TOWARDS INTELLIGENT CONTENT BASED RETRIEVAL OF WILDLIFE VIDEOS

J. Ćalić[1], N. Campbell[1], A. Calway[1], M. Mirmehdi[1], T. Burghardt[1]
S. Hannuna[1], C. Kong[1], S. Porter[2], N. Canagarajah[2], D. Bull[2]

[1] Department of Computer Science,
[2] Department of Electrical & Electronic Engineering,
University of Bristol, Merchant Venturers Building,
Woodland Road, Bristol BS8 1UB, UK

ABSTRACT

In this paper we describe a system that embarks upon the problem of efficient video processing and representation for automatic semantic classification and modelling for indexing and retrieval of large multimedia databases. The major focus of the system is the integration of a large-scale wildlife digital video archive with manually annotated semantic metadata organised in a structured taxonomy and media classification system. Our research challenges the problems of temporal analysis of video, intelligent key-frame extraction and summarisation, animal gait analysis, unsupervised semantic modelling and classification and identification of individual animals. In order to evaluate the system a demonstrator is being developed as a part of the ICBR project within the 3C Research programme of convergent technology research for digital media processing and communications.

1. INTRODUCTION

Though being one of the most important areas in the multimedia and computer vision research, the evolution of a functional multimedia management system is hindered by the "semantic gap" between the simplicity of content descriptions that can be currently computed automatically and the richness of user's queries posed for media search and retrieval. In order to bridge the semantic gap, it is necessary to focus our research towards the utilisation of the underlying processes that connect perceptual features of the multimedia and its meaning. Therefore, a large-scale system that merges the semantic text-based retrieval approach to multimedia databases with content-based feature analysis and investigates the signification links between them could shed more light on this problem.

The ICBR (Intelligent Content-based Retrieval) project [1] concentrates on the development of a large-scale centralized multimedia server enabling large quantities of media to be stored, indexed, searched, retrieved, transformed and transmitted within a framework encompassing issues related to multimedia management, content analysis and semantic multimedia retrieval. ICBR brings a unique opportunity to deal with semantic gap issues by integrating a large video database with its semantic description organised in a structured taxonomy. The ICBR project partners are Granada from Bristol, one of Europe's biggest programme producers contributing with more than 12000 hours of annotated wildlife footage and Matrix Data from London, specialists in data management, bringing into the project a large taxonomy and classification structure of approx. 75,000 concepts.

2. RELATED WORK

A common content-based multimedia retrieval system addresses low-level features like colour, textures, shape, spatial relationships to obtain fully automated numeric descriptors from objective measurements of the visual content and to support retrieval by content based on combinations of these features. Query by Example is a typical technique that utilises this paradigm [2, 3]. However, retrieval of multimedia is generally meaningful only if performed at a high level of abstraction based on semantically meaningful categories. Therefore, a great deal of research effort has been put recently into developing a system that will enable automatic semantic analysis and annotation of multimedia.

Initially, temporal structure of media is analysed by tracking spatial domain sequence features [4], or more recently compressed domain features [5]. A video sequence is summarised by choosing a single key-frame [6] or a structured set of key-frames [7] to represent the content in the best possible way. Low-level features such as colour, texture, shape, etc. are extracted and stored in the database as video metadata. Using this information, various methods of media representation have been proposed [8, 9] targeting user centred retrieval and the problem of the semantic gap. Utilising metadata information, attempts to apply semantic analysis to a limited contextual space were presented [10].

3. ICBR SYSTEM OVERVIEW

By incorporating a large manually annotated multimedia archive alongside its structured semantic classification, ICBR has a unique potential to tackle a wide spectrum of research problems in content-based multimedia retrieval. The major guideline of the development process is the semantic information, considered as the ground truth, upon which content-analysis algorithms base their internal representations and knowledge.

In the initial stage of temporal analysis and content summarisation, production knowledge and ground-truth annotation are utilised in algorithm design to achieve robust results on real-world data. Production information such as camera artefacts, overexposures, camera work as well as the shot composition type are automatically extracted enabling content summarisation and temporal representation on higher a semantic level than the conventional video summarisation methods.

In addition to the algorithms developed within predecessor projects AutoArch and Motion Ripper [11], the ICBR project brings novel methods of motion and gait analysis for content analysis and classification. Objectives set for the semantic modelling module include identification of overall scene type as well as individual object classification and identification. The approach is to design unsupervised classification algorithms trained on the rich and structured semantic ground-truth metadata that exploits specific media representations generated by the content adaptation module and a set of MPEG-7 audio/visual descriptors.

4. FEATURE EXTRACTION

The segmenter module parses digitised video sequences into shots, additionally labelling camera artefacts and tape errors on the fly. It utilises block-based correlation coefficients and histogram differences to measure the visual content similarity between frame pairs. In order to achieve real-time processing capability, a two-pass algorithm is applied. The shot boundary candidates are labelled by thresholding chi-square global colour histogram frame differences. In the second pass, more detailed analysis is applied to all candidates below a certain predetermined threshold. At this stage, hierarchical motion compensation is employed to ensure the algorithm is robust in the presence of camera and large object motion. This algorithm achieves a higher recall and precision compared with the conventional shot detection techniques [12]. In addition, the algorithm detects gradual shot transitions by measuring the difference between the visual content of two distant frames. Motion estimates obtained from the shot cut detection algorithm are used to track regions of interest through the video sequence. This enables the distinction between content changes caused by gradual transitions and those caused by camera and object motions.

An important stage of the video representation is the abstraction of a data intensive video stream as a set of still images, called key-frames. A key-frame should convey both the perceptual features and the semantic content of a shot in the best possible way. The aim is to extract a single *intelligent* key-frame and to generate low-level features from it and the adjacent frames. In order to tackle subjective and adaptive intelligent key-frame extraction, we analyse spatio-temporal behaviour of the regions present in the scene, as well as the camera work parameters. A set of heuristic rules is designed in order to detect the most appropriate frame representative [13]. To evaluate the key-frame extraction module, a set of tapes is hand labelled with good (GD), bad (BD) and very good (VG) regions for a key-frame. Results for two tapes with approximately 90 minutes of' wildlife rushes from the ICBR database are given in Table 1.

	VG	GD	NONE	BD	SHOTS	PR_1	PR_2
T01000.mov	52	121	23	14	210	93.5%	93.3%
T01002.mov	10	129	32	42	213	73.6%	80.3%

Table 1. IKF extraction precision measures

5. SEMANTIC ANALYSIS

5.1. Animal Face Tracking

In such a complex domain as the wildlife video production, it is essential to narrow the contextual space of wildlife's heterogeneous semantics. Therefore, we have developed an algorithm that tracks animal faces in wildlife rushes and populates a database with appropriate semantics. It is an adapted version of the human face detection method that exploits Haar-like features and the AdaBoost classification algorithm [14]. Detected face regions are tracked using the Kanade-Lucas-Tomasi method, fusing it with a specific interest model applied to the detected face region [15]. The model is based upon a confidence parameter that gets assigned to each instance of the interest model. It carries the detection density. Analysis of this parameter allows more robust decision making on object appearance and disappearance. Temporal and spatial integration of detections allows discarding detection outliers. The confidence parameter gets initialized with a first face detection. Once the parameter overrides an acceptance threshold the model is accepted as an animal face, even for past frames. This procedure allows the post-labelling of frame content with knowledge (ensured appearance and disappearance) gained after its actual occurrence.

This specific tracking model achieves reliable detection and temporally smooth tracking of animal faces. In addition, it creates strong priors in the process of learning animal models as well as extracting additional semantic information about the animal's behaviour and

environment. The information extracted in the detection and tracking processes is used to annotate the semantic description of a wildlife clip with the presence of a given animal specimen and its basic locomotive behaviour, like walking, running or standing.

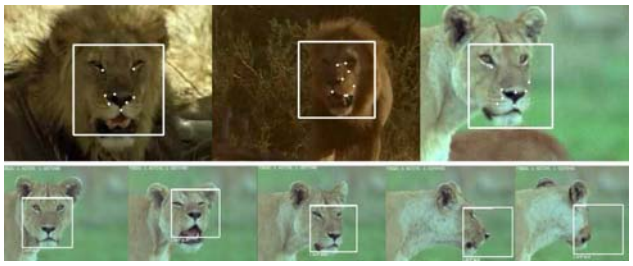

Figure 1. Chosen feature points stipulated in the centre of detected faces for further tracking of this region; Lower row shows the detected lion face, tracked through various poses;

5.2. Gait Analysis and Unsupervised Classification

One of the main challenges in the wildlife domain is that the animals naturally blend in well with their environment. This causes segmentation algorithms, which rely solely on texture information, to be unreliable. Hence motion based segmentation is particularly attractive in the wildlife context. Commonly, sparse sets of points have been segmented into different regions using a RANSAC algorithm. This yields sparse region segmentation. The sparse set of tracked points can be used to obtain trajectories for image regions. Here, we use the dense flow for foreground region that provides information about the internal motion characteristics of the foreground object(s), as given in Figure 2.

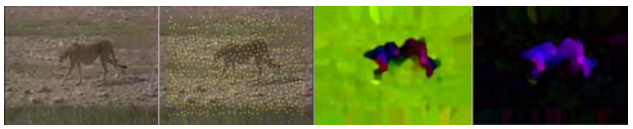

Figure 2. Extraction of a dense motion flow from a sparse point set with foreground object extraction

This information is analysed using Principal Component Analysis. PCA provides us with a set of responses, which tell us how variation in the information it models, changes over time. The frequency characteristics of these responses and the phase differences between them provide parameters for classifiers, as depicted in Figure 3. Using this technique a simple Nearest Neighbour classifier has over a 90% success rate in differentiating solitary side on gait from other animal movements, for the input training set.

This system has been automated such that the only input required is the key frame around which the algorithm should be applied [16]. For 79 key frames input into the algorithm, seven were clustered with the side on gait training data. In actuality, three were side on gait, one was a tiger running side on and another was a young tiger playfully stalking a man (who was walking). Two were misclassified. These results were particularly impressive as this tape had been previously scanned by eye whilst searching for training data, and all of these sequences had been missed. Two of the side on gait sequences featured very small foreground objects yet their motion was still successfully segmented.

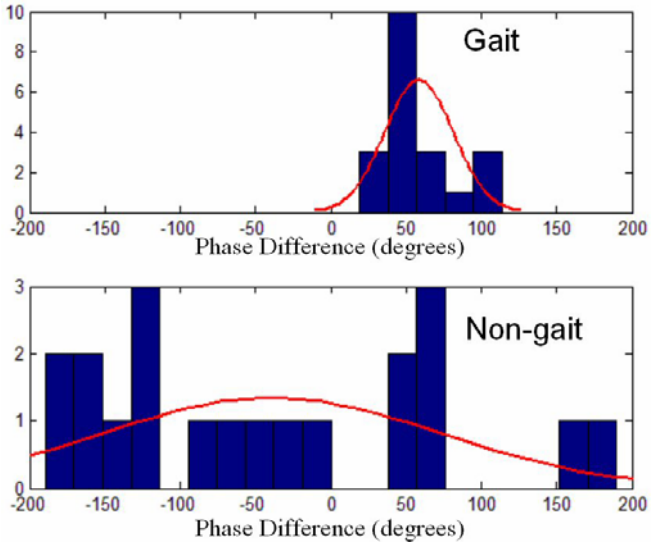

Figure 3. Histograms of phase differences between 3rd and 4th response in training set for gait and non-gait

In addition to this system, we model the 2-D motion field associated with an animal in order to extract characteristic motion information for the purpose of automatic classification. This would allow the automated generation of groundtruth information with the huge database of wildlife sequences we are dealing with in this project.

As an example, Figure 4 shows the typical regions of significant motion that might be used to represent a running hyena. As well as the leg regions, the motion is detected on regions around the head and lower back all of which help characterize the gait of the animal class.

Figure 4. Significant Motion regions of a running hyena

5.3. Individual Animal Identification

In addition to species classification and recognition, there is a demand in wildlife documentary production for a system that will be able to distinguish between particular animal individuals present in the footage. As an example system, we have developed a real-time algorithm that can identify individual African penguins (*Spheniscus demersus*) by analysing a pattern of black spots on their chests that does not change during their adult life. An extraction of the chest spot pattern allows the generation of a unique biometrical identifier for each penguin from still images as well as from video clips [17]. Using these identifiers an authentication of filmed or photographed African penguins against a population database can be performed, as depicted in Figure 5.

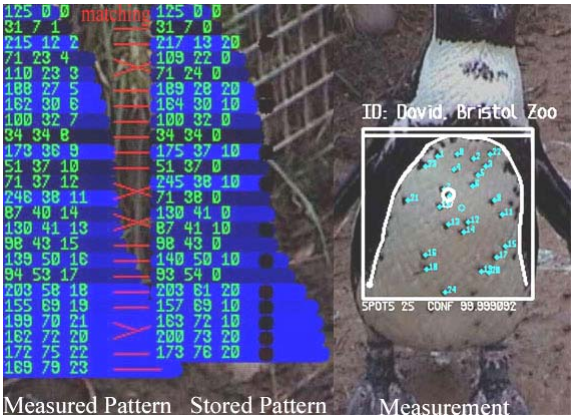

Figure 5. Extraction of the chest pattern and comparison with the unique identifier

6. CONCLUSIONS

This paper gives an outline of the ICBR project and describes the approach taken to utilise an unique opportunity in having both large real-world multimedia database and manually annotated semantic description of media organised in a structured taxonomy. The final demonstrator will be integrated into a media production system showing functionalities of a third generation content-based multimedia retrieval system such as semantic retrieval and browsing, automatic content generation and adaptation, etc. bringing novel multimedia management concepts into a real world of wildlife production.

7. ACKNOWLEGDEMENTS

The work reported in this paper has formed part of the ICBR project within the 3C Research programme of convergent technology research for digital media processing and communications whose funding and support is gratefully acknowledged. For more information please visit www.3cresearch.co.uk .

8. REFERENCES

[1] J. Calic, N. Campbell, M. Mirmehdi, B. Thomas, R. Laborde, S. Porter and N. Canagarajah, "ICBR - Multimedia management system for Intelligent Content Based Retrieval", *Proc. of CIVR 2004*, pp. 601-609, Springer LNCS 3115, 2004.

[2] M. Flickner et al., "Query by image and video content: The QBIC system", *IEEE Computer* 28, pp 23-32, September 1995.

[3] A. Pentland, R.W. Picard, and S. Sclaroff, "Photobook: Content-based manipulation of image databases", *Intern. J. Comput. Vision*, 18(3), pp 233-254, 1996.

[4] H. J. Zhang, A. Kankanhalli, and S. Smoliar, "Automatic Partitioning of Full-Motion Video", *Multimedia Systems*, Vol. 1, No. 1, pp.10-28, 1993.

[5] B.-L. Yeo and B. Liu, "Rapid Scene Analysis on Compressed Video", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 5, No. 6, pp. 533-544, December 1995.

[6] J. Calic, E. Izquierdo, "Efficient Key-Frame Extraction and Video Analysis", *2002 International Symposium on Information Technology*, 8-10 April 2002, Las Vegas, NV, USA. IEEE Computer Society, pp. 28-33, 2002,

[7] A. Girgensohn, J. S. Boreczky, "Time-Constrained Keyframe Selection Technique", *Multimedia Tools Appl*., 11(3): 347-358 (2000)

[8] M. Davis, "Knowledge Representation for Video", Proc. *Of 12th National Conference on Artificial Intelligence*, Seattle, USA, AAAI Press, pp. 120-127, 1994.

[9] C. Dorai, S. Venkatesh: Computational Media Aesthetics: Finding Meaning Beautiful. *IEEE MultiMedia* 8(4): 10-12 (2001)

[10] M. Naphade, T. Kristjansson, B. Frey, T. S. Huang, "Probabilistic Multimedia Objects Multijects: A novel Approach to Indexing and Retrieval in Multimedia Systems", *Proc. IEEE ICIP*, Volume 3, pages 536-540, Oct 1998

[11] D. Gibson, N. Campbell, B. Thomas, "Quadruped Gait Analysis Using Sparse Motion Information", *Proc. of ICIP*, IEEE Computer Society, September 2003.

[12] Sarah V. Porter, Majid Mirmehdi, Barry T. Thomas, "Temporal video segmentation and classification of edit effects*", Image Vision Comput*. 21(13-14): 1097-1106, 2003.

[13] J. Calic, B. T. Thomas, "Spatial Analysis in Key-frame Extraction Using Video Segmentation", *Proc. of WIAMIS'2004*, April 2004, Instituto Superior Técnico, Lisboa, Portugal

[14] P. Viola and M. Jones. "Robust Real-time Object Detection", *Second International Workshop on Statistical and Computational Theories of Vision*, Vancouver, USA, (2001).

[15] T. Burghardt, J. Calic and B. Thomas, "Tracking Animals in Wildlife Videos Using Face Detection", *Proc. of European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology*, November 2004.

[16] S. L. Hannuna, N. W. Campbell and D. P. Gibson, "Segmenting Quadruped Gait Patterns From Wildlife Video", to be presented at *VIE 2005 - IEE Visual Information Engineering Conference, 2005.*

[17] T. Burghardt, B. Thomas, P. J. Barham, J. Calic, "Automated Visual Recognition of Individual African Penguins", *5th International Penguin Conference, Ushuaia, Tierra del Fuego, Argentina, September 2004*.