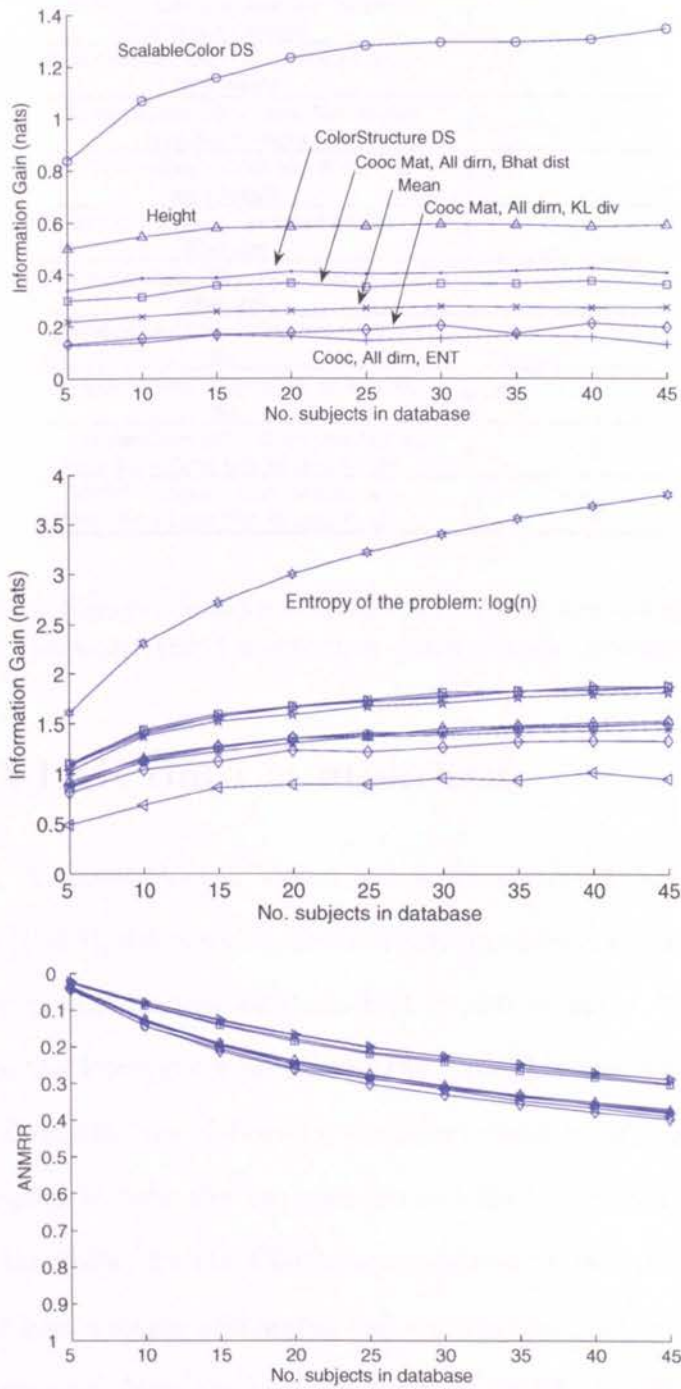


Figure 4.11: The Information Gain metric on individual features with the validity mask (top). On combined features using a GMM (Gaussian Mixture Model) (middle). The ANMRR metric is computed from the ranked results of the Information Gain metric (bottom).

	Info gain	ANMRR
MPEG-7 colour + Cooc Mat All dim, Bhat dist + Height	1	2
ScalableColor DS + Cooc Mat All dim, KL div + height	2	4
ScalableColor DS + Cooc Mat All dim, Bhat dist + Height	3	3
MPEG-7 colour + Cooc Mat All dim, KL div + Height	4	1
MPEG-7 colour + Cooc Mat All dim, Bhat dist	5	6
ScalableColor DS + Cooc Mat All dim, Bhat dist	6	5
ScalableColor DS + Cooc Mat All dim, KL div	7	8
MPEG-7 colour + Cooc Mat All dim, KL div	8	7
ScalableColor DS + Cooc Mat All dim, Bhat dist + Cooc Mat All dim, KL div	9	10
MPEG-7 colour + Cooc Mat All dim, Bhat dist + Cooc Mat All dim, KL div	10	9

Figure 4.12: The different feature combinations (left) are listed to demonstrate the correlation between the Information Gain metric (middle) and ANMRR (right).

4.5 CAVIAR data comparison

The CAVIAR (Context Aware Vision using Image-based Active Recognition) data set series [Fis04], discussed in Section 3.9, provides a further means to evaluate the experimental framework described in this section. Two data sets are generated from the Portuguese section of the CAVIAR test set. 120 individuals are extracted from the ‘cor’ (short for corridor) camera. 66 individuals are extracted who appear in both the ‘cor’ camera and the ‘front’ facing camera. These data sets give the ability for the CBIR (content-based image retrieval) evaluation to be executed over a single and across two cameras on standard data. The individuals are segmented from the background from which the MPEG-7 colour and Co-occurrence texture feature vectors are generated. The results are evaluated with the ANMRR and Information Gain metrics. It is not possible to use the Height feature in this CAVIAR evaluation because the appropriate calibration

data is not available.

4.5.1 Building the data set

The images of people are automatically extracted by segmenting the moving regions using a modified Stauffer and Grimson algorithm [KB01] supplied by OpenCV. CAVIAR Ground Truth (GT) is annotated with the CVML (Computer Vision Markup Language), described in Section 3.9. The GT identifies individuals and tracks them through the image sequences. These identifiers allow sets of images to be generated for each different person. Some noise and shadows left from the segmentation are eliminated by ignoring the area outside the GT bounding-box. The segmented images are then ordered on their pixel count and the larger ones are used for the data set. There is a minimum of one second of time between the images, *i.e.* 25 frames, and images with less than 10 pixels in either height or width are removed.

The GT annotation indexes each individual uniquely for a single camera, but the identifier is not unique across cameras. This means the outputs from the two cameras had to be matched manually, *e.g.* to ensure that person ‘A’ in camera ‘cor’ is the same as person ‘A’ in camera ‘front’. Certain people who appear in one camera do not necessarily appear in the other and subsequently the ‘front’ camera data set is smaller in size. Some poor or incorrect segmentations are also removed, based upon a subjective quality measure about how readily an individual can be identified. Additionally, the images are chosen so there is little background clutter and no moving occlusions. When certain individuals reappear in other sequences, their catalogue of images is updated with the new images.

The camera footage contains people at a medium to far range, at a low resolution, and many of the images of people are small, at about $50(w) \times 100(h)$ pixels

for the ‘cor’ camera and $30(w) \times 50(h)$ pixels for the camera ‘front’. Generally the quality of camera ‘front’ is poor and identifications are often not possible without cross-referencing the footage from camera ‘cor’. The ‘front’ camera suffers from colour artifacts affecting the blue colour channel. An example of the data set is shown in Figure 4.13.

4.5.2 Experimental procedure

The data is grouped into two indexed catalogues, one for the test set and the other for the query images. The query image for each individual is automatically chosen by selecting the image with largest masked pixel area. The experiments undertaken use data from either same camera or from both cameras. In the latter case, the queries are from the ‘front’ camera set and the test set is from the ‘cor’ camera. The image data between the cameras is normalised using the Gray World technique, as described in Section 3.6.1. Randomly varying subsets of the test set are selected in the experiments using the method described in Section 4.4.3. The varying number of images per each individual, as shown in Figure 4.13 (bottom), require modifications to the experimental apparatus. The ANMRR has the ability to accommodate such variation but, due to constraints on development time the ANMRR calculations are excluded from the evaluation using the Information Gain metric.

Three ANMRR experiments are carried out: single camera; cross-camera with; and without the Gray World normalisation. Subsets of the data sets are taken in steps of five, ranging from 10 to 120. The same four MPEG-7 descriptors are used, as described in Section 4.3, although the quantization levels for ScalableColor DS and ColorStructure DS are set at 256 for the combination experiments.

The evaluation using Information Gain uses data from the single camera data

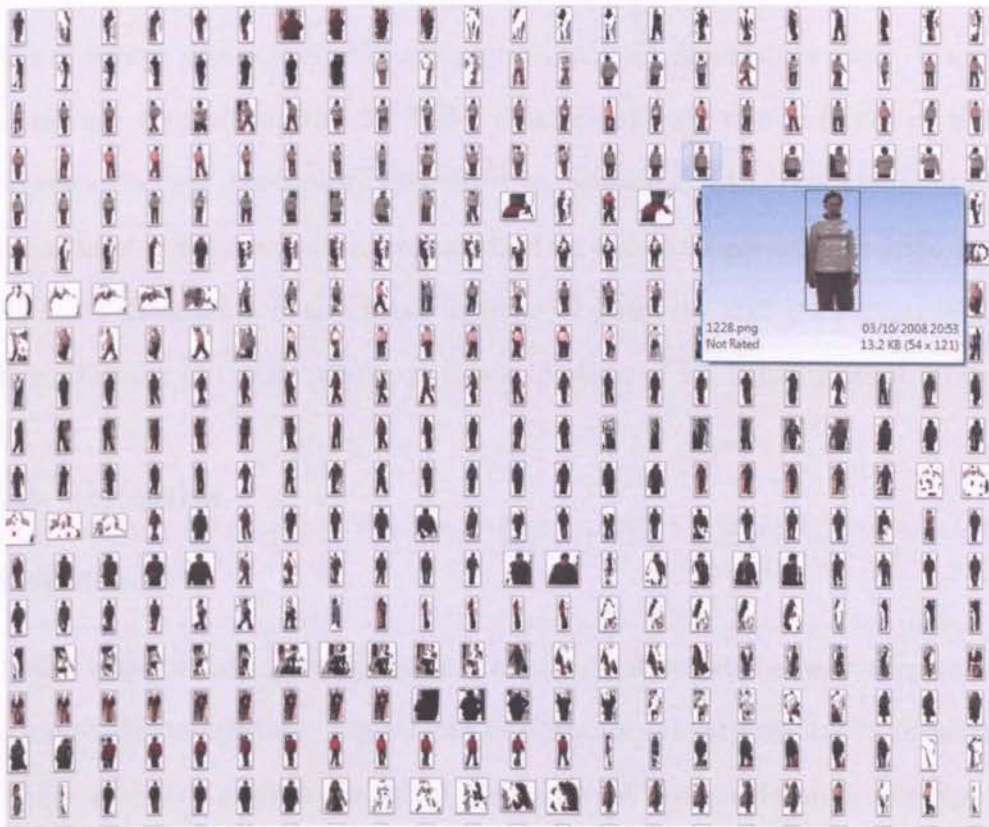


Figure 4.13: A portion of the GENERICK CAVIAR single camera data set is displayed (top). A closeup of one individual is shown. The second image (bottom) demonstrates the variability between images of an individual. For the same camera data set, the number of images per individual ranges from three to 71. Three is the minimum number for viable ANMRR calculations.

set. The procedure is identical to the procedure stated in Section 4.4, except for the different number of subjects and additional descriptors used. Confusion matrices are created for the MPEG-7 descriptors and two variants of the Co-occurrence texture descriptor, described in Section 3.4.1. The Mean, Random and Height descriptors are all excluded. The matrices are 1033×1033 in size, where each dimension is the total number of subjects and their representative images. Subsets of the data set are taken in steps of 10, ranging from 10 to 120.

4.5.3 Results

ANMRR metric

The same camera data ('cor') (Figure 4.14) (top), illustrates a near monotonic decline for all the descriptors. The overall performance is very similar to those shown in Figure 4.4, with a few exceptions. The order of best performing descriptors is the nearly as previously, with ColorStructure DS 256 giving the best performance, followed by ColorStructure DS 32, ScalableColor DS 256. ColorLayout DS outperforms DominantColor DS and ScalableColor DS 32 which may be the result of corrections to the MPEG-7 XM [ISO03c] source code. The poorest performing descriptor is ScalableColor DS 32.

The two camera experiments, with and without the Gray World processing, are displayed on Figure 4.14 at the middle and bottom, respectively. The results are again similar to those shown in Figure 4.6, in Section 4.3, whereby the performance of the descriptors across two cameras is poor. The difference being the Gray World algorithm does not noticeable improve the performance.

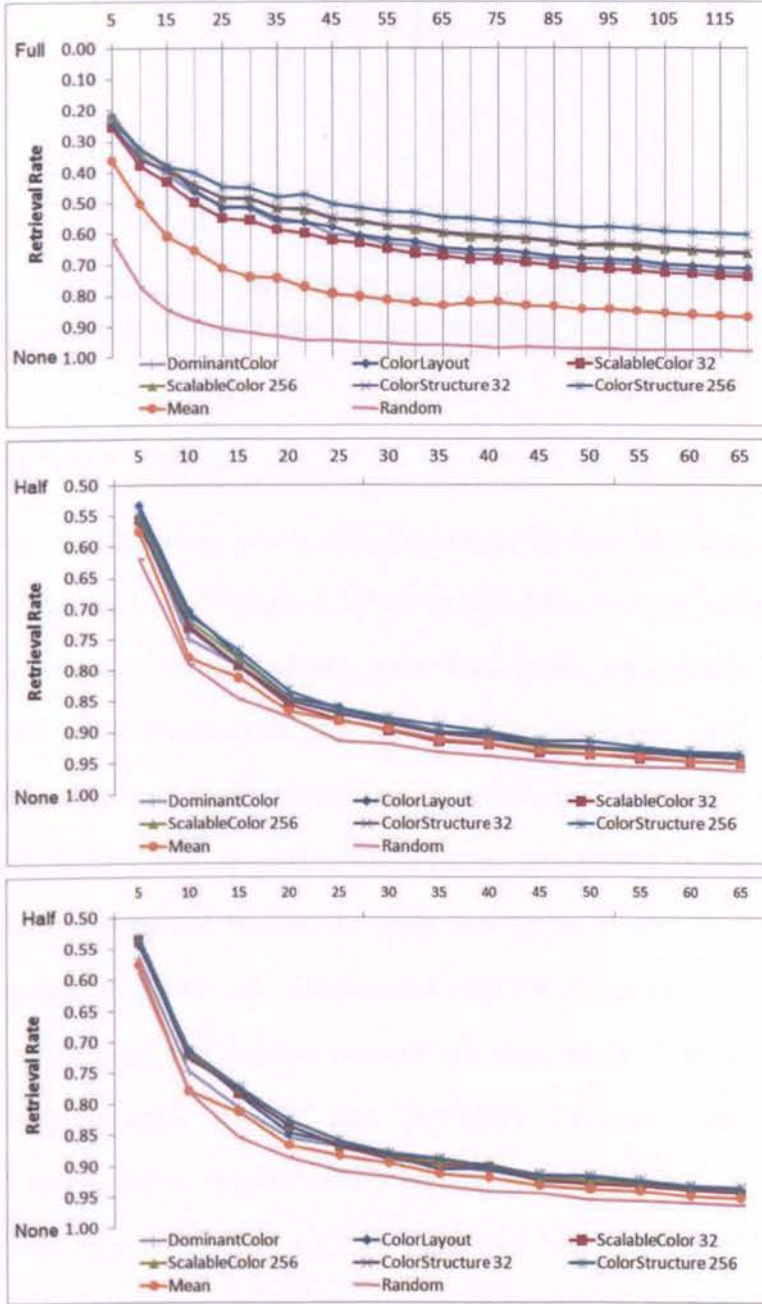


Figure 4.14: The MPEG-7 colour descriptors including the simple Mean and Random features. The results are from the same camera data (above), the two camera data with Gray World processing (middle) and without (bottom).

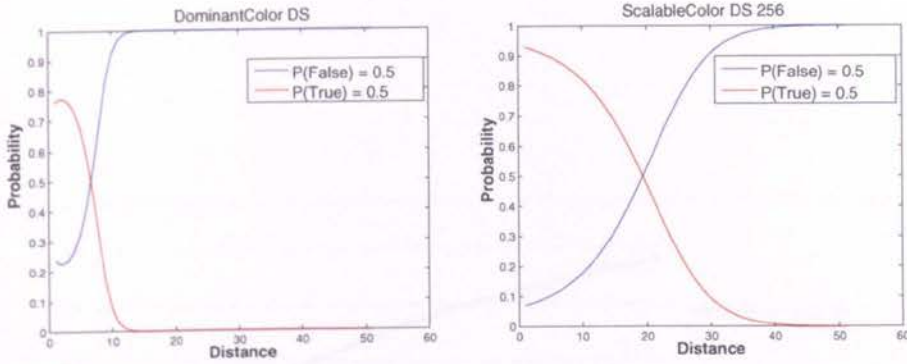


Figure 4.15: The Bayesian probability functions of DominantColor DS (left), ScalableColor DS (right) for True & False matches on CAVIAR data with an equal prior.

Information Gain metric

Examples from the Bayesian probability functions for the True and False samples are shown in Figure 4.15. Section 4.4.2 gives the details concerning this calculation. They are similar but not identical to those given previously, in Figure 4.9. The differences in the DominantColor DS could be due a new implementation in the MPEG-7 XM or due to a reduced range of colours in the data set.

The results from the Information Gain metric are shown in Figure 4.16. The Figure is split between the results for each descriptor singly (top) and combinations of descriptors (bottom). DominantColor DS is shown to have the most information singly and can remove one bit of uncertainty from 40 subjects (onwards) reducing the entropy to 2.7 nats (3.8 bits). This effectively increases the likelihood of retrieving a correct individual to one in 15 ($e^{2.7} \approx 15$). At 120 subjects the entropy is reduced to 3.79 nats (5.5 bits) giving $e^{3.79} \approx 44$. The descriptors perform slightly worse than with the previous experiments, with the exception of DominantColor DS, as shown in Figure 4.11 (top). This demonstrates the effects of poorer data quality.

The results for the combined descriptors show ScalableColor DS combined

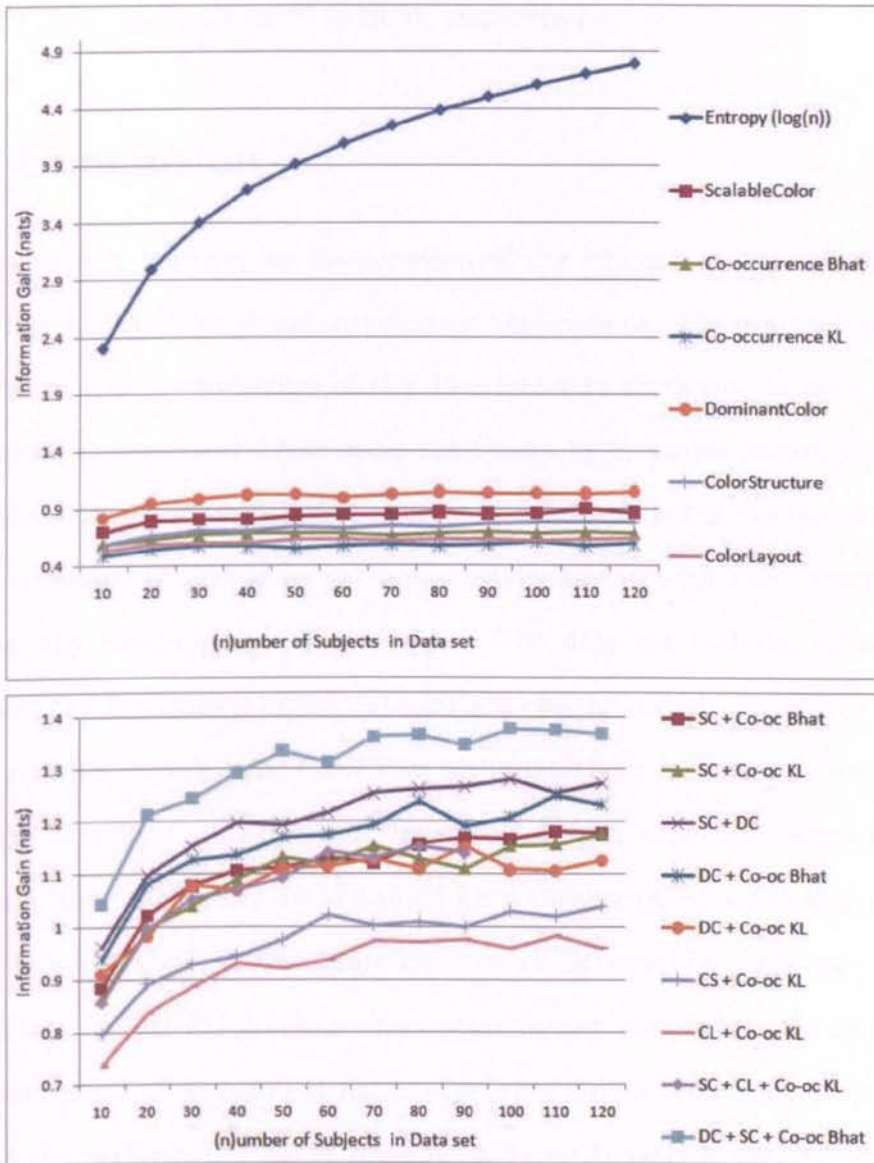


Figure 4.16: The Information Gain metric results on CAVIAR data. The results are shown for MPEG-7 Colour Descriptors and Co-occurrence texture singly (top) and combined (bottom).

with DominantColor DS and Bhattacharyya Co-occurrence texture provide the most information, reducing the entropy to 2.4 nats (3.5 bits) with 40 subjects and 3.42 (4.9 bits) with 120 subjects. The uncertainty is reduced to one in 11 ($e^{2.4} \approx 11$) and one in 31 ($e^{3.42} \approx 30.6$), respectively.

4.6 Conclusion

The experiments perform an examination of the efficacy of the colour features provided by MPEG-7 for visual surveillance applications. The evaluations demonstrate the relative performance of the descriptors in retrieving images of people in an indoor environment whilst using the Query by Example paradigm.

A data set called GENERICK is created, as described in Section 4.2. GENERICK contains 47 people in an indoor environment with two cameras which have partially overlapping fields of view. The data set includes segmentation masks and the limitations of the data set are clearly stated.

The ranking experiments show that although there are cases in which a segmentation into ‘Top’ and ‘Bottom’ improved results, especially when manually segmented, the best results are obtained for a single camera when all foreground data is input into the ColorStructure Description Scheme (DS). For multiple camera data sets the MPEG-7 colour descriptors do not outperform the simple r, g, b mean description of foreground data. Applying colour constancy preprocessing improves the performance but it remains still significantly below the level of the single camera retrieval results.

ANMRR (Average Normalised Modified Retrieval Rate) provides an unbiased and bounded indication of the performance of the retrieval process. It has particular strengths in comparing data of differing sample and Ground Truth (GT) sizes, *i.e.* the number of correct matches. It fails, however, adequately to address

certain issues familiar to Visual Surveillance researchers. For example, the rank ordering method cannot in itself provide evidence that a given query example does not appear in a data set: there will always be one element of the data set most similar to the example. Probabilistic estimates of identity for incorporation with other uncertain clues are not easily deduced from the rank method.

The performance of MPEG-7 colour descriptors is then assessed using the established information-theoretic metric Information Gain. The MPEG-7 colour features are evaluated using a confusion matrix which generates two histograms of successful and unsuccessful matches over the data set. The information gained through observation of the best performing descriptor, ScalableColor DS, is approximately constant at one nat with respect to the scale of the data set on which the query is performed. The work shows that the ScalableColor DS contains the most information of any MPEG-7 colour descriptor and has demonstrated the use of standard evaluation techniques in an innovative context.

Combining the MPEG-7 colour features with robust features for texture and height are evaluated using a GMM (Gaussian Mixture Model) model. The most effective combination is the KL (Kullback-Leibler) divergence with all-direction co-occurrence matrices and combined MPEG-7 descriptors and height. This is consistent over a variety of subsets from the data set. A comparison with the ANMRR ranking metric demonstrates that the Information Gain metric has similar performance. This framework provides a flexible way of improving the standard descriptors through aggregation using probabilities which the ANMRR metric cannot provide.

The MPEG-7 colour descriptors and the Co-occurrence texture feature are evaluated on the standard data set from the CAVIAR project [Fis04]. A new data set is generated from the CAVIAR data set of an indoor scene from a Portuguese shopping centre. The data set has two parts: one single camera part and one two

camera part. Each set contains images of a series of uniquely identifiable people. The single camera set contains 120 people and the two camera set contains 66 people. The individuals are automatically segmented from the background and the CAVIAR GT is used to help associate the segmentations with unique people and also remove background noise.

ANMRR calculations are then made on these individuals, in a similar fashion as given in Section 4.4. The main differences being the increased number of people, owing to the CAVIAR data set's size. There is also a varying number of true-positive examples of each individual. The findings for the single camera data set concur with what has been established previously. The improved performance of the DominantColor DS shows the importance of correctly performing computer algorithm implementations. For the two camera experiments, there is a noticeable drop in performance, even more than found previously (Figure 4.6). This is likely to be due to the difference in quality of the images from the 'front' camera compared with the 'cor' camera. The images from the former camera are smaller and often noticeably discoloured with a blue hue. The Gray World algorithm does not improve the results. It is clear that a level of similarity is required between multi-camera systems or image degradation will render the Gray World application ineffectual.

The Information Gain metric is applied in Section 3.3.4 to the CAVIAR data on the single camera data set only. DominantColor DS is shown to contain the most information individually (Figure 4.16) (top). This difference, as previously noted in the Figure 4.11 (top), is likely to be due to an improved implementation within the MPEG-7 XM. The combinations show DominantColor DS with ScalableColor DS and Bhattacharyya Co-occurrence texture to have the most information.

Chapter 5

The Content Description Interface

5.1 Introduction

The need for and the lack of a content description interface for CCTV (closed-circuit television) meta-data has been established by the previous Chapters. A standard video surveillance media format will facilitate systems in exchanging information about the contents of the video data as *meta-data*. The meta-data shall contain a variety of tools for basic information descriptions, annotations and automatic analysis. This Chapter explains the design of such a content description interface intended as a foundation for further standardisation and subsequent integration.

The first Section reviews the user requirements for a standard content description interface and looks at the *Functional* and *Non-functional* requirements. The former are focused upon what various interested parties want from the meta-data. The latter are concerned with the meta-data's quality attributes.

The second Section contains the details of the proposed solution architecture.

Three main paradigms are incorporated: *Technical* and *Observation* meta-data; Classification Schemes; and identity preservation. The Technical meta-data is concerned with the camera and the equipment and Observation meta-data describes the activity within the video. Classification Schemes are meta-data taxonomies, and the benefits in their use is explained and a standard method for their description is presented. The scenario is considered where these components are used to observe people by a network of cameras. The capability to add information derived from other sensors is important. This requirement is related to the issue of incorporating *uncertainty* in any given scene description. A design is presented which shows how low-level features can provide probabilistic evidence for a higher-level description of the individual's identity.

Details of the problem of refining existing technology, *i.e.* MPEG-7, so that it better accommodates the desired functionality are given followed by the structural details of the design. This includes an architecture where the meta-data is split accommodating the different semantic levels of a collection and the items it contains.

5.2 User requirements

5.2.1 Functional

The main requirement of the meta-data is that its scope is restricted to describe a video scene with a finite amount of detail. This can be compared to a more ambitious proposal, where the majority of all the possible meta-data that a scheme can describe in surveillance is accommodated. As a consequence, meta-data about authentication, data-integrity and non-video media, *e.g.* audio, is not considered.

The technical requirements, enumerated in Table 5.1, were formed in MPEG (Motion Picture Experts Group) meetings and have been reproduced from associated public domain output documents [SD07].

The elements of the list are categorised into equipment and scene activity, Technical meta-data and Observation meta-data, respectively. The former describes equipment and settings and the latter describes the video content. For items marked as ‘required’, this meta-data must be included and emphasises the importance the designers have given to the type. The remaining items are optional. A mechanism should be supported to allow the values and labels of user entered meta-data to conform to a user supplied Classification Scheme. This scheme will be specific to a particular application scenario, *e.g.* railway station, airport, shopping centre; convention for labelling cameras; or a lexicon for describing events. Finally, it is necessary to be able to describe the meta-data from different video formats, *e.g.* MPEG-4: (Part 2) or legacy formats, especially video cassettes which many surveillance systems still use.

5.2.2 Non-functional

There are certain non-functional requirements which are incorporated into the proposed meta-data interface. These characteristics are related to performance and can only be qualitatively analysed although some evaluation of these properties is required. Reviews of the basic quality attributes required for meta-data have been undertaken [Wri02]. For surveillance, the following five elements have been identified and are now described below:

Extensibility It is often useful to extend an existing standard to accommodate a new or specific requirement. Indeed it is rarely feasible for a standard to accommodate all possible scenarios and the MPEG-7 standard is already a large

1. Technical meta-data – The format must be able to store:

- 1.1 An accurate record of when the data was captured (required)
- 1.2 Equipment identifiers
 - 1.2.1 ID tag for the camera (required)
 - 1.2.2 ID tag for current, preceding, and succeeding fragments (required)
 - 1.2.3 ID tag for the cluster to which the camera belongs
 - 1.2.4 ID tags for each of the multiple streams from a single camera
- 1.3 Description of equipment and its settings
 - 1.3.1 Description of camera make and model
 - 1.3.2 Description of camera settings, *e.g.* shutter-speed values

2. Observation meta-data – The format must be able to store:

- 2.1 Annotation of events:
 - 2.1.1 In a free-text format, along with an event ID (unique in the file only) and a time-stamp
 - 2.1.2 In a semantically structured format
- 2.2 The location of an object of interest in one or more frames of video data, in image co-ordinates
- 2.3 The colour appearance of an object of interest, in order to enable retrieval of observations of that object from multiple video sources
- 2.4 The ID of objects observed in one or more video sources, to enable cross-referencing

Table 5.1: List of functional requirements for the meta-data.

body of work. Users may need to include proprietary data-structures alongside the existing meta-data.

Information compatibility The usual benefits of standardisation include: a reduction of costs associated with product development and support; and a greater chance of producing a profitable product. For meta-data to be considered compatible with systems, each common element must have an agreed interpretation, *i.e.* mandated normative behaviour, *e.g.* The ITU (International Telecommunication Union) UTC (Coordinated Universal Time) time-code [ITU02]. A *caveat* with XML standards is that it is possible only to standardise the syntax but not the content. The possibility of a wide variation of the content's semantic interpretation means standardisation of the semantics is not possible, *e.g.* text annotation and the difference between the colours described in colour descriptors.

Any extended meta-data should remain forwards and backwards compatible, as described in Section 2.3.3. This situation is not required in the commercial market where the emphasis is on product profitability. Where the data is of high value a proprietary format may be desirable, otherwise, if possible, a well-designed future proof technology should be used from the outset.

Flexibility The meta-data should be flexible enough to describe general surveillance situations while remaining compatible. Flexibility allows the same syntax to support multiple semantic differences, though, this increases the likelihood that standard products are not compatible with one another due to different interpretations. Rigidity limits the descriptive power of the meta-data but improves compatibility, *e.g.* binary data versus XML.

Computational complexity The meta-data should be encoded and decoded efficiently with a low number of CPU (central processing unit) processing cycles.

There is a conflict between quality and speed and the format should support efficient mechanisms to allow devices to ignore unused items. Recursive structures are not recommended since these require significant computer resources and are difficult for humans to understand.

Indexing for catalogues The meta-data should be able to be indexed allowing data persistence and rich data manipulation via databases. To achieve this it is necessary for a meta-data catalogue to be maintained separately from the media [Wri02]. Unique identifiers can be mapped to ‘primary keys’ which facilitates the mapping of the meta-data to database tables.

5.3 Architecture of the VSCDI

The Video Surveillance Content Description Interface (VSCDI) is the result of a body of work which comprises the following steps: the need for a standard format for visual surveillance; the importance of standard meta-data; an analysis of existing meta-data standards; the creation of a framework for encoding and describing uncertainty; and the gathering of requirements for a standard format for visual surveillance meta-data. The outcome is the VSCDI and the overall scope made by this contribution is given in Figure 5.1, containing the Technical, Observation and Classification Scheme meta-data.

A draft of the VSCDI specification is included in Appendix A. The following subsections describe the components of the VSCDI and comprise: Technical, Observation and Classification Scheme meta-data; identity preservation; profiling MPEG-7; and the division of meta-data between *file level* and *track level*.

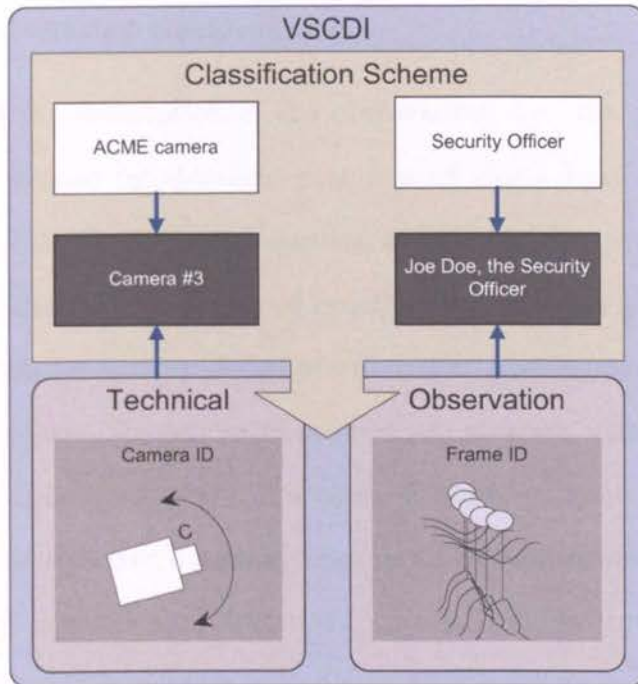


Figure 5.1: The scope of the proposed meta-data is technical (left) and observation (right). User-defined labels and identifiers for equipment and individuals (top).

5.3.1 Technical meta-data

The fundamental characteristics of the production, or technical information, are the time and identity of the sensor. Varying degrees of information will be known about the static content of the scene. A reference to a database containing geographical features, *e.g.* coastline, contours, roads, buildings, street furniture, could be used to provide spatial co-ordinates for reasoning about scene kinematics and occlusions. Additional information about the production can be included to assist with content analysis, *e.g.* the position of the sun relative to the optic axis, and weather conditions. The internal and external camera parameters that define the calibration transform between image and ground co-ordinates are considered technical.

5.3.2 Observation meta-data

The structure of the description of the observation, *i.e.* the semantic contents of the video, is defined by standard practices of video decomposition. At the lowest level, the video is defined as a group of frames. The meta-data requires a design which facilitates the process of isolating this group, *e.g.* time point, and providing the tools for further decomposition, *i.e.* frame granularity. Once the frames are decomposed the regions of interest are isolated, using a bounding box or such, and an annotation about the scene attached, if required. Annotations should be possible wherever necessary, *e.g.* as an overall comment on the video's contents. Further semantics are attached to the description through references in Classification Schemes.

5.3.3 Classification Schemes

A Classification Scheme is defined [ISO04a] as “descriptive information for an arrangement or division of objects into groups based on characteristics which the objects have in common”. A Classification Scheme is a term within the Meta-data domain, and the equivalent term *Taxonomy* is associated to phylogenetics in Biology. Two problems are solved: knowledge particular to a discipline is managed by grouping terms to form a lexicon for the application, and relations are defined between taxonomic terms or *taxa* to enhance their meaning.

The Holmes database system, developed by Surrey University and the Surrey Police, includes a Classification Scheme of the terms used by British police forces [Got06]¹. Some of the benefits of Classification Schemes are given in Section 2.4.2. As depicted by Figure 5.2, hierarchical structures organise term relationships in three possible ways:

¹The system is not to be confused with the Unisys HOLMES 2 (Home Office Large Major Enquiry System) investigation management system [HOL07]

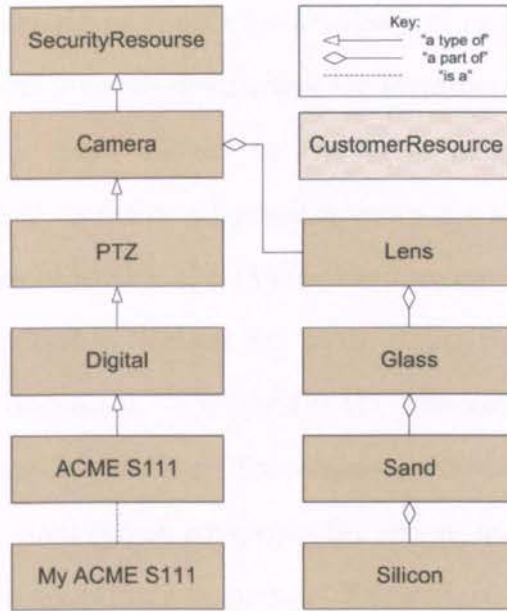


Figure 5.2: A Classification Scheme illustrating the three different possible relationships between taxa. The relationships are either *type of*, *part of*, or *is a*.

- **a type of:** A relationship to increase the aperture of meaning from wide to small, *e.g.* an *ACME S111* is **a type of** *Camera* and the relationship can be expressed as $\text{Camera} \rightarrow \text{Surveillance} \rightarrow \text{PTZ} \rightarrow \text{Digital} \rightarrow \text{ACME S111}$.
- **a part of:** A relationship expressing the direction of meaning from the general to the specific, *i.e.* a decomposition of meaning, *e.g.* *Silicon* is **a part of** a *Camera* and expressed as $\text{Camera} \rightarrow \text{Lens} \rightarrow \text{Glass} \rightarrow \text{Sand} \rightarrow \text{Silicon}$.
- **is a:** Thirdly this work proposes using the Classification Scheme dynamically to express the instantiation of *objects*. This follows the paradigm of a *class - object* relationship, *e.g.* *My ACME S111* **is a** *ACME S111* expressed as $\text{ACME S111} \rightarrow \text{My ACME camera}$.

A major design consideration is for a CCTV meta-data content description

to be extensible. It should be usable for a variety of applications in the CCTV domain. While general purpose descriptors for identifiers, time, media and features can be specified, these will not be specific to any individual application. It would be impractical to define all possible terms for all possible applications. One attractive feature of ViPER-GT (Video Performance Evaluation Resource – Ground Truth), described in Section 3.9, is its ability for the user to define an ontology within the document. The ViPER-GT *annotator* defines the types or classes in the configuration section. The classes or terms are simple taxonomic associations, *e.g.* the class *person* comprises the classes *head*, *body*, *arms* and *legs*, which are defined as subordinate to *person*. These classes are *referenced* within the document body. This is the standard use of terms defined in a Classification Scheme. Terms related to car park surveillance could be, *e.g.* ‘PTZ’ = a PTZ (pan-tilt-zoom) camera, ‘private car’ = a vehicle with 1-7 seats, ‘security officer’ = an employee to protect the premises. Using a standard Classification Scheme will result in a machine readable taxonomy recognisable to a standard decoder, but extensible for user-defined descriptions.

5.3.4 Identity preservation meta-data

Visual surveillance data will not be gathered in isolation. In many scenarios there will be other sources of information, *e.g.* swipe-card, vehicle registration data, or human operator input, to augment the information manually or automatically extracted from the visual data sources. This includes input from multiple cameras, as described in Section 3.6. All sources can be used collectively to update the relationships between low and high-level descriptions to provide the most informative representation of the scene. The topography of such a system is illustrated by Figure 5.3. The outputs of the automated processing are likely to contain

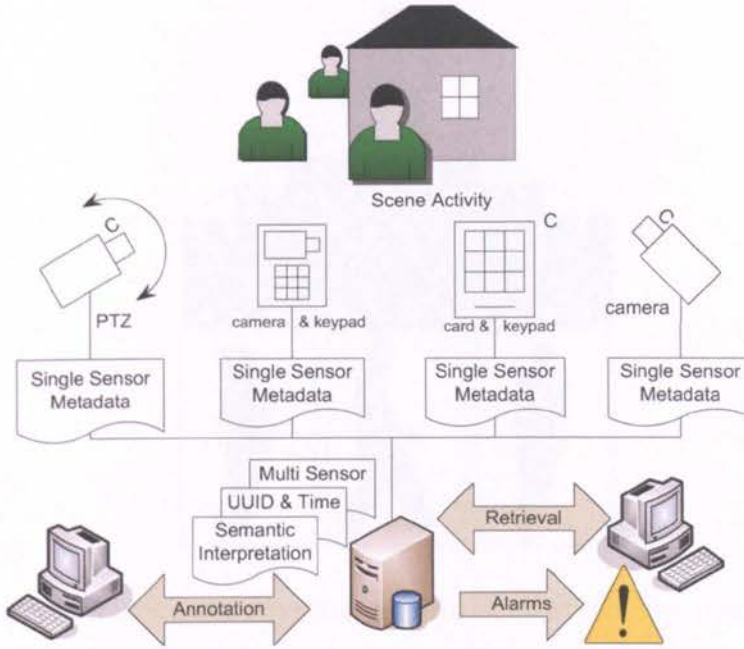


Figure 5.3: A multi-sensor CCTV system process flowchart.

errors or not be certain. With this in mind an extensible content description is defined for accommodating this uncertainty.

Low Level Objects

The features of the moving objects as observed by a single camera are considered as low-level. The spatiotemporal extension of such an object is expressed as a *segment* or a Low Level Object (LLO). This is illustrated in Figure 5.4 (top): a series of bounding boxes, spanning its temporal extension, *i.e.* the number of frames for which it is observed. It may contain certain properties, *e.g.* a colour description. As with the ViPER standard (see Section 3.9) it is possible to include concurrent multiple segments describing the same object: for example, the independently moving limbs and torso of a body. Similarly, it should be possible to include consecutive segments describing the same object, *e.g.* before and after an occlusion. In this case, the different LLOs will be associated to one another via a link to a single Global Unique Object (GUO). The probabilistic



Figure 5.4: The camera sensor produces meta-data about the Low Level Objects (LLOs).

links can be generated from Bayesian analysis (Section 4.4).

Global Unique Objects

The LLOs are of limited use to the end-user, since they are not consistent across viewpoints, and they are not associated with any persistent identification of the object (person, vehicle, *etc.*) in the scene. To generate useful semantics about the scene activity, it is helpful for the objects to be referenced by consistent unique identifiers, known as a Global Unique Object (GUO). For a multi-camera surveillance system we assume that it is an objective for each agent in the scene to be represented by exactly one GUO. If other data sources are available then external denominations of identity can be associated with the GUO. For example, electronic passes, keypads, vehicle recognition devices can all provide this type of input. This can be an automatic real-time process, or an off-line manual annotation.

The proposed strategy is to define links between each GUO and one or more

LLO from one or more cameras. At any given instant, and for any particular camera, a LLO will have more than one link to a GUO, if either a) the inference is that the LLO is a merged region, and represents two GUOs simultaneously or b) there is a degree of uncertainty about which unique identifier this region represents. For example, if two people disappear from one field of view, and appear, separately, in another field of view, it is possible that the surveillance system (human or machine) will not be able unequivocally to state which way round the people are positioned. This uncertainty is accommodated by allowing the estimate of any GUO location to be split between more than one LLO with probability of association distributed among them.

There are constraints on the links between the GUO and LLO, and also on the movement of the GUO, that may be exploited by a particular surveillance system. For example, the probabilities have unit probability mass: over each LLO or GUO graph and each LLO and GUO object relation; or a GUO cannot be in two places at the same time. These constraints are not included in the proposal, but rather can be adhered to, in order to maintain a consistent semantic interpretation.

The basic design is illustrated in Figure 5.5. The Media component (top), contains the Technical meta-data, while emphasising the calibration components. The identifier information is contained in the 'media details' box. The LLO information is shown (middle) and the GUO information is shown (bottom).

5.3.5 Profiles and levels

A critical appraisal of the existing video surveillance meta-data is described in Section 3.9, of which MPEG-7 and ViPER-GT (Video Performance Evaluation

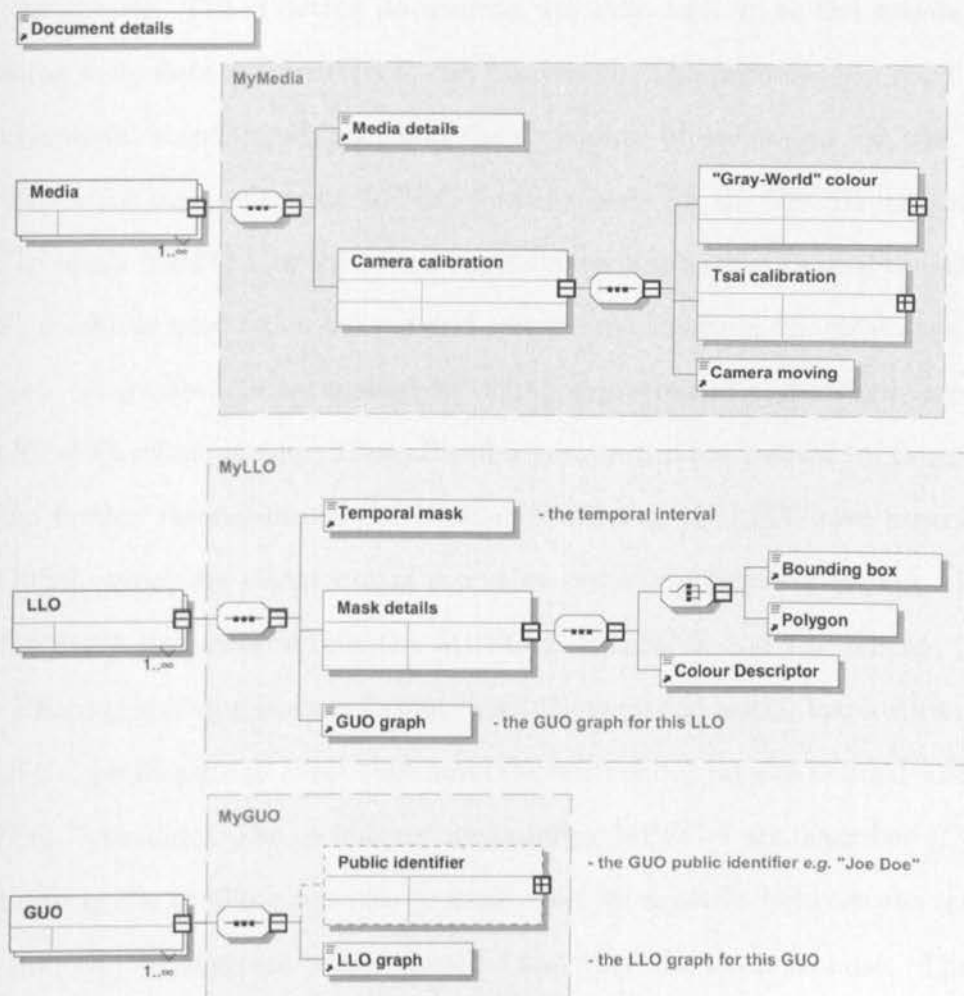


Figure 5.5: The structure of the VSCDI, from the perspective of identity preservation, showing the technical information about the camera and media (top), the LLO (middle) and GUO (bottom) observations.

Resource – Ground Truth) seem the most appropriate as a basis for a new description interface. ViPER-GT has strengths over MPEG-7 both through its acceptance within the Visual Surveillance community and its simple design. By using MPEG-7 as a dictionary the scope of MPEG-7 can be narrowed while managing its complexity. The resulting documents will still conform to the standard, and existing compliant software tools can be reused. The promise of a solid design, international standard support and the capability of restricting the size has led to the choice being made for MPEG-7 as the basis for the VSCDI interface.

To apply MPEG-7 to the visual surveillance domain the Description Schemes (DS) available need to be filtered and interpreted to select the most appropriate subset. A profile is a term used to define a portion of a standard, *e.g.* AVC [ISO03a] Baseline profile. These Profiles can be further refined by Levels, with define further restrictions of the Profile. Profiles of MPEG-7 have been defined [ISO05d], which are either out of scope, or not comprehensive enough. Profiles of MPEG-7 also exist within the MPEG-A standards and the scheme used in the Photo player application format [ISO07b] provided initial inspiration. These unofficial profiles are in effect standards themselves but are not defined within the MPEG-7 standard. The techniques for profiling MPEG-7 are described [ISO05e], and during the profiling process, in some cases an exact fit between the intended use and the requirement is not possible and there is a compromise. The main goal in the selection of the meta-data types is to keep the specification simple. The main guidelines to achieve this are as follows:

- Appropriateness of the description to the requirement
- Reuse of types where possible to minimise the proliferation of different types
- Nested structure to be kept as shallow as possible

- Overall schema length to be as short as possible
- Prohibition of circular references (to place upper limit on complexity)
- Avoidance of problematic or ambiguous DSs
- Duplication of meta-data should be unambiguous

A new profile of MPEG-7 schema is created based upon the functional requirements in Section 5.2.1 and the set of written objectives given above in this Section. Some re-engineering is required to enforce the simplified design, *e.g.* multiplicity of elements is appropriately restricted and recursive elements are removed and replaced with references. Fortunately MPEG-7 supports both methods. The end result is a data-model which reflects the scope of the application. Human understandability is improved and the significant likelihood that existing XML parsers would fail to parse MPEG-7 due to its size and complexity is reduced.

5.3.6 Structural details

The main design goal was to keep the schema as simple as possible, while incorporating useful functionality for a surveillance scene description. The VSCDI meta-data stream is designed for generation either in parallel to or after the generation of the video stream. The technical meta-data describes the equipment, *e.g.* camera type and camera identifiers. The observation meta-data describes segmented regions within the video and only when activity occurs. The Classification Scheme meta-data allows the linking of segmented regions with instances of real-world objects. Figure 5.1 shows an example a region is linked to the security officer called ‘Joe Doe’. A scene reconstruction is possible, *e.g.* the route a person takes through a building where their progress is described using *bounding box* information on a frame-by-frame level. Semantic descriptions can be applied

to the moving regions using text annotations and it is possible to use free text or a structured annotation format.

Meta-data choices

The challenge is to understand the semantics of the MPEG-7 types and to choose the simplest descriptors which accommodate the most functionality. MPEG-7 contains many types with significant overlap, there are types describing multimedia summaries, mathematical content, analytically edited video, the media, the creator, media with ink and handwriting, user behaviour and numerous feature descriptions. Many of these are not necessary for a video surveillance meta-data description. It may be tempting to choose types with sophisticated characteristics, *e.g.* trajectory descriptions using point interpolation, but the conservative nature of surveillance and the desire to keep the descriptions simple are more important. A detailed analysis is given by the author in the technical report [Ann07].

For the basic descriptions, it is not immediately clear what descriptor to use. Tools to describe the content are provided by the Creator descriptors which yield good functionality. This descriptor is focused on the requirements of production and the subtleties between the different types of timing information. This provides too much detail in the production and emphasises the creator as a significant entity within the system, when in video surveillance the creator is a simple machine. The type chosen fulfills its task well, but is not the immediate choice. Its name 'DescriptionMetadata' has a general meaning. The VSCDI scheme applies its own semantics to the elements. For example, the context for the description is not to describe the meta-data, as hinted by this type's name, but to provide an overall description of the content.

Some tools contain unnecessary complexity and are avoided, those describing

video, segments and frames. It is possible to describe a video in many ways. First it is necessary to understand what the video content is about, which in this case could not be described as a signal, a summary, or anything other than a simple video segment. Therefore it is necessary to choose the simplest element with the capability to describe some decomposition. Decomposition was required, in order to describe individual frames, but this brought up new problems. It is possible to choose between many types of decomposition, that is, temporal, spatial or both of these combined. Further choices include a discrepancy between moving region, video or still region decompositions. The described interface uses the simplest possible method of decomposing a video: a frame-by-frame decomposition, which requires a temporal decomposition and frame descriptors. There are, however, multiple ways of decomposing a frame many of which are complex and hard to comprehend. The MPEG-7 MovingRegion DS is a complex set of descriptors which can describe an object trajectory. The design of the MPEG-7 decomposition tools have been criticised and MPEG-21 (Part 17) – Fragment Identification of MPEG Resources [ISO06b] provides an improved solution. Each frame can contain a single simple region of interest, which further limits complexity. Of all the features contained within MPEG-7 the choice was limited to just two colour descriptors, based upon the results of the experiments shown in Chapter 4.

A single tool is available to describe the Classification Schemes, but its complexity and scope have been restricted. In general, most of the types in VSCDI have been restricted to manageable levels of multiplicity. This includes the removal of attributes which are not used, *e.g.* the attributes *overlap* and *gap* from the decomposition descriptors and *confidence* from the text annotation are removed, *etc.* The terms within a Classification Scheme, while being the only types within VSCDI allowed to remain recursive, are restricted to contain single descriptive elements, *e.g.* Name, Definition. These restrictions are intended to

facilitate software implementation. Of the many types available to describe a graph or mathematical model, the simplest was chosen, being the graph and relations structure. This provides the power required to describe the uncertainty with general description applicability.

The entire Semantic DS was removed even though it provides a comprehensive tool kit to describe *agents*, *objects* and *events*. It was abandoned for the simpler but adequate TextAnnotation DS. The semantic structures are complex and more suitable for manual creation because recursive structures are used throughout. The use of the Event DS is not included for a different reason as this term has become overloaded by companies in their proprietary event detection products. Inclusion would have been problematic because it would require each company to adapt their own *event* semantics to the semantics of this DS. It is less problematic for all interested parties to adapt their semantics to lower-level MPEG-7 elements.

No specific content management structures are provided: normally these are required to handle the semantics of a *collection*. The individual meta-data files contain unique identifiers and are valid on their own, which provides a loose structure. Of all the tools included in the framework the ones providing uncertainty descriptions have the most abstraction and can be overloaded to provide extra functionality, *e.g.* describing a collection. Additionally a file format could provide such a collection framework. In particular, the Collection DS is excluded to reduce complexity and also because it lacks the versatility of the MPEG-21 DID (Digital Item Declaration) [ISO05f].

5.3.7 File Level meta-data

The meta-data is split at two levels between File Level and Track Level. Figure 5.6 shows the structure of a File Level document (left). The File Level meta-data is designed for use at an over-arching level defining the circumstances under which the Track Level meta-data documents exist. This scope shall be defined when a set of Track Level documents is compiled. The core descriptive elements shall be a Universally Unique Identifier (UUID) [LMS05] and a UTC (Coordinated Universal Time) [ITU02] time-stamp. The time is recorded as the amount of time passed between the measurement and 1st January, 1601 with an allowable error of up to 100 nano-seconds or 10^{-7} seconds. It has no knowledge about any frames that were dropped. This number is converted into the MPEG-7 time format and is the earliest UTC time over all the tracks. This identifier and time-stamp allow the meta-data description to be indexed effectively.

The remaining technical meta-data comprises a comment, geographical location, and instrument settings. The comment could give the reason why the group of meta-data documents were collated and comprise multiple free text annotations and structured annotations. The MPEG-7 StructuredAnnotation DS defines the fundamental constructs within a sentence as the fields, *e.g.* *why*, *where*, *what action*, *when*, *what object*, *who* and *how*. The geographical information is the Global Positioning System (GPS) [Par96] of the compiler location. The instrument settings elements contain any information regarding the compilation processor. The equipment settings designator uses a ‘value–attribute–pair’ structure which allows multiple free text records defining the equipment settings of the compiler.

The observation meta-data comprises a graph of object relations used to annotate the GUO and LLO instances, described in Section 5.3.4. Multiple graphs

File-level meta-data		Track-level meta-data	
MPEG-7 (optional)		MPEG-7 (optional)	
Technical		Technical	
Required	ID	Required	Camera ID
	Time (ms)		Comment
Optional	Comment	Place	Instrument settings *
	Place	Cluster *	Stream ID
	Instrument settings *	Video time offset	Camera calibration *
Observation		Observation	
Optional	Graph depicting relations between file- and track-level objects *	Required	Video track
Classification Schemes			Capture time (ms)
Optional	Surveillance terms *		Duration (ms)
			Text annotation *
			Temporal group of frames (GOF) *
			GOF duration (ms) *
		Optional	GOF frame descriptions: Time, text annotation *, region (grid,box) *, colour (DC,SC) *, frame reference
			Graph depicting relations between track- and file-level objects *

Figure 5.6: The File Level (left) and Track Level meta-data (right). Bold denotes the element is required to be included in the description and * denotes multiple descriptions of this type are possible.

of relations can be stored, called *Relation* elements and the attributes of which are optional but comprise: *source*, *target*, *type* and *strength*. The Strength denotes the probability measure associated with the relation. The *type* refers to a Classification Scheme term and the *source* and *target* are used to define the objects. Zero or more Classification Schemes are also defined at File Level and the unique objects (GUO) are created here, and referenced by the Track Level meta-data documents. The basic design is illustrated in Figure 5.5, with a simplified File Level structure (bottom) showing the identifier and graph but not the equipment details.

MPEG-7 describes Classification Schemes using the extensible Classification-Scheme DS. Groups of *terms* are defined and relations can be expressed amongst them. Embedding terms within terms is permitted and allows a coupled structure. The relation allows fine-grained control in the type of relation the implied Superordinate \sim subordinate terms have, *e.g.* Superordinate \sim subordinate **Vehicle** \rightarrow **Bus** or subordinate \sim Superordinate **Bus** \rightarrow **Vehicle**. The terms can also be defined inline within elements outside the Classification Scheme or reference Classification Schemes from other standards or in other documents. External Classification Schemes can also be imported allowing reuse. The Structured Annotations can be embellished through a linkage to terms defined inline and to terms within Classification Schemes. For example, in a Structured Annotation, a *who* field could be annotated with the text *Joe Doe* and linked to the Classification Scheme describing this person's role, **Person** \rightarrow **Company** \rightarrow **Security** \rightarrow **Guard**.

When instantiated a File Level XML document may contain similar information as the example given in Figure 5.7. The example contains the basic DescriptionMetadata DS about the document *e.g.* a comment, an identifier, the creation location and time-stamp and instrument details. The Description DS contains

Description Metadata	Comment	FreeTextAnnotation	The VSAF reference software created this XML							
		StructuredAnnotation	Who	WhatObject	WhatAction	Where	When	Why	How	
	PublicIdentifier	4E9C95FA8C47428								
	CreationLocation	longitude	latitude	altitude						
		0.100000	0.100000							
CreationTime	2008-01-01T01:01:01:0F30									
Instrument	Automatically generated by VSAF Meta-data API									
		name	value							
Description	Relationships	type	source	target	strength					
		urn..process:obj:unknown1	urn..vs:infra:cam:ptz:cam1	urn..acme:people:sec:guard:joe_doe	0.8					
		urn..process:obj:unknown1	urn..vs:infra:cam:ptz:cam1	urn..acme:people:staff:fred_bloggs	0.2					
	Classification Scheme	<p>Term</p> <p>ACME: The ACME company classification scheme</p> <ul style="list-style-type: none"> • People: People recorded by the ACME surveillance <ul style="list-style-type: none"> ◦ Sec: Security are people who are work for the ACME surveillance system <ul style="list-style-type: none"> • Guard: Security guards to protect the premise and patrol <ul style="list-style-type: none"> ◦ Joe Doe: Joe Doe, the security guard Staff: Staff are the employees who work for the ACME company <ul style="list-style-type: none"> • Fred Bloggs: Fred Bloggs, the company employee <p>vs: The surveillance system</p> <ul style="list-style-type: none"> • Process: The definition of terms used in the computer processing components <ul style="list-style-type: none"> ◦ Obj: An detected object <ul style="list-style-type: none"> • Blob: The moving region within a video <ul style="list-style-type: none"> ◦ Known: The moving region within a video with a known identity ◦ Unknown: The moving region within a video with a unknown identity <ul style="list-style-type: none"> • Unknown1: The identity of an object of unknown identity Infra: The surveillance system's infrastructure <ul style="list-style-type: none"> ◦ Cam: Cameras watching and recording <ul style="list-style-type: none"> • PTZ: A camera with pan-tilt-zoom mobility <ul style="list-style-type: none"> ◦ 8A4E85D3-B2D8-4b99-A430-11032F0AAAF4: Camera over exit Clust: Camera cluster <ul style="list-style-type: none"> • 73D1C6F8-5405-4a82-B2AF-295FDD3B5D30: Cameras surveying entrances 								

Figure 5.7: An example File Level XML document containing the majority of elements available. The document describes itself, relations and a Classification Scheme.

a series of Relation elements (*type, source, target, strength*) the ClassificationScheme DS defining the terms within the surveillance system.

5.3.8 Track Level meta-data

The *track level* meta-data describes a particular instance of video media and is known as a *Camera*. Figure 5.6 shows the structure of a Track Level document (right). The basic technical meta-data shall comprise the camera's UUID and a UTC time-stamp of the recording's start time, with the same format as the

File Level time-stamp. This is the time of original capture, not the time of any subsequent recording, copying, annotation or editing. The optional information comprises the video's duration, a comment, the geographic position, equipment settings, the camera cluster identifier, the stream identifier, and camera calibration. The comment, geographic location and equipment settings have the same format as the File Level meta-data. The camera identifier may also be used as an external library catalogue reference, thereby handling video cassettes, *e.g.* an ISBN (International Standard Book Number) [ISO05a]. A camera cluster identifier could also be provided within the camera identifier. Alternatively, a more prolix description of the cluster can be annotated. The identifier of the cluster can be given along with many optional entries that reference Classification Schemes to give more context about the camera's domain. If the media is a trans-code of a non-compliant format, such as VHS (Video Home System) tape or a different coding scheme, an identifier can be provided to this external media instance along with its URI (Uniform Resource Identifier) [BLFM05]. If the camera uses more than one stream, the particular stream can be identified.

Since it is possible for the internal camera clocks to drift, the facility to accommodate a time-offset is required. This feature is designated as the *offset* field within the equipment settings meta-data. It is assumed that this drift will not be severe enough to change over the course of a single streaming session. It is intended that the media fragments should remain temporally short. Longer time periods are intended to be captured by chaining together media segments via predecessor and successor reference mechanisms.

It is possible to record the Gray World colour constancy matrix, described in Section 3.6.1, within the equipment settings. Multiple camera calibration descriptions are possible, with each one accommodating one camera position preset. The Perspective Projection Matrix (PPM) calibration technique is expected.

The capability of handling a moving camera is not officially supported but the equipment settings meta-data can handle this event by making an annotation at frame-level, shown below. Many of the descriptions, including camera settings, can be embellished with supporting entries taken from Classification Schemes, *e.g.* **Camera** → **Surveillance** → **PTZ** → **Location** → **Entrance**.

The observation meta-data shall comprise a *video track* entity. The duration of the video track is optional and describes the length of the video in terms of the time between the starting time and the ending time. The specification of the time format is as defined in Section 5.3.7. The video can contain a text annotation containing specifics about the video's contents or perhaps some arbitrary output from a bespoke algorithm, such as the number of people filmed within the sequence. This has the same format as the comment field, described above. The Classification Scheme can be used to label the annotation field, so the decoder can interpret the contents appropriately.

The granularity of the description extends to optionally defining a decomposition descriptor to annotate a group of frames (GOF) within the Camera track. Each decomposes a frame which contains a time-stamp to locate it in video time. Optionally, a text annotation, a region of interest (ROI) and a visual descriptor can be defined. The text annotation has the same format as the comment field described above. The annotations at frame-level could be used to describe the main events taking place in the video either manually or automatically. The annotation is linked to the frame with an identifier unique within the document, *e.g.* **camera#1/frame#1**. The frame or LLO can be linked to a single GUO only. To link more LLOs to the same GUO, additional frame annotations are required. The descriptive terms defined within the Classification Scheme are used. Multiple text annotations are possible and an example could be, "who(BlobA), what action(Alert), what object(Door), how(Loitering), when(2007-07-10 T14:07:00)".

The ROI can be a bounding box or a polygon. The annotation and visual descriptor, if defined, will then apply to this ROI, since a frame can only segment one region. The visual descriptor can be a `DominantColor DS` or a `ScalableColor DS`. A $M \times N$ grid can be defined, where M and N are arbitrary numbers. The visual descriptors can be applied to every cell in the grid. If it is necessary to define a new ROI for the same frame, then a new frame descriptor can be created with reference to the same media time point.

As with the File Level meta-data, probabilistic preservation of identities is supported using a graph structure, where multiple graphs of relations are possible. This time the link is upstream, *i.e.* the frames and decomposed regions are linked with the unique objects. The structure can also describe algorithm output with a probability value.

When instantiated a Track Level XML document may contain similar information as the example given in Figure 5.8. The example contains the basic `DescriptionMetadata DS` about the document *e.g.* a comment, an identifier, the creation location and time-stamp and instrument details. The `Description DS` contains a series of `Relation` elements (*type, source, target, strength*), camera calibration information related to each preset position, and the `Video` content description. The `Video` element contains the identification and location descriptors, textual annotation and a series of `StillRegion (frame)` descriptors. A single frame description is shown in the example, defined using a time-stamp and described with a text annotation. The frame contains two descriptions about the colour information within the frame. One defines a region with the `SpatialLocator DS` and the other describes the whole frame using the `GridLayoutDescriptor DS`.

Description Metadata	Comment	FreeText Annotation	Sensor details. Sensor details manually added						
		Structured Annotation	Who	What Object	What Action	Where	When	Why	How
	PublicIdentifier	4E9C95FA8C47429							
	CreationLocation	longitude	latitude	altitude					
	CreationTime	2008-01-01T01:01:01:0F25							
Instrument	Automatically generated by VSAF Meta-data API								
	name	value							
	Focus	1							
Description	Relationships	type	source	target	strength				
		urn..vs:process:obj:unknown	blob1	urn..process:obj:unknown1	0.8				
		urn..vs:process:obj:unknown	blob2	urn..process:obj:unknown1	0.9				
	Header	LocalCoordinateSystem							
		Preset 1: 0 00.700 0.400000 00.740000 0.1500000 00.742000 0.45000 PPM MappingFunction							
		Preset 2: 0 00.1200000 0.10700 00.1030000 0.100500 00.104000 0.80000 PPM MappingFunction							
		MediaIdentification				MediaProfile			
		73D1C6F8-5405-4a82-B2AF-295FDD3B5D30				file:///data/video.mpg			
		FreeTextAnnotation The details about the video							
		StructuredAnnotation							
		Video	StillRegion	FreeTextAnnotation	Spatial Locator	MediaTimePoint	Visual Descriptor	GridLayout Descriptors	
				Detected blob	1 2 3 4	2008-01-01T01:01:01:0F25	1705 4 4 0 0 0 ..		
				2008-01-01T01:01:01:0F25		011 1 1 30 6 - 76 ...			

Figure 5.8: An example Track Level XML document containing the majority of elements available. The document describes itself, relations, camera calibration information, the video and its frames.

5.4 Conclusion

This Chapter describes the creation of the Video Surveillance Content Description Interface (VSCDI). The architecture is based upon both functional and non-functional requirements and some further objectives: the means to describe the *technical* information about the timing, camera identification and other details about the equipment as well as *observations* about the content. It is anticipated that the latter will be produced both manually and automatically to describe events and object appearances. The provision for Classification Schemes is necessary to allow the definition of custom lexicons for any application providing value to operators and programmers, and the ability to define the terms for unique objects. Any visual surveillance system, manual or automatic, will sometimes need to express incomplete or uncertain conclusions, and this situation is helped by the incorporation of information from other data sources, *e.g.* RFID (Radio-frequency Identification), in to the overall strategy. The meta-data structures that express the results of the surveillance will need to accommodate probability in their structure and relations.

MPEG-7 provides the tools for these descriptions, but it is a large standard and needs to be profiled into a new schema. A core set of technical and observation meta-data is stored using MPEG-7 Description Schemes. A considerable number of standard MPEG-7 Description Schemes are not included in the proposal. The resulting profiled schema is 1032 lines long, including comments. For comparison the MPEG-7 Simple Metadata Profile (SMP) is 537 lines long, the MPEG-7 Core Description Profile (CDP) has 2364 lines and the entire MPEG-7 schema version 2004 is 9249 lines long. Distinctions are made between meta-data applying to the file as a whole and that which applies to a single track.

It is a substantial task to develop a profile of MPEG-7 that is suitable for

a significant proportion of the visual surveillance community. The VCSDI is designed to accommodate an appropriate balance between specificity and generality, in order to encourage stake-holders to adopt common approaches where possible, but nonetheless allow alternatives where necessary. The details of the implementation and examples are included in Appendix A.

Chapter 6

Evaluation of the proposed solution

This Section provides three evaluation studies of the VSCDI (Video Surveillance Content Description Interface). The first evaluation considers a scenario in which a surveillance system with multiple sensors assigns unique identifiers to pedestrians. A second evaluation examines three implementation processes that use the VSCDI. These are the MPEG-A (Part 10) Conformance and Reference Software [AS08], a comparison with a meta-data scheme from the European project ‘CARETAKER’ [IST], and the use of this VSCDI in a system entered for the ‘Grand Challenge’ project [UKG08a]. A third evaluation method concerns the DominantColor DS (Description Scheme) [ISO02c] feature component of VSCDI. A CBIR (content-based image retrieval) application is developed for an embedded system using a Query by Attribute (QBA) (Section 3.2) graphical user-interface.

6.1 Global identifiers for pedestrians

In Section 5.3.4 the preservation of the identity of objects moving under a multiple sensor network is discussed. A correct identity is established from concrete information, such as the response from electronic gates. The estimated position of the pedestrian is updated using feature descriptors from video data, based upon the method shown in Section 4.4. This method derives a probabilistic measure of the likelihood of correct identification. A probability measure is important, due to the varying quality of the visual processing. This is especially important within a multiple camera network, as discussed in Section 3.6 and illustrated in the experiments in Section 4. With the addition of timing information and local knowledge about the normal time intervals to get from place to place a likelihood map can be produced.

Chapter 5 proposes the VSCDI (Visual Surveillance Content Description Interface) which is designed to accommodate this facility of describing relations between low-level and high-level elements. This Section provides a qualitative analysis to establish how well this is achieved.

The method is illustrated in Figure 6.1 with a network of Low Level Objects (LLO) and Global Unique Objects (GUO) linked together. Three LLO instances are contained within this media segment *LLO#1*, *LLO#2* and *LLO#3*. There are also three GUO instances, *Joe Doe*, *John Bon*, and *Joe Bloggs*. Each LLO contains details of the temporal interval, a bounding box and the colour features that describe the video frame. The last LLO element is a graph of the relations to the different GUOs. The system associates different LLOs with GUOs according to the likelihood of a match based upon the data from multiple sensors. For example, *LLO#1* has a 60% chance of being *Joe Doe*, a 20% chance of being *John Bon*, and a 20% chance of being *Joe Bloggs*. In a closed-world scenario for

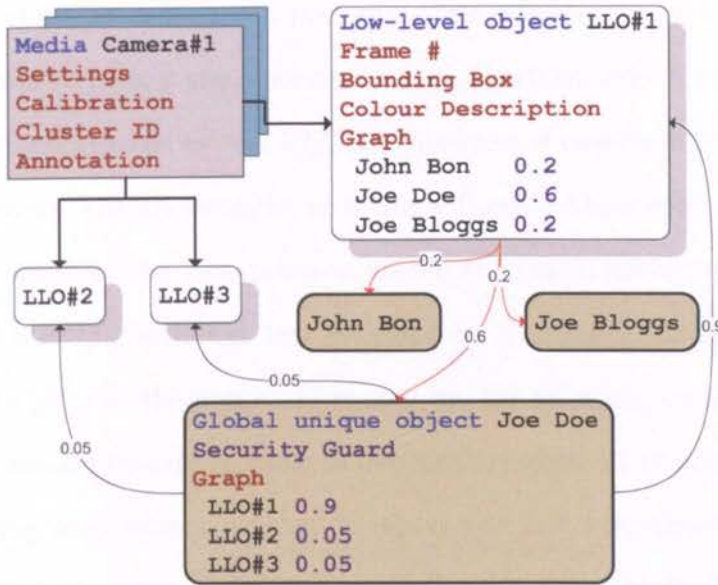


Figure 6.1: The scene is surveyed by multiple cameras. Relationships are built between the Low Level Objects (LLOs) and the Global Unique Objects (GUOs).

each LLO the sum of associations to a given GUO is unity. In an Open World this is the upper-bound. A similar situation is depicted with the GUO ‘Joe Doe’ except the GUO contains a graph of LLOs. On this basis the system can suggest to the end-user a list of possible identifications ranked in order of likelihood. Every observation will have a likelihood of being uniquely associated to a number of unique objects.

As discussed, low-level identifiers provide short-term links to a unique identifier using one or more low-level features. The LLOs express the features of the moving objects observed by a single camera, *i.e.* video segment. The video segment may have many LLOs and many LLOs may refer to the same GUO. The VSCDI records each LLO within a graph with a series of relations, each with a link to a GUO. The same goes for the GUO records. A system may maintain a history of these time instances and in this case, each frame iteration, *e.g.* five frames per second, will cause a new set of probabilistic relationships to be generated. Recording the set of relations produced at each iteration might be useful

in a data model that relies upon previous observations, *e.g.* Markov chain, or it might be useful to review the behaviour of an algorithm over time. Although not obviously accommodated by the VSCDI, this type of description can be achieved by defining every VSCDI segment as a single frame. Alternatively, an additional graph structure could be incorporated within the frame descriptor.

It is also possible, although less satisfactory, to include a single LLO to refer to multiple objects in the scene. This may be, for example, caused by a merged region that simultaneously represents two nearby objects; or caused by an error in the tracking association, so that it represents first one object, and then the other. In either situation, if detected, the reference to multiple objects should be removed where possible by splitting the feature appropriately.

Another issue is the provision of relationships between LLOs. The LLOs might be linked and form part of a group, *e.g.* a group of tourists or school children. The VSCDI can achieve this kind of description using the Classification Scheme mechanism. Multiple terms are allowed within each LLO relation, *i.e.* each Classification Scheme term allows a list of references. In the case of school children, a Relation could be linked to the GUO *Child* and the GUO *GroupA* at the same time. Alternatively, an entry within a Classification Scheme could be defined so a hierarchy is built with *GroupA* being super-ordinate to 'Child'.

6.2 Application of the VSCDI

This section provides an assessment of the performance that can be achieved using the format in real-world scenarios. Three scenarios are investigated: the reference software for MPEG-A (Part 10); to facilitate the data exchange necessary for the CARETAKER European project; the use of the VSCDI (Visual Surveillance Content Description Interface) in an entry for the Grand Challenge surveillance

competition.

6.2.1 Video surveillance application format: overview

MPEG-A (Part 10): Video surveillance application format (VSAF) [BR08], is an archival file format designed as an architecture for low-powered and inexpensive CCTV video capture devices. It contains precise timing information, invariant video data quality, and good meta-data annotation capabilities. The VSCDI is included in this standard as the meta-data component of the specification. VSAF is part of MPEG-A and extends the AVC (Advanced Video Coding) file format [ISO06a], which in turn is an extension of the ISO Base Media File Format or MPEG-4 (Part 12) [ISO05b]. Overviews of MPEG-A and the ISO Base Media File Format are given in Section 3.11.4. The VSAF is due to become an international standard in 2009.

As part of a digital recording system, a VSAF instance or *Fragment*, is a short-term video segment able to be linked with other Fragments in a ring buffer architecture. Multiple tracks can be contained within a single Fragment with each containing a video output stream from a camera. Useful for forensic analysis, a VSAF fragment allows the export and transfer of multiple video and meta-data tracks spanning a particular time period. The video coding scheme chosen is AVC with Baseline profile. The meta-data comprises custom binary types, for identifying the Fragment and camera recording equipment, and the VSCDI. The VSCDI meta-data is non-essential and defined Annex B of the VSAF standard [BR08]. Legacy and non-compliant bitstreams are indirectly supported and can be referenced, even video tapes. A Fragment has a length of a particular UTC time period, therefore all video tracks have the same length in UTC time. The track UTC time-points can encompass different time periods. The time of the

track can be stored to a high precision and binary time-stamps must accompany every encoded frame for each track.

The author has designed and developed the reference software¹ and conformance documentation for the VSAF standard [AS08]. The ISO Base Media File Format reference software requires extensions that support the VSAF binary meta-data and to pack and unpack AVC bitstreams. An API (Application Programming Interface) is required to produce XML documents conforming to the VSAF specification, *i.e.* the VSCDI. The AVC coding scheme reference software [SW08] is used to produce compliant bitstreams.

The main idea is for MPEG-A standards to use MPEG technology. This means the VSAF had a fairly limited choice, but sensibly chose AVC for the video and MPEG-7 for extended meta-data. MPEG-21 was seen as too advanced for the conservative CCTV industry. AVC is not perfect because while surveillance requires a simple coding scheme, MPEG coding schemes are designed for broadcasters. This means the emphasis is on encoder complexity and not decoder. For surveillance the requirements can be considered as the inverse, *i.e.* requiring encoder simplicity and decoder complexity. The coding scheme called DVC (Distributed Video Coding) fulfills this role [GARRM05]. VSAF uses a simplified version of AVC which doesn't contain complex error resilience. This widely used flavour of the Baseline Profile has only just been officially recognised by MPEG, and has delayed the progress of VSAF while the appropriate steps are taken to define a new AVC profile. The MPEG-7 profile (VSCDI) has unresolved issues where in order to provide a suitably simple surveillance description, it is necessary to use certain descriptors with a new semantic. The approach has been criticised because it is harmful to inter-operability.

¹The author would like to thank Houari Sabirin and Thomas Rathgen for their contributions, for the generation of the conformance files and AVC bitstream packing and unpacking software, respectively.

Technical meta-data

As discussed in Section 5.2.1 the requirements for the VSAF were drawn out during MPEG meetings. With reference to Section 5.2.2, an evaluation of the fulfillment of the quality requirements is assessed in this Section. Siemens Plc. required binary meta-data elements to define the camera identifiers and time-stamps. The advantages to this approach are simple easily processed meta-data, without the need for XML parsers. This fits the requirements for low-powered devices and ‘computational complexity’. The ‘timed meta-data’ tracks [ISO05c] containing the binary time-stamps have the ability to contain extra information. There is a space for user-data here and it is even possible for a developer to record track by track XML meta-data. The design of the ISO Base Media File Format requires ‘unknown’ boxes to be ignored, this forwards compatible design allows proprietary meta-data to be included. This satisfies the ‘flexibility’ requirement. The ‘information compatibility’ requirement is not accommodated with the binary meta-data because it is proprietary to the VSAF specification. The use of unique identifiers allows a catalogue to be built by parsing the file structure and extracting the appropriate binary meta-data boxes. The structure of the catalogue is not defined and not standardised since the meta-data is intended to be an archival format, to be incorporated within a system where standardisation is not necessarily required.

Observation meta-data

The VSCDI is an optional component allowing a low-power processor to skip its decoding. The VSCDI provides a conceptually simple method of describing a surveillance scene at a frame-by-frame level. The method for semantic annotation is via free and structured text annotation. Structured annotation

adequately provides a facility for semantic annotation with *why*, *whatAction*, *etc.* fields, while Classification Schemes improve information compatibility by defining a simple structure and referencing mechanism. At the same time a flexible description is possible, through an attribute of each element, *e.g.* `who href='my_CS:who:joe_doe'`. Computational complexity is maintained within reasonable limits by removing the ability of the MPEG-7 elements to be recursive and limiting the number of the elements that it is possible to create within the description, *e.g.* one cannot define an unlimited number of *who* elements in a structured annotation. This kind of limitation makes the development process easier. The developer does not have to iterate through many lists, examining each for possible avenues of meaning.

6.2.2 CARETAKER evaluation

The CARETAKER project, described in Section 3.9, uses meta-data to handle the output from its automated processes. The XML schema used to describe the data model is shared throughout the collaborators. There is consensus in the correct use of the schema but its design is customised for the collaborators' needs. The main top-level types are *Event*, *Object* and *Tracked Target*. An *Event* comprises an identifier and an enumerated type taken from the CARETAKER ontology. The ontology is not referenced but manually entered, requiring new editions of the schema to be distributed if there are changes. The main characteristics of the schema are listed:

- The *Object* has a bounding box location and an identifier.
- The *Tracked Target* has both two- and three-dimensional bounding boxes.
- The name of the algorithm outputting the meta-data is recorded.

- The identifiers use integers.
- All the elements are defined as root elements (top-level) in the global namespace.

The structure relies upon expressed consensus between the collaborators. The agreed hierarchy is *caretaker* \rightarrow *algorithm* or *caretaker* \rightarrow *frame*, *frame* \rightarrow *event* or *frame* \rightarrow *trackedtarget* or *frame* \rightarrow *object*. The meta-data is ‘frame-oriented’ meaning a Frame element is required first, then a new Event, Tracked Target or Object is defined within the frame. It is the developers’ responsibility to use unique identifiers to preserve the correct identities. There is no identifier for the root element, *e.g.* the CARETAKER element. This means that documents received from multiple sources using the same algorithms cannot be differentiated from one another. Problems discovered with the schema are fed into project meetings and are subsequently addressed. This iterative development cycle suits the research and development *modus operandi* of the project and the relatively small team of engineers and scientists.

In comparison, the VSCDI was developed by MPEG and companies with an interest in a well engineered standard. To change the specification after publication requires a laborious process to make an addendum to the standard which is then published as an amendment. This has advantages since the specification is made clear and unambiguous from the outset, and suits the professional domain where the scope of the application is well known. VSCDI uses the established MPEG-7 standard with excellent label identification support, thereby avoiding the problems in the CARETAKER schema design, like preserving unique identifiers between objects and documents. Similarly to CARETAKER the VSCDI meta-data is frame-oriented, and the frames contain bounding box information but a three-dimensional bounding box is only possible using additional camera

calibration information with a corresponding conversion matrix. There are no explicit types for Event, Object and Tracked Target. An Event could be described within a frame description and a text annotation that references a Classification Scheme term called 'Event'. The Classification Scheme is stored at File Level (Section 5.3.7) and could define the CARETAKER ontology, negating the need to issue updated schemas. An Object and Tracked Target could be defined in the same way. A VSCDI track has a UUID, a global identifier and the video has an identifier unique within the file, as do most of the elements. VSCDI meta-data provides additional functionality through the use of Classification Schemes, two colour features and a grid descriptor and equipment-based elements. CARETAKER meta-data is distributed in RSS (Really Simple Syndication) packets [RSS]. The VSCDI will distribute its meta-data within a VSAF fragment.

Regarding the media, CARETAKER and VSAF use different video coding schemes and CARETAKER makes use of audio. A CARETAKER media file converted into a VSAF file would require the addition of the CARETAKER MPEG-4 media bitstream outside the VSAF file. Per-frame time-stamps are used for both formats but CARETAKER time-stamps are drawn on to a black rectangle at the bottom of every video frame. This is impossible to remove from the source media and may occlude objects of interest. The VSAF standard solves this problem by using the 'timed meta-data' track.

6.2.3 Grand Challenge evaluation

The Grand Challenge project is a U.K. Ministry of Defence (MOD) competition. Entrants must provide a solution that autonomously detects threats in an urban environment. Kingston University is a partner in the Silicon-Valley team. This team is developing robotic vehicles which navigate their way through the urban

environment using GPS (Global Positioning System) and video data to gather intelligence. Kingston are implementing the video analysis components which operate on data from cameras mounted on the vehicles. One partner in the team, University of Reading, is implementing a Command and Control console. This console will display the output from the Kingston processes and requires meta-data containing information about moving objects of interest, *i.e.* bounding boxes, and information about potential threats, *e.g.* an object's speed, number of people within a group, a sniper in a window or a *Technical* in the scene. A *Technical* is a term for a vehicle with a weapon attached over its roof.

The VSCDI was evaluated for use as the means for exchanging meta-data between the partners and the components. One alternative was to design a bespoke meta-data schema for this project. The choice was made to use VSCDI and for each meta-data document to describe a sequence of video over a 10 second time period. A threat description has a probabilistic value, *e.g.* object = sniper $\times 0.8$ and these values are stored within a graph where each relation represents the entire sequence. A Classification Scheme of the terms used is referenced by the graph relations or within structured annotations. This allows values requiring an integer amount to be accommodated within the schema, *e.g.* number of people, and speed. The structured annotation element can reference a Classification Scheme to inform the programmer how to interpret the value, *e.g.* `who href='my_CS:who:sniper'`.

In the event the VSCDI was not used and a bespoke meta-data solution was developed. Due to complications with the VSCDI SDK (Software development kit) and XML a more simplistic approach was favoured.

6.3 Considerations for an embedded system

The colour descriptor analysis undertaken in Chapter 4 and the VSCDI (Visual Surveillance Content Description Interface) architecture are further evaluated by implementing a colour meta-data engine for an embedded system. The implementation gave the device CBIR (content-based image retrieval) functionality in addition to its existing motion-based image retrieval capabilities. The work involved processing test images and generating colour meta-data, developing the matching process to measure the distance between a query colour and an image and providing a user-interface through which the image retrieval tasks are performed. The meta-data was MPEG-7 and compatible with the VSCDI. This Section describes the embedded device, the meta-data production procedure and evaluation.

6.3.1 Embedded digital video server

Overview Ltd. is a small company based in London who manufacture CCTV dome cameras [Prob] (Appendix B provides a description of a dome camera). The author spent a portion of the project's time working for this company, as part of the Imaging Faraday CASE (Collaborative Awards in Science and Engineering) award. The company has developed a digital video server called DVIEO®, shown in Figure 6.2, which is a small unit about the size of a VHS (Video Home System) cassette which can record video from a single camera. Its development was part of a long-term research strategy into a modular CCTV system. The device stores annotated video footage and allows remote control via a HTTP (Hypertext Transport Protocol) interface. All short-term video up to a period of a month is stored and only longer term footage based upon 'event' data, *i.e.* if there is movement in the scene. Camera control is only possible over the bespoke



Figure 6.2: The Overview Ltd. DVIEO[®] video server.

client application, shown in Figure 6.3. The DVIEO[®] server is a Linux computer running on a prototype mother-board with an ARM processor. It has a dedicated DSP (Digital Signal Processor) daughter-board to digitise the video and produces four watermarked JPEG (Joint Photographic Experts Group) images a second, in CIF (Common Intermediate Format) [ITU93] and QCIF (Quarter CIF) format.

6.3.2 Meta-data format

The meta-data defined by the system is proprietary and a mixture of hard-coded information and database tables. The main piece of meta-data is the time-stamp followed by the motion information. The images are stored on a hard-disk in a particular structure based upon the image's time-stamp. Each component of the time-stamp becomes a directory:

camera002/2005/05/03/08/00/00/ = camera/year/month/day/hour/min/.

This design is closely tied to the system architecture and means the DSP does not do any database indexing. Its limited processing power can be fully applied to frame capture, JPEG (Joint Photographic Experts Group) file [ISO94] compression and motion vector generation. Instead, an off-line process working on the CPU (central processing unit) iterates through the file system, indexing the images by reconstructing the time-stamp using the directory structure. An

Figure 6.3: The DVIEO[®] client.

alternative to this structure is to use a single directory for all the videos and label each with a unique identifier, however, the relation between the identifier and the image would have to be defined at the image's creation time.

The stored motion vector meta-data contains an image's motion information. The off-line process populates the database with the motion information. The movement in each image is calculated by the DSP. A motion detection routine uses a component of the JPG compression process to produce motion sensitive zones, as described in Section 3.5. The image is divided up into 4×4 *Macro-blocks* and these are further divided into 8×8 cells or *Micro-blocks*. The Micro-blocks provide further granularity in the motion sensitive regions. Viewed at Macro-block level the following matrix depicts a zone called *Window* at the top-right of the image:

$$\text{Window} = \begin{cases} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{cases}$$

There are 64 Micro-blocks cells per Macro-block cell. If the 64 bit number is zero there is no detection. Full detection means 2^{63} or a series of 1's in each of the 64 blocks. To clarify, if there is only movement in the central Micro-block, the number is 2^{31} or 1073741824. The 'motion-block' codes are used to produce a mask when carrying out queries based upon the motion in a particular region of interest. This is an example of a specialised use of meta-data with low flexibility and low computational overhead (Section 5.2.2). This description could potentially be accommodated with the VSCDI GridLayout DS (Description Scheme), which can support a corresponding number of cell masks and Visual Descriptors, *i.e.* 64×64 . Additionally, the Classification Scheme could define a 'window' term and a Structured Annotation could provide a suitable description. The system records zone information in the database.

6.3.3 Developing colour search capabilities

The MPEG-7 DominantColor DS is chosen for the task of providing the system with the ability to search on colour. This descriptor is suitable for a Query by Attribute (QBA) (Section 3.2) and details of this descriptor are provided in Section 3.4.2. To perform a CBIR application the user selects a colour from a colour palette provided by the user interface. This type of query is desirable in a surveillance application where an indication of the presence of a particular colour

within the video material could be useful to identify a person.

A compact and multi-platform search and retrieval application is implemented to encode and match DominantColor DS compliant meta-data. The implementation is written entirely by the author and did not borrow from the MPEG-7 reference software, thus avoiding any licensing issues and providing training. It is important to realise that MPEG-7 standardises the form of the meta-data, and not how the meta-data should be produced, nor how matching is performed for retrieval applications. Hence, the K-Means algorithm is chosen to cluster the colours because of its simplicity, as opposed to the recommendation of the ‘Generalised Lloyd Algorithm’ [MSS02]. The only drawback is the fixed size of the clusters while other methods allow variable cluster sizes.

The motion meta-data is used to recreate a binary motion mask so the clustering is only performed on the moving area. The proposed match-measure for the DominantColor DS is the mode without variance and spatial coherence, as given in Section 3.4.2. Complications are encountered when deciding whether two colour clusters are similar. The guidebooks suggest a threshold of about 16 as the minimum distance. This default threshold is too severe a discriminator due to this data set’s small size and correspondingly limited number of colours. This problem is heightened by a QBA application which employs a single query colour. Compare this to a Query by Example (Section 3.2) application where a query image is likely to contain many colours. The solution to this problem is to iteratively increase the threshold until n images are retrieved. For the colour blue, for example, four iterations are required, *i.e.* 4×16 , before enough images are retrieved. Another problem encountered is the production of negative green values generated when converting from l, u, v to r, g, b . In this case the green values are very small and so the solution is to use their *absolute* value. The Dominant-Color DS converts each colour channel to a range of between 0 – 31 (five bits),

Search By Colour Using Meta-data

SEARCH BY COLOUR

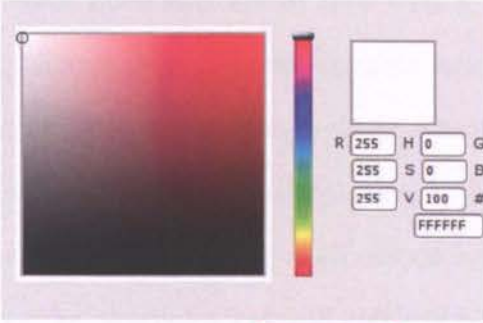
WHAT IT DOES...
 This option gives colour-based searching capabilities to the video server. All images are encoded with MPEG-7 compliant dominant colour meta-data, upon which the searches are performed.

Select Start Date and Time: 🕒

05-10-2005 14:33:30

Select End Date and Time: 🕒

05-10-2005 16:32:21



Preset Colours ▾

SEARCH

View results by: Images List

Min. rows

Copyright 2005 © Dvleo. All Rights Reserved.

Figure 6.4: The Query by Attribute colour search interface.

so these differences are removed after quantization.

The existing DVIEO[®] web user interface is extended to allow the image querying function. An existing h,s,v colour-picker [Pro07] is chosen to provide the user with a choice of any colour to search with, shown in Figure 6.4. The QBA design converts the colour into MPEG-7 format to perform the distance measuring. The colour feature encoder runs as an off-line process working through image data.

6.3.4 Evaluating the colour search tool

The performance of the implemented DominantColor DS algorithm within the system's framework was evaluated with an experiment based upon questionnaires. 15 participants took part to answer questions about how well a particular colour retrieves images containing this colour. The experiment has two parts: 1) To assess the level of agreement between participants about which label is most appropriately associated with each of 15 pre-selected colours. 2) To indicate how many elements in the set of images retrieved using each colour label was appropriate. The first criterion reflects the psychological and cultural aspects involved when humans select colours [Wik08]. The participants chosen from within the Computer Science Department at the University are a mix of different cultures, ranging from Northern Europe, Southern Europe, Asia, and South America. It is assumed that none of the participants is colour blind.

A test data set is created² of an outdoor scene where a camera is mounted on a pole and focused on a terraced house. Activity outside the front door is captured, *e.g.* a milkman delivery, children leaving and returning from school, as shown in Figure 6.5 (top). An example of the motion mask data is also illustrated (middle). The latter example provides the opportunity to retrieve individuals wearing brightly coloured school uniforms. There are approximately two hours of data. Some of the data does not contain people, but errors caused by the motion detection algorithm, as shown in Figure 6.5 (bottom).

The system is delivered to users via web server (Figure 6.4) from the demonstration platform. Table 6.1 shows the 15 pre-selected query colours used. The different colours are chosen so they spread evenly across the colour spectrum. A

²Thanks to Dr. Paul Jones and Mr. David Watkins for creating the data set.



Figure 6.5: The Overview data set showing moving people (top), corresponding motion masks (middle) and an example of errors in the motion detection (bottom).

semi-automatic technique is used to choose colours with an even spread. The resultant colours did not include grey, yellow and brown which are manually added since the data set contains images with these colours. Consideration is given to the effect different saturation and value levels may have on the hues but, such potential problems do not affect the system unduly since: a) the system always returns the closest colour whether it is similar or not; and b) DominantColor DS has qualities allowing for illumination insensitivity.

Due to the potential influence on the participant from one colour to the next, the experiment is structured so that a colour selection operation is followed by a retrieval operation. For the latter, the screen layout changes and the user must concentrate on the retrieved images. Even though the majority of neighbouring colours do have evenly increasing hues, it is considered that the operational effects of the experiment lessen any undue influence.

6.3.5 Results

The graphs shown in Figure 6.6 are derived from the 225 (15×15) queries. The top graph shows the results of the average retrieval success (one for success, nought for failure) for each of the colours over all participants. The most successful colour is red, which is the only one to score 100%. The colour orange scores $\approx 10\%$ and the agreement about the correctness of this label has nearly the same value. The latter result is understandable since there are no orange colours within the data set and the colour is an un-saturated orange. It is clear that the label 'orange' carries a meaning of a bright and intense colour to the participants. The Spearman correlation coefficient between the retrieval results and the correctness of the label gives a high correlation of 0.8492.

The results plotted from each participant (Figure 6.6) (top) show the majority

1	grey	240,5,74	182,182,192	B6B6C0
2	white	0,0,100	255,255,255	FFFFFF
3	black	0,0,0	0,0,0	000000
4	brown	30,50,50	128,96,64	806040
5	red	0,0,75	192,48,48	C03030
6	orange	30,75,75	192,120,48	C07830
7	yellow	56,95,93	236,221,12	ECDD0C
8	green	90,75,75	120,192,48	78C030
9	turquoise	180,60,47	48,120,120	307878
10	aquamarine	180,75,75	48,192,192	30C0C0
11	light blue	210,75,75	48,120,192	3078C0
12	dark blue	240,75,75	48,48,192	3030C0
13	purple	270,75,75	120,48,192	7830C0
14	violet	300,75,75	192,48,192	C030C0
15	pink	330,75,75	192,48,120	C03078

Table 6.1: The 15 query colours for the colour constancy experiment. Each row comprises a colour label and values in h,s,v , r,g,b and hexadecimal.

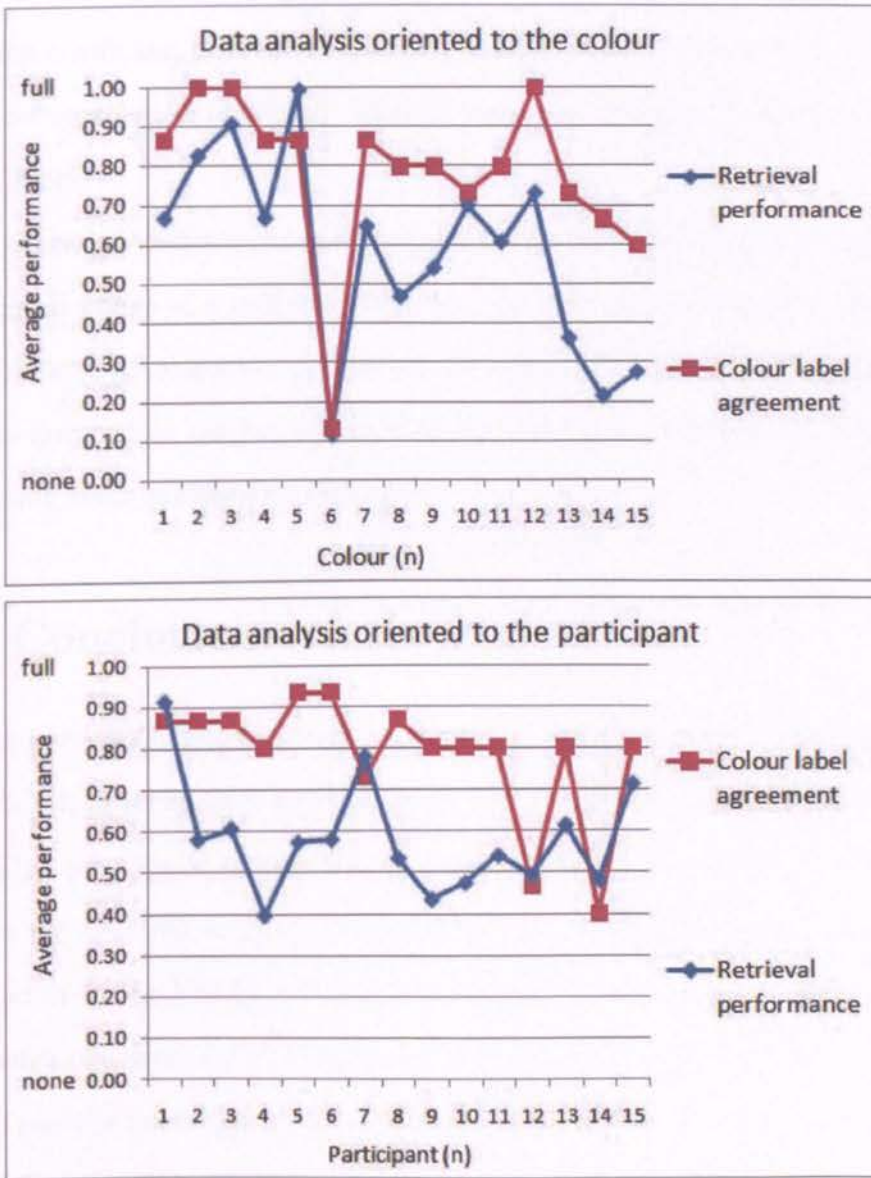


Figure 6.6: The results for the success rate of retrievals on each of the 15 colours are shown (top). The units on the x-axis relate to Table 6.1. The average results for the 15 participants are plotted (bottom). In addition, both graphs plot the agreement data.

of participants give a good rating to \approx half of the results. Two of the 15 had a tendency to give poorer results. This potentially could have been due to poorly calibrated monitors or other factors such as colour blindness. The Spearman correlation coefficient between the retrieval results and the correctness of the label gives a low correlation of 0.2857. This shows the participant is not influenced by a colour label.

The effects of shadows caused the participants difficulties. For example, when retrieving an image of a milkman who wears a white coat, if a grey colour query is used then psychologically it appears incorrect. The machine will detect the colour as grey, while the human participants are preconditioned to consider the colour using their previous experience.

6.4 Conclusion

The first part considers the object identity preservation technique proposed in Section 5.3.4, in terms of preserving pedestrian identity in a multiple sensor network. The question is asked about how it will work in practice; and how the language for *creating* a particular ontology, as opposed to using one, can be embedded in to the VSCDI (Visual Surveillance Content Description Interface). The benefits of a probabilistic framework which uses Classification Schemes seems clear: to provide flexibility in the definitions used for the transient Low Level Objects (LLO) and Global Unique Objects (GUO); and allow any relationship type between any kind of entity. How this system scales is unknown at present, but the maintenance of 1000 LLOs and GUOs, for example, may be problematic.

The second part investigates the VSCDI schema applied to three applications. First is a description of the VSAF (Video surveillance application format) and

the reference software implementation. Through its creation, the details of implementing an API (Application Programmer's Interface) and conformance files have exposed the design to scrutiny. There is a compromise between the extensibility and flexibility of the optional VSCDI meta-data and the mandatory binary meta-data. The idea is for the VSCDI component to provide extra optional features which are not mandatory. Components such as Colour descriptors and frame decomposition or Classification Schemes provide a complexity level which is not in demand, at present. The inclusion of this MPEG-7 meta-data can be used to provide a mechanism of supporting extra features. One example is using the XML to store information to correct a device's faulty or incorrectly calibrated internal clock.

The comparison between CARETAKER and VSCDI meta-data shows several elements of the CARETAKER approach are not compatible. Whether this compatibility is necessary as part of academic research and development projects and concept demonstrators is debatable. It is shown, however, that the entire CARETAKER ontology can be represented as a specific special case of VSCDI meta-data. The evaluation of VSCDI continues with the Grand Challenge project. For the fundamental descriptive elements, *e.g.* time, place, identity and decompositional structure, it is asserted that a suitable set of MPEG-7 Description Schemes has been identified. The Classification Schemes allow the description of variables which must be passed from one process to another. Identifiers for the variables are defined within the Classification Scheme because the tools can accommodate probability and integer valued variables. In the event the VSCDI was not adopted and this illustrates the need for a simple bespoke schemes easy which are easy to implement.

The colour search tool delivers tangible benefits to the DVIEO[®] product at application and engineering levels. The use of MPEG-7 descriptors provides

a compact and powerful description of colour meta-data. The impact on the processing power of the embedded device remains uncertain due to the evaluation conditions. The colour constant evaluation demonstrates the approach may have use in reducing the work-load of the operator when inspecting recorded data for the occurrence of a particular colour. The sensitivity of the search tool means false positives are more likely than false negatives which is more beneficial to the CCTV operator. One such insight is that a clearer definition of how shadows affect colour understanding will help deal with associated problems with this application. The examination of the techniques for meta-data design and integration at systems level on an embedded platform provides a useful insight into how feasible it is to use the proposed meta-data specification.

Chapter 7

Conclusions and future work

7.1 Summary of achievements

This body of work presents a new content description interface for video surveillance – The Video Surveillance Content Description Interface (VSCDI). It is shown a meta-data interface in the CCTV (closed-circuit television) domain shall act as a conduit for information exchange between digital computer processes. As standards are likely to become important in the surveillance community this work is timely and can act as a starting point for future development. The use of meta-data in CCTV is illustrated in Chapter 2 and reviewed in Chapter 3. For general purpose multimedia meta-data descriptions, MPEG-7 is the most advanced scheme available.

The investigation into MPEG-7 continues with an analysis of content-based image retrieval (CBIR) and its colour descriptors in Chapter 4. The efficacy of these descriptors is tested on bespoke and standard data. The GENERICK data set is created and contains pedestrians in an indoor scene. In addition a new data set of pedestrians is generated from the standard CAVIAR data set. A novel aspect of the CBIR work is to split the data into Top and Bottom categories

to improve the retrieval results of the experiments conducted in pedestrian re-identification. Generating posterior probabilities allowed the Information Gain metric to be brought to the domain which deals with Open World scenarios and this metric is suggested as a future standard for use in video surveillance. It is shown colour calibration is important for inter-camera processing and how benefit of applying this calibration only to the moving objects of interest is beneficial.

Chapter 5 describes the processes involved in profiling MPEG-7. The need for the descriptions of uncertainty is emphasised for surveillance and the means to describe them is defined. Customisable Classification Scheme descriptors are presented as a new tool for use in the surveillance domain. The MPEG-A: Part 10 – Video surveillance application format, which was developed in parallel to this work, helped define the details of the VSCDI specification. Qualitative and quantitative studies evaluate the VSCDI which include the development of a colour search tool for an embedded system.

7.2 Discussion

With application to surveillance, CCTV is a remote visualisation tool used for deterrence, visualisation and criminal apprehension. Today we have a problem similar to the one presented in Section 2.2, where a mass of video data is generated from millions of CCTV cameras which cannot easily be processed. Automatic processing can provide the means to improve the situation. Automated processing systems already exist but operate with limited application, *e.g.* ANPR (automatic numberplate recognition) systems. There is an increasing use of more sophisticated systems and these have a greater reliance on meta-data. Simple systems use simple meta-data and complex automated systems require meta-data with structure and a set of description tools. Designing the meta-data's structure

and processing tools is a complex process. A standard content description interface could benefit the interested parties in the CCTV industry, *e.g.* end-users, operators, system integrators and manufacturers. A standard interface will address issues related to quality, requirements by law, functional improvements to legacy systems and precise operation requirements, *e.g.* timing information. In particular, proprietary video analysis modules could work together by sharing a common format.

The concept of CBIR (content-based image retrieval) system is explained. Of the three different query paradigms reviewed a Query by Example (QBE) is important to this thesis. A variant of the QBE concept is introduced called a Query by Attribute (QBA). The Query by Symbol (QBS) uses an ontology to translate a semantic query to the data signals and is beyond the scope of this work. Meta-data is an essential component of the features used in a CBIR system and MPEG-7 provides a suite of meta-data feature descriptors each with feature generation and matching algorithms. The performance of the features can be evaluated using a binary yes or no, a ranking or a probability measure. Segmentation of moving foreground objects and camera calibration are necessary tools to provide quality meta-data features. Taxonomies and ontologies are used in behaviour recognition.

Six key surveillance applications are listed and their potential meta-data usage is analysed. Visualisation requires inter-operable annotation meta-data, while ANPR, Search and retrieval, pattern discovery and alarms all require varying levels of signal processing.

The main meta-data frameworks used in CCTV are in the commercial domain from IBM (Section 3.9) and Sony (Section 2.3), and it is questionable about how open these are. Meta-data standards have more use in the research community which use them for performance evaluation. It can be seen that there is some

commonality between the meta-data used, *e.g.* bounding boxes, identifiers and annotations. The descriptions become more general when high-level semantics are used. Outside CCTV (Section 3.10) the standardisation bodies often concentrate on standards built upon established descriptors, *e.g.* SMPTE (Society of Motion Picture and Television Engineers) time-codes and Dublin Core. STANAGs are detailed military standards and STANAG 4609 is a detailed standard for military surveillance. MPEG-7 (Section 3.11) has industrial adoption with TV-Anytime. Only certain parts of the system framework MPEG-21 have found adoption. Current file formats (Section 3.12) demonstrate sophisticated mechanisms to archive both media and meta-data. Mostly these are focused upon an application, *e.g.* SMPTE MXF, but MPEG-4: Part 12 show versatility as a relatively simple general purpose media format.

Signal based meta-data is not in widespread because the algorithms are not presently reliable enough. Despite this, MPEG-7 provides advanced content-based descriptors. The evaluations of MPEG-7 colour descriptors in Chapter 4 demonstrates that these features have efficacy for video surveillance. While the ANMRR (Average Normalised Modified Retrieval Rate) is a useful ranking metric, its main use is in performance evaluation. The Information Gain metric is useful in CCTV Open World conditions and allows descriptors to be combined together to improve retrieval performance.

A content description interface for visual surveillance meta-data is proposed called the VSCDI (Video Surveillance Content Description Interface) (Section 5). Its aim is to provide a simple interface which defines the essential meta-data elements required by a typical surveillance application. Generated by profiling MPEG-7, a dramatic reduction is made in the size of the MPEG-7 schema. The elements are chosen based upon a list of requirements generated by industrial and academic researchers. The scheme includes two colour descriptors whose choice

is based upon the outcomes of the evaluations in this work. Tools to describe uncertainty and define Classification Schemes are included. Uncertainty values are derived from the outputs of probability metrics, such as, CBIR style algorithms. A framework is proposed which defines relationships between objects, connecting low-level blobs and objects with certain identities. Classification Schemes provide a powerful way to define taxa in a structured way. These Term definitions can be used throughout and applied to the descriptors in a flexible manner. The VSCDI has an architecture of Technical and Observation components split at File Level and Track Level allowing a Collection and Item conceptual hierarchy. The Technical meta-data is about the equipment and the Observation is about the video semantics.

The framework for the relationships between certain and uncertain objects is evaluated qualitatively (Section 6.1). The use of the VSCDI in various scenarios is evaluated (Section 6.2). The choice to build a bespoke meta-data solution is a widely taken approach but not without issue. The VSCDI is designed to be as simple as possible, while being applicable to analytical video processing. A component of the VSCDI is implemented within an embedded system to provide it with the capability to search images based upon their colour. The implementation demonstrates the meta-data used by such a system and how the DominantColor DS (Description Scheme) has applicability in a constrained embedded environment. The importance of an operator's psychological response to colours is also noted.

Ethical considerations

The reasons for the use of surveillance in society are given in Section 2.2. There is a potential for the abuse of the tools installed for this purpose and it is correct to

have safeguards to protect the public. The ethical issues of using cameras to monitor people in public spaces are especially significant if the data is recorded. As a consequence the European member states must implement laws governing the management of sensitive data. The U.K. version of the EU directive 95/46/EC is the Data Protection Act (DPA) of 1998 [DPA98]. One product of the DPA is the establishment of the Information Commissioner who has published a Code of Practice for CCTV [UKG08b]. The code helps the end-users of public CCTV systems stay within the law. The potential for standards to increase compatibility can increase information flow. The implications of a freer flow of CCTV information on society will require consideration. If the information travels beyond national boundaries, international data protection laws will be required, *e.g.* the Prüm Treaty of European Parliament [UKG07a].

Governed by the Kingston University Digital Imaging Research Centre's Code of Practice, the GENERICK data set (Section 4.2) created for this work was ethically cleared by the university for use with research purposes including publication on the Internet. A letter was drawn up inviting students to participate in the data gathering process and each of the undergraduate participants in the data set provided written consent that the data could be used in this way.

7.3 Future work

7.3.1 Content features

Further improvements could be made in the MPEG-7 evaluations by combining features from more of the MPEG-7 library, *e.g.* texture, shape, *etc.* Improved segmentation of objects into subcomponents and generating feature vectors for each would improve the overall signature and allow searches on subcomponents.

The identification of subcomponents would allow semantic searches upon components comprising a subject's appearance, *e.g.* shirt and trousers.

The poor performance of the colour descriptors over multiple cameras with un-normalised colour gamuts demonstrated by the relatively good performance of the simple Mean descriptor, clearly indicates a useful direction for future work, *i.e.* the specification of an improved preprocessing method through which the multi-camera retrieval rate can be further enhanced.

Principal component analysis (PCA) could be used to evaluate the correlation between the data set and the features. This will provide more information on the strengths of each feature. Additionally, comparisons between the texture feature used and MPEG-7 texture features will provide further useful analysis. In the evaluation the data was partitioned into training and test sets using a 'soft partition'. Increasing the size of the data set and separating the data into test and training sets and into categories of different sensors would provide interesting and realistic cross-sensor retrieval performance data.

The DominantColor DS (Description Scheme) evaluation in Section 6.3.5 shows that by selecting the exact colour from an image within a set of retrieved images, subsequent searches based upon this new colour refined the search results. More formal relevance feedback approaches could be investigated in order to improve the retrieval performance.

Computational efficiency tests would also be beneficial. The effects made by the use of different reference whites used in the colour conversion processes, as described in Section 3.4.2, could be quantitatively analysed.

7.3.2 Identity preservation

The identity preservation framework would benefit from a quantitative evaluation. This would include an assessment about how well the system performs with many GUOs (Global Unique Objects) and LLOs (Low Level Objects). The performance could be compared against Closed and Open World scenarios. The evaluation could involve a multi-sensor system which generates the data and meta-data. Such a system requires considerable overhead in terms of implementation time and cost, especially if it involves non-visual sensors, such as swipe cards.

7.3.3 Schema extensions

For future work several extensions to the domain of standardisation could be considered. Audio and its associated meta-data are useful in some surveillance contexts. Content protection technology potentially has a role in surveillance. MPEG is considering developing an advanced version of the MPEG-A: Part 10 – Video surveillance application format with these features.

It will be helpful to identify how the relation with other control elements of a surveillance system, *e.g.* PTZ (pan-tilt-zoom) and access control can be standardised. There are low-level protocols for communicating with the camera level, *e.g.* Sony's commonly used VISCA (Video System Control Architecture)¹ protocol allows PTZ control over a serial communications link. More suitable to this application are network control protocols for controlling the camera over the network, *e.g.* for video conferences the ITU-T recommendation H.281 [ITU94]. As computing power continues to increase there is potential for H.281 XML format.

Further both qualitative and quantitative evaluations of the VSCDI (Visual

¹The VISCA protocol is found printed on Sony products specification sheets.

Surveillance Content Description Interface) meta-data would be beneficial in assessing the performance characteristics of the specification. The evaluation would also benefit from a larger scale translation of alternative meta-data descriptions to the VSCDI.

Final words

It is inevitable for meta-data to become an integral part of future CCTV systems and inevitable that certain media and meta-data formats will become standard. The VSAF (Video surveillance application format) is the first such standard and the VSCDI (Visual Surveillance Content Description Interface), its meta-data component, provides a way of describing surveillance meta-data.

Appendix A

Video surveillance application format – MPEG-7 profile

A.1 Table of contents and Annex B (Schema omitted)

These pages taken from Committee Draft Document ISO/IEC FCD 23000-10 Video surveillance application format are reproduced with the permission of the International Organization for Standardization, ISO: www.iso.org.

Contents

Page

Foreword	
Introduction	
1 Scope	
2 Normative references	
3 Overview of MPEG Standards Used	
3.1 MPEG-4 Advanced Video Coding	
3.2 ISO Base Media File Format	
3.3 MPEG-7 Multimedia Description Scheme	
3.4 MPEG-7 Visual	
3.5 AVC File Format	
4 Using the Video Surveillance AF	
4.1 General	
4.2 File Structure	
4.3 File Contents	
4.4 Track Structure	
4.5 Derivation from the ISO Base Media File Format	
5 Video Coding Definition	
5.1 Introduction	
5.2 AVC Profile and Level	
6 Metadata	
6.1 Introduction	
6.2 File Level Metadata	
6.3 Track Level Metadata	
6.4 Timed Metadata	
Annex A (informative) Use cases of Video Surveillance AF	
Annex B (normative) Metadata Specification	185
B.1 Introduction	185
B.2 Metadata Definition	185

Annex B (normative)

Metadata Specification

B.1 Introduction

This annex contains the metadata specification for the additional Meta Boxes, on file and track level, which may be included in a Video Surveillance AF fragment. (This is separate to the specification for the required Meta Boxes at file and Track level, which are to be found in subclauses 6.2.1 and 6.3.1 respectively).

B.2 Metadata Definition

The following table summarizes the elements of MPEG-7 schema that are conformance to the requirements of Video Surveillance AF. The interpretation is as follows:

- “element/attribute/attributeGroup” – this MPEG-7 metadata shall be instantiated in the metadata of a Video Surveillance compliant file
- `xsi:type="[TypeName]”` – the element shall have this attribute, when instantiated in the item-level metadata. Therefore, it shall only be instantiated with type [TypeName])
- `minOccurs="n”` – at least *n* occurrences of the element shall be instantiated in the item-level metadata
- `maxOccurs="m”` – no more than *m* occurrences of the element shall be instantiated in the item-level metadata
- Elements are referenced using MPEG-7 id attributes

Textual description of metadata:

1. General text annotation

- a. Textual annotation should be added using either free text or structured annotation. All elements within structured annotation are limited to zero or one.

2. For file level metadata:

a. ID

- i. The UUID of the VSAF file shall be repeated in the `PublicIdentifier` of the `DescriptionMetadata` element. There shall be one of these descriptors.
- ii. It is assumed the UUID of the camera already contains the information of the cluster the camera belongs to.

b. Time

- i. The VSAF start-time should be repeated at file-level (`CreationTime` of the `DescriptionMetadata` element). This indicates the creation time of the fragment.

- c. Textual annotations should be added using `Comment` of the `DescriptionMetadata` element. Zero or one of the types described in 1.a. shall be used.

- d. Classification schemes and `Terms` can be defined and used as described in ISO/IEC 15938-5:2003 section 7.4. The cardinality of the `Definition` element of the `TermDefinitionBaseType` shall be zero or one. The cardinality of the `Name` element of the `TermDefinitionBaseType` shall be zero or

one. There shall be zero of the preferred attribute of the Name element of the TermDefinitionBaseType.

e. Maintaining object references

- i. Object references are grouped using the Graph DS and referenced using Relation DS elements. The objects can be any DS, as used within the XML document, e.g. a still region of video at track level.
- ii. The attributes within a Relation are restricted to contain unary values. E.g. source, target and type can only contain a single reference to a Classification Scheme Term, id reference, etc. i.e. the values of the termReferenceType.

1. General text annotation

The TextAnnotationType shall have zero of the confidence and relevance attributes.

The term TermUseType shall be restricted to contain:

- o zero of the TermUseType elements (no recursion)
- o zero or one of the Name element
- o zero or one of the Definition element

Description	Elements of MPEG-7 Schema	Constraints
General text annotation		
1.a Free Text	FreeTextAnnotation	minOccurs = "0"
1.b Structured Annotation	StructuredAnnotation/Who	minOccurs = "0"
	StructuredAnnotation/WhatObject	minOccurs = "0"
	StructuredAnnotation/WhatAction	minOccurs = "0"
	StructuredAnnotation/When	minOccurs = "0"
	StructuredAnnotation/Where	minOccurs = "0"
	StructuredAnnotation/Why	minOccurs = "0"
	StructuredAnnotation/How	minOccurs = "0"

2. File-level table

Description	Elements of MPEG-7 Schema	Constraints
Information associated with captured data		
2.a ID	Mpeg7/DescriptionMetadata/PublicIdentifier	
2.b Time information	Mpeg7/DescriptionMetadata/CreationTime	
Metadata for content		
2.c Annotations that apply to file-level		
i. In a free text	Mpeg7/DescriptionMetadata/Comment/FreeTextAnnotation	Choice

format		minOccurs = "0" maxOccurs = "unbounded"
ii. In a semantically structured format	Mpeg7/DescriptionMetadata/Comment/StructuredAnnotation	
2.d Classification schemes	Mpeg7/Description[@xsi:type="ClassificationSchemeDescriptionType"]/ClassificationScheme	minOccurs = "0" maxOccurs = "unbounded"
2.e The identity of objects observed in one or more video sources		
i. Object reference descriptions	Mpeg7/Description[@xsi:type="ContentEntityType"]/Relationships	minOccurs = "0" maxOccurs = "unbounded"
	Mpeg7/Description[@xsi:type="ContentEntityType"]/Relationships/Relation	minOccurs = "0" maxOccurs = "unbounded"
	Mpeg7/Description[@xsi:type="ContentEntityType"]/Relationships/Relation[source,target,type]	minOccurs = "0"
ii. Restricting the cardinality of the Relation attributes	Mpeg7/Description[@xsi:type="ContentEntityType"]/Relationships/Relation[source="xxx"],target,type]	(xxx) maxOccurs = "1"
	Mpeg7/Description[@xsi:type="ContentEntityType"]/Relationships/Relation[target="xxx"]	
	Mpeg7/Description[@xsi:type="ContentEntityType"]/Relationships/Relation[type="xxx"]	

3. For each track:

- a. Meta-data for each track are described using the Video Segment DS. *E.g.* VideoType. Only one of these types shall be used.
- b. ID.
 - i. The camera id shall be repeated from the 'cami' box in PublicIdentifier of the DescriptionMetadata element. There shall be one of these descriptors.
- c. Equipment
 - i. The camera / cluster settings should be given (Instrument of the DescriptionMetadata element). If present, zero or one of these types shall be used.
 - ii. Additional information regarding the cluster to which the camera belongs to should be given by EntityIdentifier and its VideoDomain element. If present, one of the EntityIdentifier types shall be used and zero or more of the VideoDomain types should be used. The VideoDomain elements reference entries from a Classification Scheme (ClassificationScheme).
 - iii. The camera stream should be identified with StreamID. Zero or more of these types shall be used. It is necessary to include the element InstanceIdentifier, although this can be kept empty.

- iv. The camera geographic position should be given using `CreationLocation`. Zero or one of these types shall be used.
- v. Camera calibration should be provided with the `Spatial2DCoordinateSystemType`. A description of more than one of these descriptors allows a calibration function for each preset for PTZ cameras to be calibrated. Zero or one of these types shall be used
- vi. If the media is outside of the VSAF fragment and referenced using the Data Reference Box (dref) the `MediaURI` shall contain the same reference.
If no Data Reference Box (dref) is present the `MediaURI` should contain a valid reference to a media instance. It is necessary to include the element `InstanceIdentifier`, although this can be kept empty.

d. Time

- i. Video offset has no specific element, so the Description Metadata DS (`DescriptionMetadata`), `Instrument` and its `Tool Setting` elements should be used. The setting name is "offset". If present, one of these types shall have the format and precision as given in Section 6.
- ii. The time of the video shall be given with a media time element (`MediaTime`). Within this element one time point shall be given. Duration information can be also given here. To specify the duration of a video decomposition `MediaDuration` should be used. This is a representation of the duration of the track as given in Section 6.
- iii. To isolate where the `StillRegion` exists in the video, the `MediaTimePoint` shall be used.

e. Decomposition

- i. Groups of frames should be defined within the video using the `TemporalDecomposition`. If present, one of these types shall be used
- ii. Single frames should be decomposed using the `StillRegion` DS. If a frame (`StillRegion` DS) is decomposed its time position shall be specified.
- iii. To isolate a region within a frame, a choice shall be made between a `Box` or `Polygon`. Zero or one of these can be described per `StillRegion`. If more regions are required for a frame, then another `StillRegion` can be instantiated, referencing the same media time point (`mediaTimePoint`).

f. Visual Descriptions

- i. Colour should be described in the `StillRegion` DS by the `VisualDescriptor` DS or the `GridLayout` DS. The `VisualDescriptor` shall include one colour descriptor. The `GridLayout` can specify an arbitrary number of cells, each should contain one colour descriptor.
- ii. `DominantColor` and `ScalableColor` shall be the only descriptors present from the `VisualDescriptor` DS and `GridLayout` DS.

g. Semantic descriptions

- i. A camera track should define semantic descriptions using `TextAnnotation` – see 1."General text annotation". This is possible with `FreeTextAnnotation` at `DescriptionMetadata`, `Video` and `StillRegion` levels.
- ii. A camera track should define semantic descriptions using `TextAnnotation` – see 1."General text annotation". This is possible with a structured annotation using the `StructuredAnnotation` at `DescriptionMetadata`, `Video` and `StillRegion` levels.

iii. In order to provide detailed meaning to semantic descriptions, *Terms* should be referenced from Classification schemes (*ClassificationScheme*) – see 2.d.

h. Maintaining object references

i. Object references are defined as described in 2.e.

Track-level table

Description	Elements of MPEG-7 Schema	Constraints
Information associated with captured data		
3.a Video	Description[@xsi:type="ContentEntityType"]/MultimediaContent[@xsi:type="VideoType"].	
3.b Identification tag for the camera	Mpeg7/DescriptionMetadata/PublicIdentifier	
3.c Description of equipment used and equipment settings		
i. Camera settings (Aperture, shutter-speed values, peak-to-peak voltage, etc.)	Mpeg7/DescriptionMetadata/Instrument	minOccurs = "0"
	Mpeg7/DescriptionMetadata/Instrument/Settings	minOccurs = "0" maxOccurs = "unbounded"
ii. Additional information regarding the cluster	Mpeg7/Description[@xsi:type="ContentEntityType"]/MultimediaContent[@xsi:type="VideoType"]/MediaInformation/MediaIdentification/EntityIdentifier	
	Mpeg7/Description[@xsi:type="ContentEntityType"]/MultimediaContent[@xsi:type="VideoType"]/MediaInformation/MediaIdentification/EntityIdentifier/VideoDomain	minOccurs = "0" maxOccurs = "unbounded"
iii. Identification tags for each of the multiple streams from a single camera.	Mpeg7/Description[@xsi:type="ContentEntityType"]/MultimediaContent[@xsi:type="VideoType"]/MediaInformation/MediaProfile/MediaInstance/MediaLocator/StreamID	minOccurs = "0"
	Mpeg7/Description[@xsi:type="ContentEntityType"]/MultimediaContent[@xsi:type="VideoType"]/MediaInformation/MediaProfile/MediaInstance/InstanceIdentifier	
iv. camera geographic position	Mpeg7/DescriptionMetadata/CreationLocation/GeographicPosition	minOccurs = "0"
v. Camera calibration	Mpeg7/Description[@xsi:type="ContentEntityType"]/MultimediaContent[@xsi:type="VideoType"]/Header[@xsi:Spatial2DCoordinatesType]	minOccurs = "0"
vi. Media URI	Mpeg7/Description[@xsi:type="ContentEntityType"]/MultimediaContent[@xsi:type="VideoType"]/MediaInformation/MediaProfile/MediaInstance/MediaLocator/MediaURI	minOccurs = "0"

	Mpeg7/Description[@xsi:type="ContentEntityType"]/MultimediaContent[@xsi:type="VideoType"]/MediaInformation/MediaProfile/MediaInstance/InstanceIdentifier	
3.d Timing Information		
i. Video time offset	Mpeg7/DescriptionMetadata/Instrument/Tool/Setting[Name, Value]	minOccurs = "0"
ii. Time of the video	Mpeg7/Description[@xsi:type="ContentEntityType"]/MultimediaContent[@xsi:type="VideoType"]/Video/MediaTime/MediaTimePoint	
	Mpeg7/Description[@xsi:type="ContentEntityType"]/MultimediaContent[@xsi:type="VideoType"]/Video/MediaTime/MediaDuration	minOccurs = "0"
iii. Time of a frame	Mpeg7/Description[@xsi:type="ContentEntityType"]/MultimediaContent[@xsi:type="VideoType"]/Video/TemporalDecomposition/StillRegion/MediaTimePoint	
3.e Decomposition		

i. Groups of frames	Mpeg7/Description[@xsi:type="ContentEntityType"]/MultimediaContent[@xsi:type="VideoType"]/Video/TemporalDecomposition	minOccurs = "0"
ii. Single frames	Mpeg7/Description[@xsi:type="ContentEntityType"]/MultimediaContent[@xsi:type="VideoType"]/Video/TemporalDecomposition/StillRegion	minOccurs = "0" maxOccurs = "unbounded"
iii. Isolate a region	Mpeg7/Description[@xsi:type="ContentEntityType"]/MultimediaContent[@xsi:type="VideoType"]/Video/TemporalDecomposition/StillRegion/SpatialLocator/Box	choice minOccurs = "0"
	Mpeg7/Description[@xsi:type="ContentEntityType"]/MultimediaContent[@xsi:type="VideoType"]/Video/TemporalDecomposition/StillRegion/SpatialLocator/Polygon	choice minOccurs = "0"
3.f. Visual Descriptors		
i. Colour can be described in a frame using the StillRegion or the GridLayout	Mpeg7/Description[@xsi:type="ContentEntityType"]/MultimediaContent[@xsi:type="VideoType"]/Video/TemporalDecomposition/StillRegion/VisualDescriptor	Choice
	Mpeg7/Description[@xsi:type="ContentEntityType"]/MultimediaContent[@xsi:type="VideoType"]/Video/TemporalDecomposition/StillRegion/GridLayoutDescriptors/Descriptor	minOccurs = "0"
ii. ScalableColor and	Mpeg7/Description[@xsi:type="ContentEntityType"]/MultimediaContent[@xsi:type="VideoType"]/Video/TemporalDecomposition	

<p>DominantColor</p>	<p>n/StillRegion/VisualDescriptor[@xsi:type="ScalableColorType"]</p> <p>Mpeg7/Description[@xsi:type="ContentEntityType"]/MultimediaContent[@xsi:type="VideoType"]/Video/TemporalDecomposition/StillRegion/GridLayoutDescriptors/Descriptor[@xsi:type="ScalableColorType"]</p>	
	<p>Mpeg7/Description[@xsi:type="ContentEntityType"]/MultimediaContent[@xsi:type="VideoType"]/Video/TemporalDecomposition/StillRegion/VisualDescriptor[@xsi:type="DominantColorType"]</p> <p>Mpeg7/Description[@xsi:type="ContentEntityType"]/MultimediaContent[@xsi:type="VideoType"]/Video/TemporalDecomposition/StillRegion/GridLayoutDescriptors/Descriptor[@xsi:type="DominantColorType"]</p>	
<p>3.g. Semantic descriptions</p>		
<p>i. In a free text format</p>	<p>Mpeg7/DescriptionMetadata/Comment/FreeTextAnnotation</p>	<p>Choice minOccurs = "0"</p>
	<p>Mpeg7/Description[@xsi:type="ContentEntityType"]/MultimediaContent[@xsi:type="VideoType"]/TextAnnotation/FreeTextAnnotation</p>	<p>maxOccurs = "unbounded"</p>
	<p>Mpeg7/Description[@xsi:type="ContentEntityType"]/MultimediaContent[@xsi:type="VideoType"]/Video/TemporalDecomposition/StillRegion/TextAnnotation/FreeTextAnnotation</p>	
<p>ii. In a semantically structured format</p>	<p>Mpeg7/Description[@xsi:type="ContentEntityType"]/MultimediaContent[@xsi:type="VideoType"]/TextAnnotation/StructuredAnnotation</p>	<p>Choice minOccurs = "0"</p>
	<p>Mpeg7/DescriptionMetadata/Comment/StructuredAnnotation</p>	<p>maxOccurs = "unbounded"</p>
	<p>Mpeg7/Description[@xsi:type="ContentEntityType"]/MultimediaContent[@xsi:type="VideoType"]/Video/TemporalDecomposition/StillRegion/TextAnnotation/StructuredAnnotation</p>	
<p>iii. Referencing Classification Schemes</p>	<p>Mpeg7/Description[@xsi:type="ContentEntityType"]/MultimediaContent[@xsi:type="VideoType"]/Video/TemporalDecomposition/StillRegion/TextAnnotation/StructuredAnnotation/Who/Term[termID]</p> <p>Mpeg7/Description[@xsi:type="ContentEntityType"]/Relationships/Relation[type,target,source]</p>	<p>minOccurs = "0"</p>
<p>3.h. Maintaining object references</p>		
	<p>Mpeg7/Description[@xsi:type="ContentEntityType"]/Relationships</p>	<p>minOccurs = "0"</p> <p>maxOccurs =</p>

		"unbounded"
	Mpeg7/Description[@xsi:type="ContentEntityType"]/Relationships/Relation	minOccurs = "0" maxOccurs = "unbounded"

An example of a File Level document

```
<?xml version="1.0" encoding="UTF-8" standalone="yes" ?>
<Mpeg7 xmlns="urn:mpeg:mpeg7:schema:2001"
xmlns:mpeg7="urn:mpeg:mpeg7:schema:2001"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="urn:mpeg:mpeg7:schema:2001 Mpeg7-2001_vsaf-07-2008.xsd
">
  <DescriptionMetadata>
    <Comment>
      <FreeTextAnnotation>The VSAF reference software created this
XML</FreeTextAnnotation>
      <StructuredAnnotation>
        <Who href="urn:mpeg:mpega:vsaf:author:james_annesley"/>
        <WhatObject>
          <Name>The VSAF reference software</Name>
        </WhatObject>
      </StructuredAnnotation>
    </Comment>
    <PublicIdentifier>4E9C95FA8C47428</PublicIdentifier>
    <CreationLocation>
      <GeographicPosition>
        <Point latitude="0.100000" longitude="0.100000" />
      </GeographicPosition>
    </CreationLocation>
    <CreationTime>2008-01-01T01:01:01:0F30</CreationTime>
    <Instrument>
      <Tool>
        <Name>Automatically generated by VSAF Meta-data API</Name>
      </Tool>
    </Instrument>
  </DescriptionMetadata>
  <Description xsi:type="ClassificationSchemeDescriptionType">
    <!-- The relationship between Joe Doe (on patrol) and Fred Bloggs
    (working on the computer) and -->
    <!-- a moving object detected by the surveillance camera 8A4EB5D3-
    B2D8-4b99-A430-11032F0AAAF4 -->
    <!-- These values will be based upon the output of all cameras and
    time etc. -->
    <Relationships id="rship1">
      <Relation id="r1"
        source="urn:mpeg:mpega:vsaf:example2:vs:infrastructure:camera:ptz:camera1"
        type="urn:mpeg:mpega:vsaf:example2:process:object:unknown1"
        target="urn:mpeg:mpega:vsaf:example1:acme:people:security:guard:joe_doe"
        strength="0.8"/>
      <Relation id="r2"
        source="urn:mpeg:mpega:vsaf:example2:vs:infrastructure:camera:ptz:camera1"
        type="urn:mpeg:mpega:vsaf:example2:process:object:unknown1"
        target="urn:mpeg:mpega:vsaf:example1:acme:people:staff:fred_bloggs"
        strength="0.2"/>
    </Relationships>
  </Description>
</Mpeg7>
```



```

</Relationships>
<ClassificationScheme uri="urn:mpeg:mpega:vsaf:example1" id="cs1">
  <Term id="t1" termID="acme">
    <Name>ACME</Name>
    <Definition>The ACME company classification scheme</Definition>
  <Term id="t2" termID="people">
    <Name>People</Name>
    <Definition>People recorded by the ACME
surveillance</Definition>
  <Term id="t3" termID="security">
    <Name>Security</Name>
    <Definition>Security are people who are work for the ACME
surveillance system</Definition>
  <Term id="t4" termID="guard">
    <Name>Guard</Name>
    <Definition>Security guards to protect the premise and
patrol</Definition>
  <Term id="t5" termID="joe_doe">
    <Name>Joe Doe</Name>
    <Definition>Joe Doe, the security
guard</Definition>
  </Term>
</Term>
</Term>
<Term id="t6" termID="staff">
  <Name>Staff</Name>
  <Definition>Staff are the employees who work for the ACME
company</Definition>
  <Term id="t7" termID="fred_bloggs">
    <Name>Fred Bloggs</Name>
    <Definition>Fred Bloggs, the company
employee</Definition>
  </Term>
</Term>
</Term>
</Term>
</ClassificationScheme>
<ClassificationScheme uri="urn:mpeg:mpega:vsaf:example2" id="cs2">
  <Term id="t8" termID="vs">
    <Name>vs</Name>
    <Definition>The surveillance system</Definition>
  <Term id="t9" termID="process">
    <Name>Process</Name>
    <Definition>The definition of terms used in the computer
processing components</Definition>
  <Term id="t10" termID="object">
    <Name>Object</Name>
    <Definition>An detected object</Definition>
  <Term id="t11" termID="blob">
    <Name>Blob</Name>
    <Definition>The moving region within a
video</Definition>
  </Term>
  <Term id="t12" termID="known">
    <Name>Known</Name>
    <Definition>The moving region within a video with a
known identity</Definition>
  </Term>
  <Term id="t13" termID="unknown">
    <Name>Unknown</Name>
    <Definition>The moving region within a video with a
unknown identity</Definition>
  <Term id="t14" termID="known1">
    <Name>Unknown1</Name>

```

```

        <Definition>The identity of an object of unknown
identity</Definition>
    </Term>
  </Term>
</Term>
  <Term id="t15" termID="infrastructure">
    <Name>Infrastructure</Name>
    <Definition>The surveillance system's
infrastructure</Definition>
    <Term id="t16" termID="camera">
      <Name>Camera</Name>
      <Definition>Cameras watching and recording</Definition>
      <Term id="t17" termID="PTZ">
        <Name>PTZ</Name>
        <Definition>A camera with pan-tilt-zoom
mobility</Definition>
      <Term id="t18" termID="Carmeral">
        <Name>8A4EB5D3-B2D8-4b99-A430-11032F0AAAF4</Name>
        <Definition>The camera overlooking
entrance</Definition>
      </Term>
    </Term>
  </Term>
  <Term id="t19" termID="cluster">
    <Name>Cluster</Name>
    <Definition>Camera cluster</Definition>
    <Term id="t20" termID="cluster1">
      <Name>73D1C6F8-5405-4a82-B2AF-295FDD3B5D30</Name>
      <Definition>The cluster of cameras surveying all
entrance points</Definition>
    </Term>
  </Term>
</Term>
</ClassificationScheme>
<ClassificationScheme uri="urn:mpeg:mpeg7:vsaf:author" id="cs3">
  <Term id="t21" termID="james_annesley"/>
</ClassificationScheme>
</Description>
</Mpeg7>

```

An example of Track Level document

```

<?xml version="1.0" encoding="UTF-8" standalone="yes" ?>
<Mpeg7 xmlns="urn:mpeg:mpeg7:schema:2001"
xmlns:mpeg7="urn:mpeg:mpeg7:schema:2001"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="urn:mpeg:mpeg7:schema:2001 Mpeg7-2001_vsaf-07-
2008.xsd">
  <DescriptionMetadata>
    <Comment>
      <FreeTextAnnotation>Sensor details</FreeTextAnnotation>
      <StructuredAnnotation>
        <Who
href="urn:mpeg:mpeg7:vsaf:example1:acme:people:security:guard:joe_doe"/>
        <WhatObject
href="urn:mpeg:mpeg7:vsaf:example2:vs:infrastructure:camera:ptz:cameral"/>
        <WhatAction>
          <Name>Suspicious activity</Name>
        </WhatAction>
        <Where>
          <Name>To the left side of preset 1</Name>

```



```

    </Where>
    <When>
      <Name>2007-12-01T01:10:01:40F25</Name>
    </When>
    <Why>
      <Name>Threshold for movement detector triggered</Name>
    </Why>
    <How>
      <Name>Algorithm X</Name>
    </How>
  </StructuredAnnotation>
  <FreeTextAnnotation>Sensor details manually
added</FreeTextAnnotation>
</Comment>
<PublicIdentifier>4E9C95FA8C47429</PublicIdentifier>
<CreationLocation>
  <GeographicPosition>
    <Point latitude="0.340000" longitude="0.100000" />
  </GeographicPosition>
</CreationLocation>
<CreationTime>2008-01-01T01:01:01:0F25</CreationTime>
<Instrument>
  <Tool>
    <Name>Automatically generated by VSAF Meta-data API</Name>
  </Tool>
  <Setting name="Focus" value="1" />
  <Setting name="Offset" value="25" />
</Instrument>
</DescriptionMetadata>
<Description xsi:type="ContentEntityType">
  <!-- The relationship between the movement detected by the
surveillance camera 8A4EB5D3-B2D8-4b99-A430-11032F0AAAF4 -->
  <!-- and the known people -->
  <!-- this will be based upon appearance -->
  <Relationships id="rship1">
    <Relation id="r1" source="blob1"
      target="urn:mpeg:mpega:vsaf:example2:process:object:unknown1"
      type="urn:mpeg:mpega:vsaf:example2:vs:process:object:unknown"
      strength="0.8"/>
  </Relationships>
  <Relationships id="rship2">
    <Relation id="r2" source="blob2"
      target="urn:mpeg:mpega:vsaf:example2:process:object:unknown1"
      type="urn:mpeg:mpega:vsaf:example2:vs:process:object:unknown"
      strength="0.9"/>
  </Relationships>
  <Relationships id="rship3">
    <Relation id="r3" source="blob3"
      target="urn:mpeg:mpega:vsaf:example2:process:object:unknown1"
      type="urn:mpeg:mpega:vsaf:example2:vs:process:object:unknown"
      strength="0.9"/>
  </Relationships>
  <MultimediaContent xsi:type="VideoType">
    <Header id="coord1" xRepr="0"
xsi:type="Spatial2DCoordinateSystemType" yRepr="0">
      <LocalCoordinateSystem name="preset1">
        <Pixel>0 0</Pixel>
        <CoordPoint>0.700 0.40000</CoordPoint>
        <Pixel>0 0</Pixel>
        <CoordPoint>0.740000 0.150000</CoordPoint>
        <Pixel>0 0</Pixel>
        <CoordPoint>0.742000 0.45000</CoordPoint>
      </LocalCoordinateSystem>
    </MultimediaContent>
  </MappingFunc>ProjectionMatrixMappingFunction</MappingFunc>

```

```

        </LocalCoordinateSystem>
    </Header>
    <Header id="coord2" xRepr="0"
xsi:type="Spatial2DCoordinateSystemType" yRepr="0">
        <LocalCoordinateSystem name="preset2">
            <Pixel>0 0</Pixel>
            <CoordPoint>0.1200000 0.1070</CoordPoint>
            <Pixel>0 0</Pixel>
            <CoordPoint>0.1030000 0.10050</CoordPoint>
            <Pixel>0 0</Pixel>
            <CoordPoint>0.104000 0.80000</CoordPoint>

    <MappingFunct>ProjectionMatrixMappingFunction</MappingFunct>
    </LocalCoordinateSystem>
</Header>
<Video>
    <MediaInformation>
        <!-- Camera cluster information -->
        <MediaIdentification>
            <EntityIdentifier>73D1C6F8-5405-4a82-B2AF-
295FDD3B5D30</EntityIdentifier>
            <VideoDomain
href="urn:mpeg:mpega:vsaf:example2:vs:infrastructure:camera:ptz:camera1" />
            <VideoDomain
href="urn:mpeg:mpega:vsaf:example2:vs:infrastructure:cluster:cluster1" />
        </MediaIdentification>
        <MediaProfile>
            <MediaInstance>
                <InstanceIdentifier/>
                <MediaLocator>
                    <MediaUri>file:///data/video.mpg</MediaUri>
                    <StreamID>1</StreamID>
                </MediaLocator>
            </MediaInstance>
        </MediaProfile>
    </MediaInformation>
    <TextAnnotation>
        <FreeTextAnnotation>The details about the
video</FreeTextAnnotation>
    </TextAnnotation>
    <TemporalDecomposition>
        <StillRegion id="blob1">
            <TextAnnotation>
                <FreeTextAnnotation>Detected blob -- probably
human</FreeTextAnnotation>
                <StructuredAnnotation>
                    <Who
href="urn:mpeg:mpega:vsaf:example2:process:object:unknown1"/>

                </StructuredAnnotation>
            </TextAnnotation>
            <SpatialLocator>
                <Box xmlns:mpeg7="urn:mpeg:mpeg7:schema:2001"
mpeg7:dim="2 2">1 2 3 4</Box>
            </SpatialLocator>
            <MediaTimePoint>2008-01-
01T01:01:01:0F25</MediaTimePoint>
            <VisualDescriptor xsi:type="DominantColorType">
                <SpatialCoherency>17</SpatialCoherency>
                <Value><Percentage>0</Percentage>
                <Index>5 4 4 </Index>
                <ColorVariance>0 0 0 </ColorVariance>
                </Value>
                <Value><Percentage>0</Percentage>

```



```

        <Index>11 10 10 </Index>
        <ColorVariance>0 0 0 </ColorVariance>
        </Value>
        <Value><Percentage>0</Percentage>
        <Index>16 15 17 </Index>
        <ColorVariance>1 0 1 </ColorVariance>
        </Value>
        <Value><Percentage>0</Percentage>
        <Index>22 16 14 </Index>
        <ColorVariance>1 0 0 </ColorVariance>
        </Value>
        <Value><Percentage>0</Percentage>
        <Index>25 24 28 </Index>
        <ColorVariance>1 1 1 </ColorVariance>
        </Value>
    </VisualDescriptor>
</StillRegion>
<StillRegion id="blob2">
    <SpatialLocator>
        <Polygon>
            <Coords xmlns:mpeg7="urn:mpeg:mpeg7:schema:2001"
mpeg7:dim="2 2">5 6 7 8</Coords>
        </Polygon>
    </SpatialLocator>
    <MediaTimePoint>2008-01-
01T01:01:01:0F25</MediaTimePoint>
    <VisualDescriptor numofBitplanesDiscarded="0"
numofCoeff="256" xsi:type="ScalableColorType">
        <Coeff>
            77 7 -39 34 -18 -24 -11 13 -31 -37 -16
13 -25 10 -1 3 7 3 -3 6 15 3 -3 4 -14 2 -10 5 -3 5 1 -4
3 3 3 3 2 -3 0 3 0 7 -4 0 -4 0 3 0 -6 -3 0 0 -14 5 4
-2 0 0 0 -2 1 0 -3 -3 -1 0 -2 -1 1 1 -1 1 0 1 3 0 -1
2 1 2 0 -1 -3 -7 -1 -3 -3 0 0 -5 2 1 3 3 2 0 1 0 -2 -
1 -3 -1 -1 -1 2 2 2 1 3 3 1 0 -3 -3 0 1 0 -3 -3 -5 1
1 1 0 1 0 -1 1 0 -1 1 0 0 1 1 -1 1 0 -1 1 1 0 0 -1 2
1 0 -1 0 1 2 -3 1 1 1 1 -1 -1 1 -1 1 0 2 -1 1 2 0
1 1 -3 -2 0 0 0 -2 2 1 0 -1 0 0 1 -2 0 3 0 0 0 0 1 -
1 -1 0 0 2 -1 1 1 -1 3 0 0 -1 0 0 1 -1 0 0 1 -1 0 1
1 -2 0 0 0 0 1 0 -1 -2 -1 1 0 0 1 0 -2 1 0 1 0 0 0
0 -1 0 0 0 0 1 0 -1 0 0 0 0 1 0 1
        </Coeff>
    </VisualDescriptor>
</StillRegion>
<StillRegion id="blob3">
    <MediaTimePoint>2008-01-
01T01:01:01:0F25</MediaTimePoint>
    <GridLayoutDescriptors
descriptorMask="110000000000000000000000" numofPartX="5" numofPartY="5">
        <Descriptor xsi:type="DominantColorType">
            <SpatialCoherency>0</SpatialCoherency>
            <Value>
                <Percentage>1</Percentage>
                <Index>1 1 1</Index>
            </Value>
        </Descriptor>
        <Descriptor numofBitplanesDiscarded="0"
numofCoeff="256" xsi:type="ScalableColorType">
            <Coeff>
                30 6 -76 48 -61 -26 -6 6 -31 -24 -6
9 -14 5 8 12 -7 -6 -3 9 1 2 0 0 -13 -1 1 6 -3 5 1 -4 3
0 3 3 1 1 1 5 1 -3 0 3 1 2 4 5 2 -2 -4 -1 -2 0 1 -2
0 0 0 -2 1 0 -3 -3 1 -1 -3 -6 -1 -1 -3 -3 -2 -1 0 1 2
2 2 1 2 -1 -3 -7 -1 -3 -3 -4 -1 -1 1 1 3 3 2 0 0 -3 -2
            </Coeff>
        </Descriptor>
    </GridLayoutDescriptors>

```

```

-2 -3 -3 0 -1 2 2 2 1 3 3 1 0 -6 -2 0 1 -1 -1 1 -3 1
1 1 0 1 0 -1 1 1 -1 1 1 0 0 2 1 1 -1 -2 1 0 -1 0 1 1
0 -2 -1 0 0 1 -1 0 0 0 0 0 0 1 0 0 -1 0 0 1 0 0 1 3
1 -2 -1 0 0 0 1 -1 1 -1 -1 0 0 1 -2 0 0 0 0 0 0 1 -1
-2 0 0 -1 1 0 0 0 -2 0 -2 -1 0 0 1 0 0 0 1 -1 0 1 1
-2 0 0 0 0 0 1 0 -1 -2 -1 1 0 -1 1 0 0 0 0 1 0 0 0 0
-1 0 0 0 0 0 1 0 -1 0 0 0 0 0 0 1 0 1
      </Coeff>
      </Descriptor>
    </GridLayoutDescriptors>
  </StillRegion>
</TemporalDecomposition>
</Video>
</MultimediaContent>
</Description>
</Mpeg7>

```

Appendix B

CCTV technology

The different cameras available include fixed, PTZ (pan-tilt-zoom), dome [Prob] and omnidirectional types. The fixed camera is mounted on a bracket within a casing and can move only when the bracket's position is changed. The PTZ camera is a development of the fixed camera design and has a moving bracket, in pan and tilt directions, while the camera provides its zoom functionality, from within its casing. A dome camera has a hemispherical housing within which the camera sits behind a perspex window. This design allows discreet installation and shielding from the elements. As camera technology improves, smaller designs are possible and the dome design becomes more practical. Pan and tilt functionality can be provided by moving the camera within the dome. An omnidirectional camera has a 360° view and is usually installed in a ceiling location within a dome for a top-down view. A catadioptric camera has a similar 360° view but has a different lens type. Both these 360° cameras require a process to undistort and render the scene within the format of a rectangular image and some distortion will remain at the image's periphery. The siting of cameras is important and can require some engineering work. Large PTZ cameras may

be mounted on a pole which required the digging of foundations. Electricity lines and communications cables will also be required. Extra heavy armoured camera housings exist to protect the camera from vandalism.

Different cameras have different specifications which depend upon the target deployment. The variables include frame rate, colour depth and resolution. For example, a police camera used with a motor-vehicle must have a high frame rate and be resistant to vibration. The PTZ movement commands are issued by operators from a remote control room. The cameras recognise proprietary communications protocols and a communications box with a joystick controller will send the correct signals. A software communications protocol may also have been developed to allow control from a PC (Personal Computer). The communications box may act as a communications multiplexor allowing one box to control many cameras. The communications link may also be able to adjust the camera's settings, *e.g.* picture settings, *etc.*

Appendix C

Writing competition entry

If a crime isn't caught on CCTV, did it really happen? (2005)

Awarded joint 1st prize in Faraday Partnerships writing competition

by James Annesley and Alex Starling

Closed-circuit television (CCTV) surveillance is big in Britain – the average person is filmed many times each day going about their daily lives. What is the purpose of it? Is this intrusion into our private lives justified?

These days, the police often waste hours trawling through recorded footage of video tapes. They are looking for evidence in an effort to investigate and, indirectly, to deter crime. In the Brixton nail bomb attack investigation for example, hours and hours of CCTV footage were searched. In this particular case, Brixton's large number of well installed CCTV cameras were invaluable in convicting the perpetrator, but all too often the reality is different owing to badly sited cameras or grainy image quality caused by archaic video recorders, worn tapes or a host of other reasons. Some

of these problems can be addressed by replacing existing analogue CCTV systems with digital technology.

Digital CCTV surveillance systems introduce the possibility of improved image quality, incorporating computer processing, easier installation, better data transmission and, in some cases, a level of artificial intelligence, *e.g.* using triggers to record ‘events’, such as cars driving into car parks, or more exotically, the violation of electronic ‘tripwires’ (just think of a suitable Hollywood blockbuster). As well as causing immediate alarms, these video events can be stored in a database to allow a quick review and additional information supplied by machine vision algorithms can be stored too. We call this information ‘meta-data’ and it can provide abstract information about a scene, such as the presence of objects and any characteristics of these objects. For example, when a robbery is committed and a witness gives a description, the colour of the criminal’s clothes can be used to narrow down likely suspects.

Adding a bit of intelligence to the data gathering process throws up all sorts of interesting options, including the ability to be a bit more selective about what data to record, *i.e.* not recording when nothing happens. It is far better – from a storage point of view – of only recording moving events. Movement is detected by signal processing and, using a model of a background scene, an intelligent CCTV system can understand when significant foreground movement occurs. Meta-data is generated by removing the background and analysing the extracted information. In other words, a description of information that people instinctively understand, such as an image’s colour makeup, is extracted to form a grammar that a computer can interpret. A mathematical procedure generates the values

and these are stored as text, and this meta-data – along with time stamp and event identification – is stored in a database. Examples of uses for this meta-data include quick browsing, event-based searches ('who has walked through a particular doorway within a specified time period') and similarity matching ('have other events like this happened recently?').

A standard way of defining such meta-data addresses problems of compatibility. The meta-data standard MPEG 7, designed by the Moving Pictures Expert Group, when used in conjunction with the MPEG 21 standard, can achieve the secure delivery of compatible meta-data descriptions over computer networks, such as the Internet. Both contain descriptors for protecting the integrity of the digital data, which is otherwise easy to copy and adulterate.

Other methods used in surveillance are object tracking, biometrics to automatically identify people, cooperative camera networks and neural networks. This research is compatible with all such approaches and will contribute to the development in this field.

Implications of successful implementation of this technology are potentially huge. Consider alarms delivering, possibly via mobile Internet or telephone networks, information on the location and the colour signature of objects. Police could review stored data more quickly, not only browsing by events but also using stored meta-data, *i.e.* looking for particular characteristic, such a suspect wearing a red shirt or black trousers.

Contrary to what one might believe from having seen Hollywood's *Enemy of the State*, there is much to achieve before such systems are in everyday use: problems of low-light levels, colour constancy between cameras, occlusions of foreground objects and

changing conditions in an outside environment must still be addressed. One must also consider the many complex – but not insurmountable – implementation issues that would go hand-in-hand with bringing this technology to the market, such as data protection, security and transmission. Then there is also the question of whether the public will accept the use of this technology anyone for a date with big brother?

Bibliography

- [AA05] H. W. Agius and M. C. Angelides. COSMOS-7: Video-oriented MPEG-7 scheme for modelling and filtering of semantic content. *The Computer Journal*, 48(5):545–562, 2005.
- [ABF06] E. L. Andrade, S. Blunsden, and R. B. Fisher. Modelling crowd scenes for event detection. In *Proceedings of the 18th IEEE International Conference on Pattern Recognition (ICPR)*, volume 1, Aug. 2006.
- [Abr87] A. Abramson. *The History of Television, 1942 to 2000*. McFarland, 1987.
- [Ado08] Adobe. TIFF 6.0 specification, 2008. <http://partners.adobe.com/public/developer/tiff/index.html> : last accessed March 2008.
- [ALC⁺06] J. Annesley, V. Leung, A. Colombo, J. Orwell, and S. A. Velastin. Fusion of multiple features for identity estimation. In *Proceedings of the IEE International Conference on Imaging for Crime Detection and Prevention (ICDP)*, June 2006.
- [Ann07] J. Annesley. MPEG-7: Two useful restrictions of Visual Surveillance. Technical Report DIRC-TR-2007-02, Kingston University, 2007.
- [ARG07] F. Angella, L. Reithler, and F. Gallesio. Optimal deployment of cameras for video surveillance systems, Sept. 2007.
- [AS08] J. Annesley and H. Sabirin. ISO/IEC 23000-10/Amd WD2.0 Video surveillance application format conformance and reference software, April 2008.

- [AWW05] I. F. Akyildiz, X. Wang, and W. Wang. Wireless mesh networks: a survey. *Computer Networks*, 47(4):445–487, March 2005.
- [BBC03] BBC. Could X-ray scanners work on the street?, Jan. 2003. <http://news.bbc.co.uk/1/hi/magazine/6309917.stm> : last accessed August 2007.
- [BBC⁺07] S. Boll, T. Burger, O. Celma, C. Halaschek-Wiener, E. Mannens, and R. Troncy. Multimedia vocabularies on the Semantic Web. Technical report, W3C, July 2007. <http://www.w3.org/2005/Incubator/mmsem/XGR-vocabularies-20070724/> : last accessed Feb. 2008.
- [BdWH⁺03] I. Burnett, R. Van de Walle, K. Hill, J. Bormans, and F. Pereira. MPEG-21: Goals and achievements. *IEEE multimedia magazine*, 10:60, 2003.
- [BEM04] J. Black, T. J. Ellis, and D. Makris. A hierarchical database for visual surveillance applications. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, volume 3, June 2004.
- [BFC02] K. Barnard, B. Funt, and V. Cardei. A comparison of computational colour constancy algorithms – Part I: Methodology and experiments with synthesized data. *IEEE Transactions in Image Processing*, 11(9):972–984, Sept. 2002.
- [BHC⁺05] L. M. Brown, A. Hampapur, J. Connell, M. Lu, A. W. Senior, C. Shu, and Y. Tian. IBM Smart Surveillance System (S3): An open and extensible architecture for smart video surveillance. In *Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Sept. 2005.
- [BHNB01] V. Bruce, Z. Henderson, C. Newman, and A. M. Burton. Matching identities of familiar and unfamiliar faces caught on CCTV images. *Journal of Experimental Psychology: Applied*, 7(3):207–218, Sept. 2001.
- [Bim99] A. Del Bimbo. *Visual information retrieval*. Morgan Kaufmann Publishers, 1999.

- [BLFM05] T. Berners-Lee, R. Fielding, and L. Masinter. IETF rfc3986 Uniform Resource Identifier (URI): Generic Syntax, 2005. <http://tools.ietf.org/html/rfc3986> : last accessed March 2008.
- [BPB03] W. P. Berriss, W. G. Price, and M. Z. Bober. The use of MPEG-7 for intelligent analysis and retrieval in video surveillance. In *Proceedings of the IEE Symposium on Intelligent Distributed Surveillance Systems (IDSS)*, Feb. 2003.
- [BR08] G. Bäse and T. Rathgen. ISO/IEC 23000-10 study text of FCD – Information technology – Multimedia Application Format (MPEG-A) – Part 10: MPEG video surveillance application format, April 2008.
- [BS06] W. Bailer and P. Schallauer. Detailed Audiovisual Profile: Enabling interoperability between MPEG-7 based systems. In *Proceedings of the 12th International Conference on Multi Media Modelling (MMM)*, Jan. 2006.
- [BTF⁺05] M. Borg, D. Thirde, J. Ferryman, F. Fusier, F. Brémond, and M. Thonnat. An integrated vision system for aircraft activity monitoring. In *Proceedings of the IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, Jan. 2005.
- [Buc80] G. Buchsbaum. A spatial processor model for object colour perception. *Journal of the Franklin Institute*, 310:1–26, 1980.
- [Car07] J. Carlson. METS Metadata Encoding and Transmission Standard: Primer and Reference Manual, 2007. http://www.loc.gov/standards/mets/METS_Documentation_final_070930_msw.pdf : last accessed Feb. 2008.
- [CBT03] F. Cupillard, F. Brémond, and M. Thonnat. Behaviour recognition for individuals, groups of people and crowds. In *Proceedings of the IEE Symposium on Intelligent Distributed Surveillance Systems (IDSS)*, Feb. 2003.
- [CH98] C. Connolly and H. Palus. *The colour image processing handbook*, chapter 7. Chapman and Hall, 1998.

- [Cie01] L. Cieplinski. MPEG-7 Color Descriptors and their applications. In *Proceedings of the International Conference of Computer Analysis of Images and Patterns (CAIP)*, Sept. 2001.
- [CIE07] CIE. Standard 014-4/e:2007, 2007. [http:// www.cie.co.at/ publ/ abst/ s014_4.html](http://www.cie.co.at/publ/abst/s014_4.html) – last accessed Sept. 2008.
- [COE05] A. Cavallaro, O. Steiger, and T. Ebrahimi. Semantic video analysis for adaptive content delivery and automatic description. *IEEE Transactions on Circuits and Systems for Video Technology*, 15(10):1200–1209, 2005.
- [CT91] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 1991.
- [DAE07] F. Dufaux, M. Ansorge, and T. Ebrahimi. Overview of JPSearch: A standard for image search and retrieval. In *Proceedings of the 5th International Workshop on Content-Based Multimedia Indexing (CBMI)*, June 2007.
- [DCI04] A. Dorado, J. Calic, and E. Izquierdo. A rule-based video annotation system. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(5):622–633, May 2004.
- [DCM08] DCMI. Metadata Terms, Jan. 2008. [http:// dublincore.org/ documents/ dcmi-terms/](http://dublincore.org/documents/dcmi-terms/) : last accessed Feb. 2008.
- [DD03] A. Doulamis and N. Doulamis. Performance evaluation of euclidean and correlation-based relevance feedback algorithms in content-based image retrieval systems. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, volume 1, Sept. 2003.
- [DG06] J. Davis and M. Goadrich. The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning (ICML)*, pages 233–240, 2006.
- [DIG] DIG. DIG35 Initiative Group home page. [http:// www.i3a.org/ i_dig35.html](http://www.i3a.org/i_dig35.html) : last accessed Feb. 2008.

- [DM00] D. Doermann and D. Mihalcik. Tools and techniques for video performances evaluation. In *Proceedings of the 15th IEEE International Conference on Pattern Recognition (ICPR)*, Sept. 2000.
- [Doo03] Doom9. What's on a DVD?, Oct. 2003. <http://www.doom9.org/> (The basics, DVD Structure): last accessed May 2008.
- [DPA98] DPA. Data protection act, 1998. http://www.opsi.gov.uk/Acts/Acts1998/ukpga_19980029_en_1 : last accessed Nov 2008.
- [DPC05] K. Diepold, F. Pereira, and W. Chang. MPEG-A: Multimedia Application Formats. *IEEE Multimedia*, 12(4):34–41, Oct.–Dec. 2005.
- [EBU07a] EBU. Tech 3295-v2 – P-META Semantic Metadata Schema 2.0 Metadata Library, July 2007.
- [EBU07b] EBU/ETI. TS 102 822-2 – Broadcast and On-line Services: Search, select, and rightful use of content on personal storage systems (“TV-Anytime”); Part 2: Phase 1 – System description, Nov. 2007.
- [Eid00] H. Eidenberger. Query model based content-based image retrieval. In *Proceedings of the 8th ACM International Conference on Multimedia*, Nov. 2000.
- [FHMO00] G. D. Finlayson, S. D. Hordley, J. A. Marchant, and C. M. Onyango. Colour invariance at a pixel. In *Proceedings of the 11th British Machine Vision Conference (BMVC)*, Sept. 2000.
- [Fis04] R. B. Fisher. PETS '04 Surveillance Ground Truth Data Set. In *Proceedings of the 6th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, May 2004.
- [FNB07] H. Furntratt, H. Neuschmied, and W. Bailer. MPEG-7 Library : MPEG-7 C++ API implementation, Aug. 2007. <http://iiss039.joanneum.at/cms/fileadmin/mpeg7/files/mp7Jrs2.2.pdf> : last accessed Feb. 2008.
- [GARRM05] B. Girod, A.M. Aaron, S. Rane, and D. Rebollo-Monedero. Distributed video coding. *Proceedings of the IEEE*, 93(1):71–83, Jan 2005.

- [GB06] A. Gilbert and R. Bowden. Tracking objects across cameras by incrementally learning inter-camera colour calibration and patterns of activity. *Lecture Notes in Computer Science*, pages 125–136, 2006.
- [GLMP03] K. Grant, A. T. Lindsay, M. Mainds, and A. Perrott. RETRIEVE: Realtime tagging and retrieval of images eligible for use as video surveillance. In *Proceedings of the IEE Symposium on Intelligent Distributed Surveillance Systems (IDSS)*, Feb. 2003.
- [GM07] D. M. Gavrilu and S. Munder. Multi-cue pedestrian detection and tracking from a moving vehicle. *International Journal of Computer Vision*, 73(1):41–59, 2007.
- [GMMP02] S. Gupte, O. Masoud, R. F. K. Martin, and N. P. Papanikolopoulos. Detection and classification of vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 3(1):37–47, March 2002.
- [Goo04] B. J. Goold. *CCTV and policing: Public area surveillance and police practices in Britain*. Oxford University Press, 2004.
- [Gor00] I. Gordon. AAF: An industry-driven open standard for multimedia authoring. Technical report, AAF, 2000.
- [Got06] P. Gottschalk. Stages of knowledge management systems in police investigations. *Knowledge-Based Systems*, 19(6):381–387, Oct. 2006.
- [GRJ02] D. Greenhill, P. Remagnino, and G. A. Jones. *VIGILANT: Content-querying of video surveillance streams*, pages 193–204. Video-based Surveillance Systems – Computer Vision and Distributed Processing. Kluwer Academic, 2002.
- [GROJ08] D. Greenhill, J.R. Renno, J. Orwell, and G.A. Jones. Occlusion analysis: Learning and utilising depth maps in object tracking. *Image and Vision Computing, Special Issue on the 15th Annual British Machine Vision Conference (BMVC)*, 26(3):430–441, March 2008.
- [GTD⁺07] M. Gruhne, R. Tous, J. Delgado, M. Doeller, and H. Kosch. MP7QF: An MPEG-7 Query Format. In *Proceedings of the 3rd International Conference on Automated Production of Cross Media Content for Multi-Channel Distribution (AXMEDIS)*, Nov. 2007.

- [Hay04] P. Hayes. RDF Semantics, Feb. 2004. [http:// www.w3.org/ TR/ rdf-mt/](http://www.w3.org/TR/rdf-mt/) : last accessed Feb. 2008.
- [HBC⁺04] A. Hampapur, L. M. Brown, J. Connell, M. Lu, H. Merkl, S. Pankanti, A. W. Senior, C. Shu, and Y. Tian. The IBM Smart Surveillance System. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2004.
- [HKK04] M. Hahnel, D. Klunder, and K. F Kraiss. Color and texture features for person recognition. In *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN)*, volume 1, July 2004.
- [HL01] J. Hunter and C. Lagoze. Combining RDF and XML schemas to enhance interoperability between metadata application profiles. In *Proceedings of the 10th International Conference on World Wide Web (WWW)*, May 2001.
- [Hof06] G. Hoffmann. CIE (1931) Color Space, 2006. [http:// www. fho-empden.de/ hoffmann/ ciexyz29082000.pdf](http://www.fho-empden.de/hoffmann/ciexyz29082000.pdf) : last accessed Feb. 2008.
- [HOL07] HOLMES. Unisys HOLMES 2 investigation management system, 2007. [http:// www.holmes2.com/ holmes2/ whatish2/](http://www.holmes2.com/holmes2/whatish2/) : last accessed Feb. 2008.
- [HSD73] R. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, 3:610–621, 1973.
- [HTWM04] W. Hu, T. Tan, L. Wang, and S. Maybank. A survey on visual surveillance of object motion and behaviours. *IEEE Transactions on Systems, Man, and Cybernetics*, 34(3):334–352, 2004.
- [IEE] IEEE. Development of VHS a world standard for home video recording, 1976. [http:// www.ieee.org/ web/ aboutus/ history_center/ vhs.html](http://www.ieee.org/web/aboutus/history_center/vhs.html) : last accessed Feb. 2008.

- [ISO93a] ISO/IEC. 11172-2:1993 – Information technology – Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s – Part 2: Video, 1993.
- [ISO93b] ISO/IEC. 11172-3:1993 – Information technology – Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s – Part 3: Audio, 1993.
- [ISO94] ISO/IEC. 10918-1:1994 – Digital compression and coding of continuous-tone still images: Requirements and guidelines, 1994.
- [ISO98] ISO/IEC. 11172-1:1998 – Information technology – Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s – Part 1: Systems, 1998.
- [ISO00a] ISO/IEC. 13818-2:2000 – Information technology – Generic coding of moving pictures and associated audio information – Part 2: Video, 2000.
- [ISO00b] ISO/IEC. 13818-3:1998 – Information technology – Generic coding of moving pictures and associated audio information – Part 3: Audio, 2000.
- [ISO02a] ISO/IEC. 15938-1:2002 – Information technology – Multimedia Content Description Interface – Part 1: Systems, 2002.
- [ISO02b] ISO/IEC. 15938-2:2002 – Information technology – Multimedia Content Description Interface – Part 2: Description Definition Language, 2002.
- [ISO02c] ISO/IEC. 15938-3:2002 – Information technology – Multimedia Content Description Interface – Part 3: Visual, 2002.
- [ISO02d] ISO/IEC. 15938-4:2002 – Information technology – Multimedia Content Description Interface – Part 4: Audio, 2002.
- [ISO03a] ISO/IEC. 14496-10:2003 – Information technology – Coding of audio-visual objects – Part 10: Advanced Video Coding, 2003.

- [ISO03b] ISO/IEC. 15938-5:2003 – Information technology – Multimedia Content Description Interface – Part 5: Multimedia description schemes, 2003.
- [ISO03c] ISO/IEC. 15938-6:2003 – Information technology – Multimedia Content Description Interface – Part 6: Reference Software, 2003.
- [ISO04a] ISO/IEC. 11179-1:2004(E) – Information technology – Metadata registries (MDR) Part 1: Framework, 2004.
- [ISO04b] ISO/IEC. 14496-1:2004 – Information technology – Coding of audio-visual objects – Part 1: Systems, 2004.
- [ISO04c] ISO/IEC. 14496-2:2004 – Information technology – Coding of audio-visual objects – Part 2: Visual, 2004.
- [ISO04d] ISO/IEC. 14496-3:2005 – Information technology – Coding of audio-visual objects – Part 3: Audio, 2004.
- [ISO04e] ISO/IEC. TR 21000-1:2004 – Information technology – Multimedia framework (MPEG-21) – Part 1: Vision, Technologies and Strategy, 2004.
- [ISO05a] ISO. 2108:2005 – Information and documentation – International standard book number (ISBN), 2005.
- [ISO05b] ISO/IEC. 14496-12:2005 – Information technology – Coding of audio-visual objects – Part 12: ISO Base Media File Format, 2005.
- [ISO05c] ISO/IEC. 14496-12:2005/Amd 1:2007 – Information technology – Coding of audio-visual objects – Part 12: ISO Base Media File Format – Support for timed metadata, non-square pixels and improved sample groups, 2005.
- [ISO05d] ISO/IEC. 15938-11:2005 – Information technology – Multimedia Content Description Interface – Part 11: MPEG-7 profile schemas, 2005.
- [ISO05e] ISO/IEC. 15938-9:2003 – Information technology – Multimedia Content Description Interface – Part 9: Profiles and levels, 2005.

- [ISO05f] ISO/IEC. 21000-2:2005 – Information technology – Multimedia framework (MPEG-21) – Part 2: Digital Item Declaration, 2005.
- [ISO05g] ISO/IEC. 21000-9:2005 – Information technology – Multimedia framework (MPEG-21) – Part 9: File Format, 2005.
- [ISO05h] ISO/IEC. TR 15938-11:2005 – Information technology – MPEG System Technologies – Part 1: Binary MPEG Format for XML, 2005.
- [ISO06a] ISO/IEC. 14496-15:2006 – Information technology – Coding of audio-visual objects – Part 15: AVC File Format, 2006.
- [ISO06b] ISO/IEC. 21000-17:2006 – Information technology – Multimedia framework (MPEG-21) – Part 17: Fragment Identification of MPEG Resources, 2006.
- [ISO06c] ISO/IEC. 23000-2:2006 – Information technology – Multimedia application format (MPEG-A) – Part 2: MPEG music player application format, 2006.
- [ISO07a] ISO/IEC. 13818-1:2007 – Information technology – Generic coding of moving pictures and associated audio information – Part 1: Systems, 2007.
- [ISO07b] ISO/IEC. 23000-3:2007 – Information technology – Multimedia application format (MPEG-A) – Part 3: MPEG photo player application format, 2007.
- [ISO07c] ISO/IEC. TR 23000-1:2007 – Information technology – Multimedia application format (MPEG-A) – Part 1: Purpose for multimedia application formats, 2007.
- [IST] IST. FP6-027231 CARETAKER: Content Analysis and Retrieval Technologies to Apply Knowledge Extraction to massive Recording. [http:// www.ist-caretaker.org/](http://www.ist-caretaker.org/) : last accessed Feb. 2008.
- [ITE] ITEA. SERKET: Security Keeps Threats away. EU Project, [http:// www.research.thalesgroup.com/ software/ cognitive _solutions/ Serket/](http://www.research.thalesgroup.com/software/cognitive_solutions/Serket/) : last accessed March 2008.

- [ITU93] ITU. -T Rec. H.261: Video codec for audiovisual services at p x 64 kbit/s, 1993. [http:// www.itu.int/ rec/ T-REC-H.261- 199303-I/en](http://www.itu.int/rec/T-REC-H.261-199303-I/en).
- [ITU94] ITU. -T Rec. H.281: A far end camera control protocol for videoconferences using H.224, 1994. [http:// www.itu.int/ rec/ T-REC-H.281 -199411-I/ en](http://www.itu.int/rec/T-REC-H.281-199411-I/en) : last accessed March 2008.
- [ITU02] ITU. -R TF.460: Standard-frequency and time-signal emissions – Annex 1, 2002. [http:// www.itu.int/ rec/ R-REC-TF/ e](http://www.itu.int/rec/R-REC-TF/e).
- [JB02] A. Y. Johnson and A. F. Bobick. Relationship between identification metrics: Expected confusion and area under a ROC curve. In *Proceedings of the 16th IEEE International Conference on Pattern Recognition (ICPR)*, volume 3, Aug. 2002.
- [JEI03] JEITA. Image file format for digital still cameras: Exif version 2.21 (Amendment ver2.2) (English version), Sept. 2003.
- [JRJ04] J. Orwell J.R. Renno and G.A. Jones. Evaluation of shadow classification techniques for object detection and tracking. In *Proceedings of the International Conference on Image Processing (ICIP)*, volume 1, pages 143–146, Oct. 2004.
- [JRSS03] O. Javed, Z. Rasheed, K. Shafique, and M. Shah. Tracking across multiple cameras with disjoint views. In *Proceedings of the 9th IEEE International Conference on Computer Vision (ICCV)*, volume 02, page 952, 2003.
- [Kai67] T. Kailath. The divergence and Bhattacharyya distance measures in signal selection. *IEEE Transactions on Communication Technology*, 15(1):52, Feb. 1967.
- [KB01] P. Kaewtrakulpong and R. Bowden. An improved adaptive background mixture model for realtime tracking with shadow detection. In *Proceedings of 2nd European Workshop on Advanced Video Based Surveillance Systems (AVBS)*, Sept. 2001.
- [Kru06] H. Kruegle. *CCTV SURVEILLANCE: Analog and Digital Video Practices and Technology*, chapter 16, pages 411–413. BUTTERWORTH HEINEMANN, 2006.

- [KS03] S. Khan and M. Shah. Consistent labeling of tracked objects in multiple cameras with overlapping fields of view. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10):1355–1360, Oct. 2003.
- [LF04] T. List and R. B. Fisher. CVML: An XML-based Computer Vision Markup Language. In *Proceedings of the 17th IEEE International Conference on Pattern Recognition (ICPR)*, volume 1, Aug. 2004.
- [LG05] M. Lux and M. Granitzer. Retrieval of MPEG-7 based semantic descriptions. In *Proceedings of the 11th Symposium for Database Systems in Business Technology and Web, Workshop WebDB Meets IR (BTW)*, March 2005.
- [Lin03] B. Lindbloom. Useful color equations, 2003. [http:// www.brucelindbloom.com/](http://www.brucelindbloom.com/) : last accessed Feb. 2008.
- [LM71] E. H. Land and J. J. McCann. Lightness and retinex theory. *Journal of the Optical Society America A*, 61(1):1–11, 1971.
- [LMRMJ06] N. Lazarevic-McManus, J. Renno, D. Makris, and G. A. Jones. Designing evaluation methodologies: The case of motion detection. In *Proceedings of the 9th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, June 2006.
- [LMRMJ08] N. Lazarevic-McManus, J.R. Renno, D. Makris, and G.A. Jones. An object-based comparative methodology for motion detection based on the F-measure. *Computer Vision and Image Understanding*, 111(1):74–85, July 2008.
- [LMS05] P. Leach, M. Mealling, and R. Salz. IETF rfc4122 a Universally Unique Identifier (UUID) URN Namespace, 2005. [http:// www.ietf.org/ rfc/ rfc4122.txt](http://www.ietf.org/rfc/rfc4122.txt) : last accessed March 2008.
- [Mad07] R. Mader. From video surveillance to video analytics, Aug. 2007. [http:// www.nrf-arts.org/ stores/ 200708.pdf](http://www.nrf-arts.org/stores/200708.pdf) – last accessed March 2008.
- [Mar04] J. M. Martinez. MPEG-7 Overview. Technical Report N6828, ISO/IEC JTC1/SC29/WG11, Oct. 2004.

- [MFHS02] D. L. McGuinness, R. Fikes, J. Hendler, and L. A. Stein. DAML + OIL: An ontology language for the Semantic Web. *IEEE Intelligent Systems*, 17(5):72–80, Sept.–Oct. 2002.
- [MIS07] MISP. DoD/IC/NSG MISB Motion Imagery Standards Profile, Sept. 2007.
- [MKP02] J. M. Martinez, R. Koenen, and F. Pereira. MPEG-7: the Generic Multimedia Content Description standard, Part 1. *Multimedia, IEEE*, 9(2):78–87, April–June 2002.
- [MMP⁺02] V. Y. Mariano, J. Min, J. H Park, R. Kasturi, D. Mihalcik, H. Li, D. Doermann, and T. Drayer. Performance evaluation of object detection algorithms. In *Proceedings of the 16th International Conference on Pattern Recognition (ICPR)*, Aug. 2002.
- [MSS02] B. S. Manjunath, P. Salembier, and T. Sikora. *Introduction to MPEG-7*. John Wiley and Sons, Ltd., 2002.
- [MvH04] D. L. McGuinness and F. van Harmelen. OWL Web Ontology Language overview, February 2004. <http://www.w3.org/TR/owl-features/> : last accessed Feb. 2008.
- [MWLG02] P. Muneesawang, H. S. Wong, J. Lay, and L. Guan. *Learning and adaptive characterization of visual contents in image retrieval systems*, chapter 11. Handbook of Neural Network for Signal Processing. CRC Press, 2002.
- [NA99] C. Norris and G. Armstrong. *The maximum surveillance society : the rise of CCTV*. Oxford: Berg, 1999.
- [NAT07a] NATO. AEDP-8 Motion Imagery (MI) STANAG 4609 (Edition 2) – Implementation Guide, June 2007.
- [NAT07b] NATO. NSA/0554–AIR STANAG 4609 (Edition 2) – Digital Motion Imagery Standard, 2007.
- [NBTV07] A.-T. Nghiem, F. Bremond, M. Thonnat, and V. Valentin. ET-ISEO, performance evaluation for video surveillance systems. In *Proceedings of IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, September 2007.

- [NHB05] R. Nevatia, J. Hobbs, and B. Bolles. VERL: An ontology framework for representing and annotating video events. *IEEE multimedia magazine*, 12(4):76–86, 2005.
- [Nil99] M. Nilsson. ID3 Tag Version 2.3.0, 1999. [http:// www.id3.org/id3v2.3.0](http://www.id3.org/id3v2.3.0) : last accessed Feb. 2008.
- [NNRM⁺00] P. Ndjiki-Nya, J. Restat, T. Meiers, J. R Ohm, A. Seyferth, and R. Sniehotta. Subjective evaluation of the MPEG-7 retrieval accuracy measure : ANMRR. Technical Report M6029, ISO/IEC JTC1/SC29/WG11, 2000.
- [OMG] OMG. Common Object Request Broker Architecture (CORBA/IOP). [http:// www.omg.org/ technology/ documents/ corba_spec_catalog.htm](http://www.omg.org/technology/documents/corba_spec_catalog.htm) : last accessed March 2008.
- [PA03] F.M. Porikli and A.Divakaran. Multi-camera calibration, object tracking and query generation. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, volume 1, pages 653–656, July 2003.
- [Par96] B.W. Parkinson. *Introduction and Heritage of NAVSTAR, the Global Positioning System*, chapter 1, pages 3–28. Global Positioning System: Theory and applications. American Institute of Aeronautics and Astronautics, 1996.
- [PET] PETS. Performance Evaluation of Tracking and Surveillance workshops. <http://www.cvg.rdg.ac.uk/slides/pets.html>: last accessed Sept. 2008.
- [PF97] F. Provost and T. Fawcett. Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining (KDDM)*, pages 43–48, 1997.
- [Pie80] J. R. Pierce. *An Introduction to Information Theory: Symbols, Signals & Noise*. Courier Dover Publications, 1980.

- [Por03] F.M. Porikli. Inter-camera color calibration by cross-correlation model function. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, volume 2, pages 133–136, Sept. 2003.
- [Proa] Products. Grandeye Corporation – Product descriptions. <http://www.grandeye.com/products.html> : last accessed March 2008.
- [Prob] Products. Overview Ltd. – Product descriptions. <http://www.overview.co.uk/> – last accessed March 2008.
- [Pro04] Products. Toshiba Corporation – What’s DVD?, 2004. <http://www3.toshiba.co.jp/dvd/e/whats/index.htm> : last accessed Feb. 2008.
- [Pro06a] Products. British Telecom Corporation Redcare – Kent Police ANPR case study, 2006. http://www.redcare.bt.com/new_downloads/PDF/Casestudies/CS_Kent_police.pdf.
- [Pro06b] Products. BT redcare to be the ‘electronic eye’ for Arsenal Football Club’s new Emirates Stadium, 2006. <http://www.btplc.com/News/Articles/ShowArticle.cfm?ArticleID=071b4c1a-1c86-42c5-b4c3-0b6814a0fc86> : last accessed June 2008.
- [Pro06c] Products. Sony Corporation – Sony advances intelligent video analytics for IP-based security systems, Sept. 2006. http://news.sel.sony.com/en/press_room/b2b/security/release/25163.html : last accessed Feb. 2008.
- [Pro07] Products. Yahoo Corporation and dynamic drive Corporation – YUI Color Picker, 2007. <http://www.dynamicdrive.com/dynamicindex11/yuicolorpicker/index.htm> : last accessed Feb. 2008.
- [Pro08a] Products. Adobe Corporation – Extensible Metadata Platform (XMP), 2008. <http://www.adobe.com/products/xmp/> : last accessed Feb. 2008.
- [Pro08b] Products. Aralia Ltd. – Product descriptions, 2008. <http://www.araliasystems.com/> : last accessed May 2008.

- [Pro08c] Products. Army Technology – Optronics, Surveillance and Sighting Systems, 2008. [http:// www. army-technology.com/ contractors/ surveillance/ gallery.html](http://www.army-technology.com/contractors/surveillance/gallery.html) : last accessed Feb. 2008.
- [Pro08d] Products. JAI Corporation – Corporate web page, 2008. [http:// www. jai. com/ EN/ Pages/ home. aspx](http://www.jai.com/EN/Pages/home.aspx) : last accessed Feb. 2008.
- [Pro08e] Products. ObjectVideo Corporation – Product descriptions, 2008. [http:// www. objectvideo. com/ products/](http://www.objectvideo.com/products/) : last accessed Jan. 2008.
- [Ren03] M. Renard. ADVISOR: Final evaluation report. Technical Report IST-1999-11287 Deliverable R8.4, European Union, 2003. [http:// www-sop. inria. fr/ orion/ ADVISOR/ finalreport. html](http://www-sop.inria.fr/orion/ADVISOR/finalreport.html) : last accessed March 2008.
- [RH99] T. Randen and J. H. Husoy. Filtering for texture classification: A comparative study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(4):291, April 1999.
- [Ric03] I. E. G. Richardson. *H.264 and MPEG-4 Video Compression: Video Coding for Next-generation Multimedia*. Wiley, 2003.
- [Rij79] C.J. Van Rijsbergen. *Information Retrieval*, chapter 7. Dept. of Computer Science, University of Glasgow, 2nd edition, 1979.
- [ROJ02] J. Renno, J. Orwell, and G. A. Jones. Learning surveillance tracking models for the self-calibrated ground plane. In *Proceedings of the British Machine Vision Conference (BMVC)*, Sept. 2002.
- [RSS] RSS. Advisory board – RSS 2.0 specification. [http:// www. rssboard. org/ rss-specification](http://www.rssboard.org/rss-specification) : last accessed March 2008.
- [SA96] M. Stokes and M. Anderson. A standard default color space for the Internet: sRGB. Technical report, Hewlett-Packard and Microsoft, 1996. [http:// www. w3. org/ Graphics/ Color/ sRGB](http://www.w3.org/Graphics/Color/sRGB) : last accessed March 2008.
- [Sac07] S. Sachoff. Italgo to distribute Eptascape’s MPEG-7 technology in Italy, Aug. 2007. [http:// intelligentvideo. blogspot. com/ 2007/ 08/ eptascape-and-italgo-spa-partner-to. html](http://intelligentvideo.blogspot.com/2007/08/eptascape-and-italgo-spa-partner-to.html) : last accessed Feb. 2008.

- [SBH⁺07] A. W. Senior, L. M. Brown, A. Hampapur, C. Shu, Y. Zhai, R. S. Feris, Y. Tian, S. Borger, and C. Carlson. Video analytics for retail. In *Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Sept. 2007.
- [SD07] F. Schreiner and K. Diepold. MPEG Application Format overview. Technical Report N8942, ISO/IEC JTC1/SC29/WG11, 2007.
- [Sek86] A. Sekula. *The Body and the Archive*, volume 39. MIT Press, 1986.
- [SG99] C. Stauffer and W. E. L. Grimson. Adaptive background mixture models for realtime tracking. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, June 1999.
- [SMP99] SMPTE. 12M-1999 Television – Audio and Film Time and Control Code, 1999.
- [SMP01a] SMPTE. 266M-2001 Television – 4:2:2 Digital Component Systems Digital Vertical Interval Time Code, 2001.
- [SMP01b] SMPTE. 335M-2001 Television – Metadata Dictionary Structure, 2001.
- [SMP01c] SMPTE. 366M-2001 Television – Data Encoding Protocol using Key-Length-Value, 2001.
- [SMW07] H. Schwarz, D. Marpe, and T. Wiegand. Overview of the Scalable Video Coding extension of the H.264/AVC standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 17:1103–1120, Sept. 2007.
- [Som01] I. Sommerville. *Software Engineering*, chapter 11, page 251. Harlow : Addison-Wesley, 2001.
- [SS06] C. Soanes and A. Stevenson. *Concise Oxford English Dictionary*. Oxford University Press, eleventh (revised) edition, 2006. Surveillance.
- [SW06] SW. Netlab neural network software, 2006. [http:// www.ncrg. aston.ac.uk/ netlab/ index.php](http://www.ncrg.aston.ac.uk/netlab/index.php) : last accessed June 2008.

- [SW08] SW. H.264/AVC JM reference software, 2008. <http://iphome.hhi.de/suehring/tml/> : last accessed May 2008.
- [TBH⁺06] R. Troncy, W. Bailer, M. Hausenblas, P. Hofmair, and R. Schlatte. *Enabling Multimedia Metadata Interoperability by Defining Formal Semantics of MPEG-7 Profiles*, volume 4306 of *Semantic Multimedia*, pages 41–55. Springer Berlin / Heidelberg, 2006.
- [Til93] N. Tilley. Understanding car parks, crime and CCTV: Evaluation lessons from safer cities. Technical Report 42, U.K. Home Office Police Research Group, 1993.
- [TKBM99] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers. Wallflower: Principles and practice of background maintenance. In *Proceedings of the 7th IEEE International Conference on Computer Vision (ICCV)*, Sept. 1999.
- [TRE08] TREC. Video retrieval evaluation, 2008. <http://www-nlpir.nist.gov/projects/trecvid/> : last accessed May 2008.
- [UKG06] UKGov. Royal Borough of Kingston upon Thames – Code of Practice for operation of CCTV enforcement cameras in the Royal Borough of Kingston upon Thames, 2006. http://www.kingston.gov.uk/code_of_practice_3.2.pdf : last accessed Feb. 2008.
- [UKG07a] UKGov. 18th Report of session 2006/07 House of Lords Paper 90 EU Committee – Prüm: An effective weapon against terrorism and crime?, May 2007.
- [UKG07b] UKGov. CameraWatch web page, 2007. <http://www.camerawatch.org.uk/> : last accessed Feb. 2008.
- [UKG07c] UKGov. U.K. Home Office – Imagery library for Intelligent Detection Systems (i-LIDS), 2007. <http://scienceandresearch.homeoffice.gov.uk/hosdb/cctv-imaging-technology/video-based-detection-systems/i-lids/> : last accessed March 2008.
- [UKG08a] UKGov. MOD Grand Challenge, 2008. <http://www.challenge.mod.uk/> : last accessed May 2008.

- [UKG08b] UKGov. U.K. Information Commissioner – CCTV Code of Practice, 2008.
- [VID08] VIDE. ViDe Group web page, 2008. [http:// www.vide.net/](http://www.vide.net/) : last accessed Feb. 2008.
- [VLS04] S. A. Velastin, B. Lo, and H. Sun. A flexible communications protocol for a distributed surveillance system. *Journal of Network and Computer Applications*, 27(4):221–253, Nov. 2004.
- [WD98] E. Wallace and C. Diffley. CCTV: Making it work – CCTV Control Room ergonomics. Technical Report 14-98, U.K. Police Scientific Development Branch, 1998.
- [Whe02] J. Wheeler. Airport security. Technical report, Department for Transport, UK, 2002.
- [Wik08] Wiki. Color, 2008. [http:// en.wikipedia.org/ wiki/ Color](http://en.wikipedia.org/wiki/Color) : last accessed Nov 2008.
- [Wis06] A. Wiseberg. CCTV counters casino crime, June 2006. [http:// www. sourcesecurity.com/ news/ articles/ 628.html](http://www.sourcesecurity.com/news/articles/628.html) : last accessed Feb. 2008.
- [WP98] A. Watt and F. Policarpo. *The computer image*, chapter 14, pages 325–331. Addison Wesley Longman Limited, 1998.
- [WP04] K. Wong and L. Po. MPEG-7 Dominant Color descriptor based relevance feedback using merged palette histogram. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 3, May 2004.
- [Wri02] R. Wright. Some considerations on using P_META and Dublin Core, 2002. EBU Project Group P/Meta, [http:// www.ebu.ch/ en/ technical/ trev/ trev_294-dublin_core.pdf](http://www.ebu.ch/en/technical/trev/trev_294-dublin_core.pdf) : last accessed May 2008.
- [XE01] M. Xu and T. J. Ellis. Illumination-invariant motion detection using colour mixture models. In *Proceedings of the British Machine Vision Conference (BMVC)*, Sept. 2001.

- [XR] X-Rite. Colorchecker chart. http://www.xrite.com/product_overview.aspx?ID=820 – last accessed Sept. 2008.
- [YWHT04] S. Yu, L. Wang, W. Hu, and T. Tan. Gait analysis for human identification in frequency domain. In *Proceedings of the 3rd International Conference on Image and Graphics (ICIG)*, Dec. 2004.