Analysing Animal Behaviour in Wildlife Videos Using Face Detection and Tracking

Tilo Burghardt, Janko Ćalić

Department of Computer Science

University of Bristol, Bristol BS8 1UB, UK

e-mail: {burghard, janko}@cs.bris.ac.uk

Abstract

This paper presents an algorithm that categorises animal behaviour using detection and tracking of animal faces in wildlife videos. As an example, the algorithm is applied to lion faces. The detection algorithm is based on a human face detection method, utilising Haar-like features and AdaBoost classifiers. The face tracking is implemented using the Kanade-Lucas-Tomasi tracker and by applying a specific interest model to the detected face. By combining the two methods in a specific tracking model, a reliable and temporally coherent detection/tracking of animal faces is achieved. The information generated by the tracker is used to automatically annotate the animal's behaviour. The annotation classes of locomotive processes for a given animal species are predefined by a large semantic taxonomy on wildlife domain. The experimental results are presented.

1 Introduction

The problem of semantic annotation in such a complex domain as wildlife video has highlighted the importance of efficient and reliable algorithms for animal detection and tracking. Not only to recognise the presence of an animal and determine its species but to narrow the contextual space of wildlife's heterogeneous semantics. However, there have been only a few attempts to solve this problem, mainly focused at a particular and narrow domain rather than offering a more general solution. Walther et al. [1] utilise saliency maps to minimize multi-agent tracking of low-contrast translucent targets in underwater footage. Haering et al. [2] attempts to detect high-level events like hunts by classifying and tracking moving object blobs using a neural network approach. Aiming at multiple object tracking, Tweed and Calway [3] develop a periodic model of animal motion and exploit conditional density propagation to track flocks of birds. An interesting approach, by Ramanan and Forsyth [4], takes into account the temporal coherency and builds appearance models of animals. Though dealing only with human faces, the algorithm by Everingham et al. [5] combines a minimal manually labelled set with an object tracking technique to gradually improve the detection model. Trying to tackle the problem of animal behaviour classification, Gibson et al. [6] and Hannuna et al. [7] have detected and classified animal gait by applying statical analysing to a sparse motion information extracted from wildlife footage.

In this paper we present an algorithm that tracks animal faces in wildlife rushes and populates a database [8] with appropriate semantics about their locomotive behaviour. The detection algorithm is an adapted version of a human face detection method that exploits Haar-like features and the AdaBoost classification algorithm [9]. The tracking is implemented using the Kanade-Lucas-Tomasi method, fusing it with a specific interest model applied to the detected face region. This specific tracking model achieves reliable detection and temporally smooth tracking of animal faces. Furthermore, the tracking information is exploited to classify locomotive behaviour of the tracked animal, e.g. lion walking left or trotting towards the camera. Finally, the extracted metadata about the tracked animal specimen and its behaviour creates strong priors in the process of learning animal models as well as in extracting the additional semantic information about the animal's behaviour and environment. The presented algorithm is a part of a large content-based retrieval system [10] that focuses on challenges in the wildlife documentary production. Therefore the information on the existence and behaviour of a specific animal is vital in the process of video media reuse from an large digital video repository.

This paper is organised as follows. In Section 2, a method for animal face detection that uses Haar-like features and AdaBoost classifiers is presented.

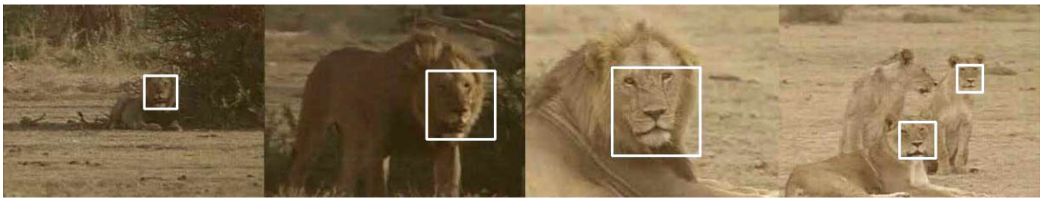


Figure 1: Lion face detection and tracking: images show scale invariance, slight posture invariance and multiple detections of the algorithm

Section 3 describes the algorithm that combines detection with tracking in a joint interest model. The semantic classification using this approach is presented in Section 4, while the final conclusions are given in Section 5.

2 Animal Face detection using Haar-like features

To measure image support for the presence of an animal face we utilize an approach developed for the recognition of human upright faces introduced by Viola and Jones [9]. The original algorithm exploits local contrast feature configurations of the luminance channel. We use a modified version where features are extracted from a colour opponent map calculated as the difference of the red and green component. This map shows robustness against shadows and illumination changes *for natural scenes* as presented by Troscianko et. al. [11]. We apply the technique for detecting animal faces as an example application, depicted in Figure 1.

The procedure employs a pool of Haar-like characteristics as primary feature space. Each Haar-like feature f represents a rectangular local contrast property outlining the existence of either an edge, line or point (see Figure 2) in the colour difference map. As given in Equation 1, f consists of N rectangular components $r = (x, y, w, h)$ that contribute with their average pixel value $S(r)$ weighted by v.

$$f(I) = \sum_{n \in N} v_n \cdot S(r_n) \tag{1}$$

Viola and Jones show in [9] that each $S(r)$ can be computed in a highly efficient way using only four accesses on the integral image (II) as defined in Equation 2. For adjacent features the number decreases respectively.

$$\begin{aligned} S(r) = II(x-1, y-1) + II(x+w-1, y+h-1) \\ -II(x+w-1, y-1) - II(x-1, y+h-1) \end{aligned} \tag{2}$$

The integral image can be derived in linear time complexity from the original image I using an iterative approach, as given in Equation 3.

$$\begin{gathered} II(x, y) = \sum_{j=0}^{j=y} \sum_{i=0}^{i=x} I(i, j) \\ II(-1, y) = 0 \ \wedge \ II(x, y) = II(x-1, y) + H(x, y), \\ H(x, -1) = 0 \ \wedge \ H(x, y) = H(x, y-1) + I(x, y). \end{gathered} \tag{3}$$

Gentle AdaBoost [12] is utilized to compose the most characteristic of these Haar-like features into a strong classifier trained on a labelled set of various positive and negative sample patches. As shown in Figure 2, most unique

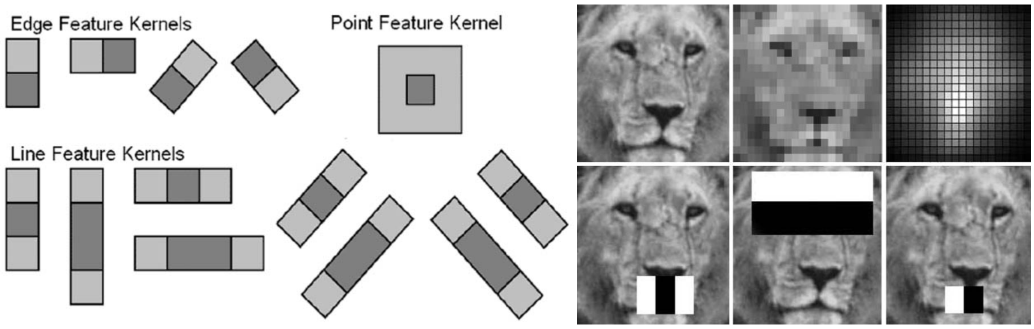


Figure 2: Haar-like Feature Kernels: (left) Selection of Haar-like feature pool introduced by Lienhart and Maydt in [12]. We use this extended set. (right) Sample patch, quantized down sampled patch, feature density and three most characteristic features chosen by Gentle AdaBoost.

features for lion faces are picked by the algorithm around distinctive and meaningful areas, e.g. the nose, the eyes and the jaw. Overall, the classifier combines 250 features and thereby achieves a very small false positive rate below 10^{-4} at a hit rate of 93%.

The ROC learning curve and a comparison with luminance based classifiers is given in Figure 3. Following this approach, detectors are constructed for both upright frontal and side views.

3 Tracking Extension

The Haar-detector performs scale invariant real-time recognition but is limited to a finite number of trained poses. Motion prediction for frames of interim non-detection using motion models is difficult since most animal movements follow a dynamically changing non-linear pattern. Instead, we try to reconstruct the animal trajectory through frames of non-detection

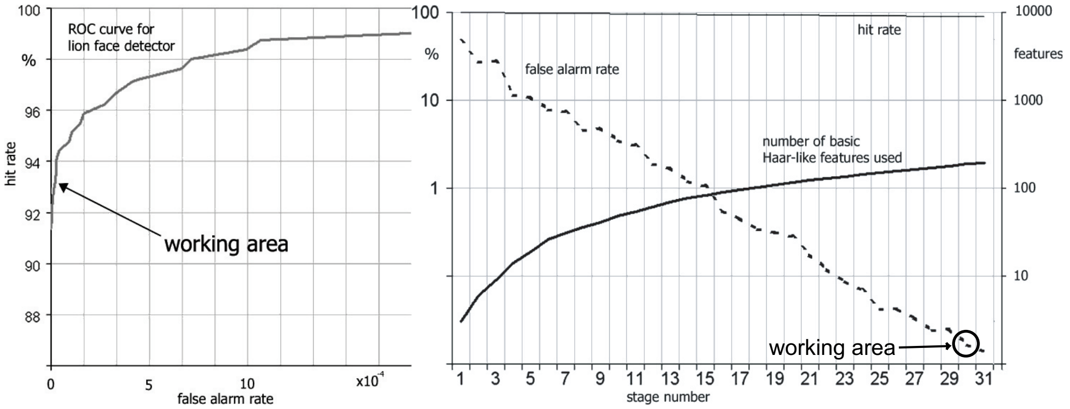


Figure 3: (left) ROC learning curve for lion face training using 680 positive examples images and 1000 negative images; (right) Outline of the learning process focused on the decrease of false alarms compared to the number of necessary single Haar-like features used to achieve this reduction

based on observations of low-level features. Tracking lo-level features can be performed independently from high-level detection success.

Rectangular interest models m_i are established at image locations where lion faces are first detected. The central area of an interest model is then populated with a sparse set of points carrying strongest available gradient in both image directions (see Figure 4) following a suggestion by Shi and Tomasi [13]. The stipulated points are tracked utilizing a pyramidal implementation of the Kanade-Lucas-Tomasi tracker. Single feature points within the cloud are continuously updated, any feature point that is lost or leaves the interest rectangle is discarded and replaced by a newly chosen point close to the center of the interest model.

We observe the centre of mass c_i of the cloud associated with interest model m_i. Its frame-to-frame position change $\mathrm{c}_i(t)-c_i(t-1)$ gives an estimate

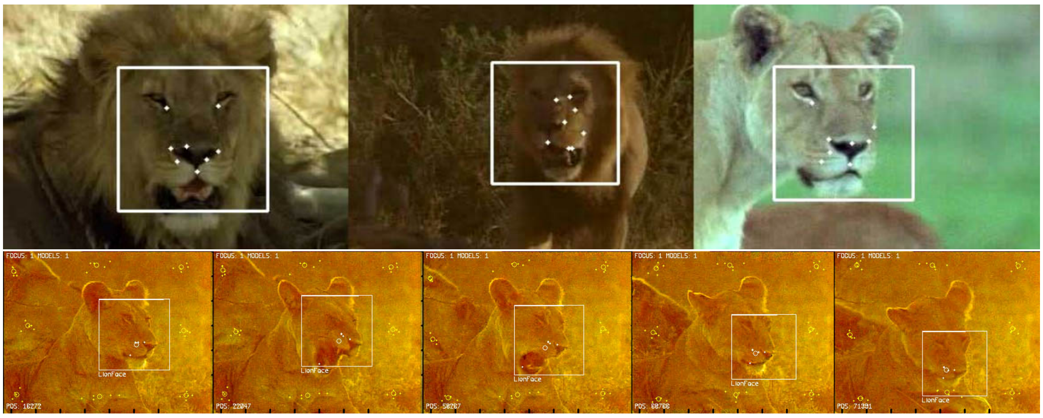


Figure 4: (top) Chosen feature points stipulated in the center of detected faces for further tracking of this region; (bottom) Detected lion face and its tracking through various poses.

for the motion of the interest model m_i. In case of a nearby face detection $d_i(t)$, the estimate is corrected towards this high-level detection in order to avoid model drift. An experimentally chosen parameter τ controls the strength of the correction. The process is summarized by Equation 4 and graphically illustrated in Figure 5.

$$m(t) = \left[(1+\tau, 1, \tau) \begin{pmatrix} c(t) \\ m(t-1) \\ d(t) \end{pmatrix} \right] \cdot (1+\tau)^{-1} \qquad (4)$$

The dynamic nature of the point cloud update prevents the interest model from drifting and generates some robustness for shaky camera action, fast animal motion, partial occlusions and insufficient image gradient for particular tracked points. Illustrative examples of such cases are given in Figure 6. The

Figure 5: ***>>>Combination of Detection and Tracking: graph shows x and y values of frontal lion face detection (points) and positioning of the interest model (curves) established on the bases of detections and point cloud tracking<<<***

Figure 6: ***>>>Combination of Detection and Tracking: graph shows x and y values of frontal lion face detection (points) and positioning of the interest model (curves) established on the bases of detections and point cloud tracking<<<***

proposed procedure creates continuous and coherent trajectory information for the interest model even through frames without detectable animal face.

3.1 Confidence accumulation

Temporal density of animal face detections along the extracted trajectory gives an estimate for the likelihood p of the actual presence of an animal face. Decisions about animal appearance and disappearance are made based on p. To calculate the temporal density a confidence parameter gets assigned to each interest model m_i.

The parameter gets initialized with a first detection, decreases over time and increases accumulating detections. Once the parameter overrides an initial acceptance threshold the model is accepted as an animal face, even for past frames (see Figure 7). This procedure allows the post-labeling of frame content with knowledge (ensured appearance and disappearance) gained after its actual occurrence. For accepted models, parts of the trajectory tracked without detection are validated by the next occurring detector response.

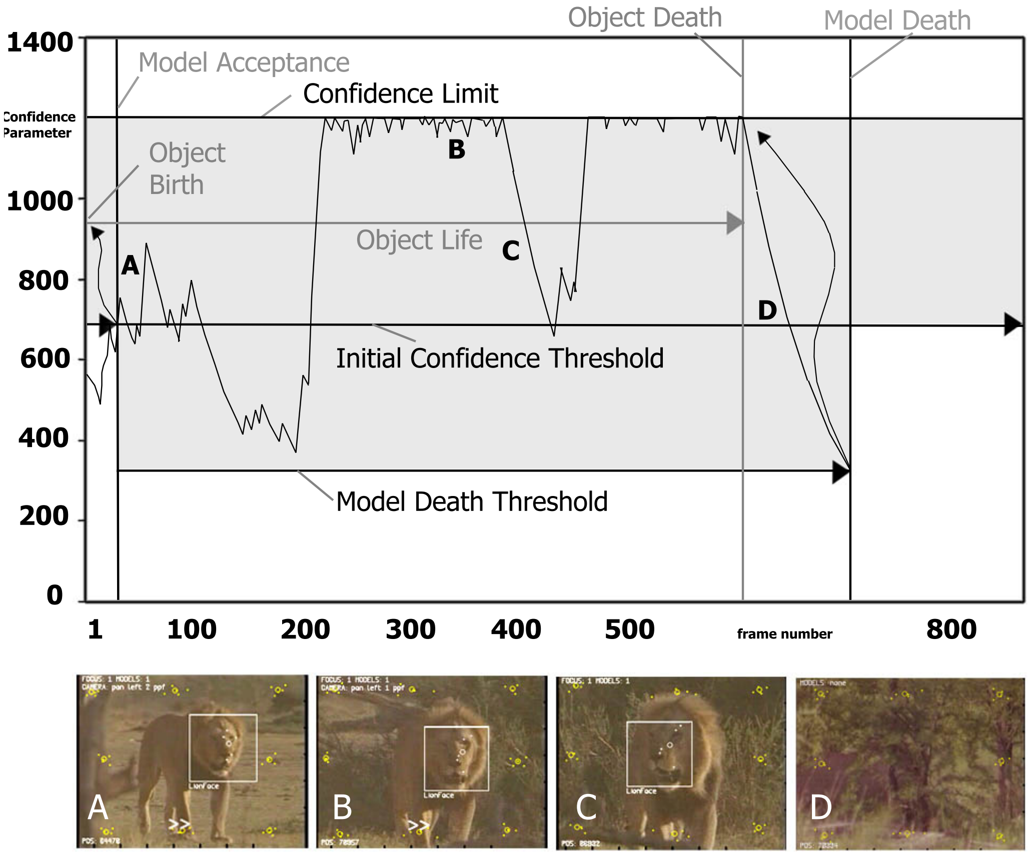


Figure 7: Robust Post-labeling using a Confidence Parameter (TTL-Curve): First occurrence of the lion face (object birth) as well as its disappearance (object death) are confirmed later than the actual incidence; Model acceptance implies object birth and model death implies object death

TCin	TCout	Caption
00:25:32:00	00:26:02:00	lions walking over dry grassland and lying down
00:08:46:00	00:09:15:00	lioness walking towards camera, lies down on ground
00:45:03:00	00:46:31:00	lion & lioness stalking across grass, chasing a gazelle

Table 1: Typical hand-labelled semantic descriptions of wildlife footage

4 Behaviour analysis

The information extracted in the process of detection and tracking is used to generate the semantic description of a wildlife clip with the presence of a given animal specimen and its basic locomotive behaviour, like walking, trotting or standing. Table 1 shows the typical log of manually labelled semantic descriptions of wildlife footage performed by video production professionals. It is noticeable that the main object, an animal in wildlife videos, and information about its behaviour form a core of the clip's semantic description.

The detector gives sufficient information to ***>>> recognise the animal specimen in clips containing frontal or side animal faces.<<<*** Moreover, the extracted trajectories offer a source of information to distinguish between different classes of locomotion. Since the tracked data belongs to a specific, merely rigid body part, e.g. the animal head, the generated trajectory is exceptionally accurate and does not interfere with other motion components at different parts of the animal body. The following method exploits this trajectory information in order to differentiate between locomotion classes.

Generally, we employ two concepts to achieve the classification task. The vertical head frequencies are unique characteristics of certain modes of loco-

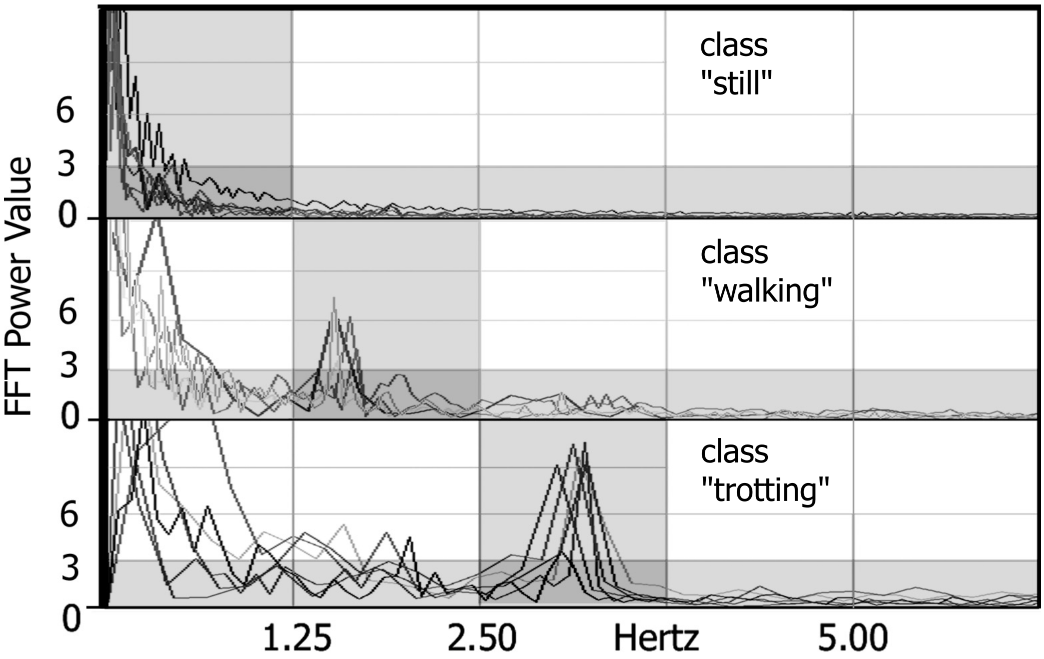


Figure 8: FFT power spectrum of vertical head motion taken from 20 sample clips containing mid distance shots of lions.

motion. The animal's horizontal movement against the background indicates locomotion.

The vertical component of the head trajectory is transferred into the frequency domain using the FFT algorithm, followed by a normalization to compensate for differences in clip length. The power spectrum is then parsed from 5Hz downwards for the first distinctive peak using predefined thresholds. This generates three classes as shown in Figure 8. Clips containing trotting lions show a first distinctive peak in the power spectrum around 3Hz while clips containing walking lions have peak around 1.6Hz. This property remained stable for the majority of investigated clips.

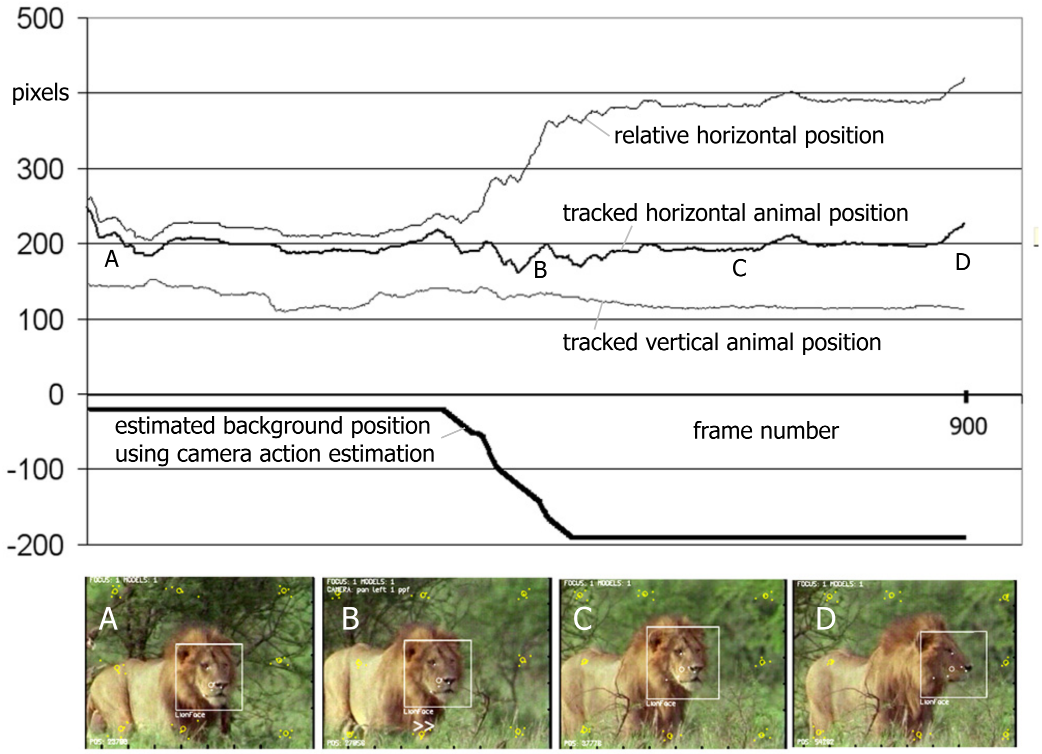


Figure 9: Extraction of horizontal animal motion component employing camera pan example sequence of estimation.

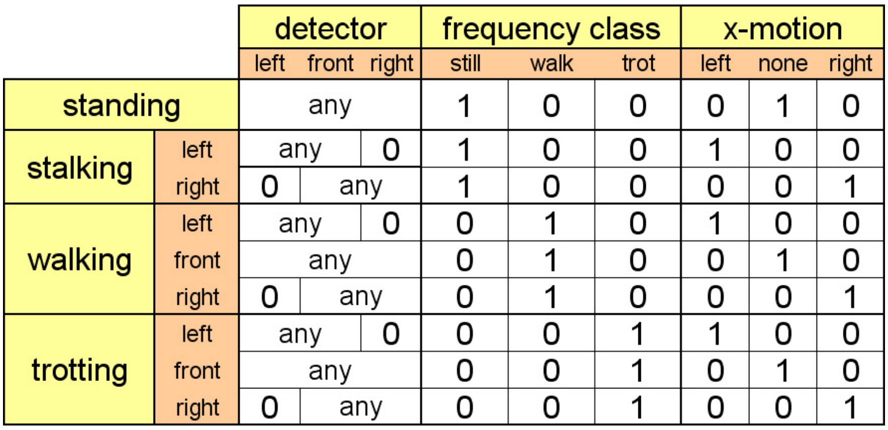
However, stalking lions actively compensate their head motion to hide from potential prey. Under the described classification model the action 'stalking' falls into the class 'still'. In order to separate stalking from standing for non-frontal shots we investigate the horizontal motion of the animal. For the deduction of the horizontal translation component we first clean the trajectory signal from horizontal camera motion by compensating with the estimated camera pan [10]. Figure 9 depicts this process on an example of a stalking sequence.

The horizontal head translation still contains a component that is created by head movements of the animal regardless its locomotion. Since head movements are constrained by the animal's physiology, translations exceeding a certain horizontal expanse are treated as locomotion. The parameters necessary for the establishment of a threshold based on this assumption include the extent of the horizontal motion and the size of the head in the frame approximated by the detection size d of the interest model. The coefficient ρ describes the extend of possible horizontal head movements. A characteristic function $f(x,t)$ for motion in x direction is given in Equation 5. The direction of motion is extracted from the gradient of the low pass filtered horizontal head translation.

$$f(x,t) = \begin{cases} 1, & if \;\; \| \max_{t=1}^{t=N}(x(t)) - \min_{t=1}^{t=N}(x(t)) \| > d \cdot \rho \\ 0, & else \end{cases} \tag{5}$$

Combining the above techniques as depicted in Figure 10 the locomotive behaviour of the detected animal is classified into nine classes of semantic descriptions, namely *standing, walking, stalking and trotting* in three direction: *left, front and right.* The semantic labels are derived from a large semantic classification structure within the ICBR research project [8] in content-based indexing and retrieval of wildlife media.

The experiments are conducted on 25 sample QuickTime MPEG-4 CIF clips from the ICBR media database, with total length of 10min. The results are presented in the Figure 11 in the form of a confusion matrix. The shaded regions mark the unavailable types of locomotive behavior in our experi-



		detector			frequency class			x-motion		
		left	front	right	still	walk	trot	left	none	right
standing		any			1	0	0	0	1	0
stalking	left	any		0	1	0	0	1	0	0
	right	0	any		1	0	0	0	0	1
walking	left	any		0	0	1	0	1	0	0
	front	any			0	1	0	0	1	0
	right	0	any		0	1	0	0	0	1
trotting	left	any		0	0	0	1	1	0	0
	front	any			0	0	1	0	1	0
	right	0	any		0	0	1	0	0	1

Figure 10: Combining outputs of the detection/tracking algorithm and frequency analysis into the predefined set of annotation classes

mental dataset. The misclassified samples are generally due to continuous transition from one locomotive behaviour to another, i.e. starting to trot, walking sideways, very slow stalking, etc. These special cases of transitions and non-uniform behaviour will be in the focus of our future work.

In a wider context, the extracted information can be exploited to establish the supervisory information in the training process of various semantic classifiers. The algorithm generates the semantics such as: existence of the particular animal in the shot; its behaviour; detection of multiple animals and their interrelations. This information boosts the priors in the learning process of an animal model or a classifier. Wildlife video is too complex a domain for applying the paradigm of unsupervised model learning and therefore this method offers a reliable way of narrowing the context of such a complex scenario.

		standing	stalking		walking			trotting		
			left	right	left	front	right	left	front	right
standing		6	0	0	0	0	0	0	0	0
stalking	left	▨								
	right	1	0	3	0	0	0	0	0	0
walking	left	0	0	0	2	1	0	0	0	0
	front	0	0	0	0	3	0	0	0	0
	right	0	0	0	0	0	3	0	0	0
trotting	left	▨								
	front	0	0	0	0	0	0	0	2	0
	right	0	0	0	0	0	1	0	0	3

Figure 11: ***>>>Will be changed - Experimental results given as confusion matrix: annotation labels on the left represent the ground truth<<<***

5 Conclusions

In this paper, we have presented an algorithm for tracking animals in wildlife video footage and classifying their locomotive behaviour into multiple semantic categories. The technique is based upon a face detection algorithm combined with a tracker and uses a novel interest model that enables continuous and smooth animal tracking. The method is illustrated on lions. The face detection method utilises a set of Haar-like features in the AdaBoost classification algorithm. Once detected, face regions are tracked by applying the Kanade-Lucas-Tomasi technique and an interest model is created. By continuous monitoring of detections and model parameters, a rectangular interest model is updated and repositioned to achieve smooth and accurate animal face tracking. The tracking information is utilised in the classification module that assigns a locomotive behaviour to the tracked animal. This high-level information is used to semantically annotate the raw wildlife footage. The focus of future work will be twofold. Firstly, methods that minimise the required number of hand labelled images while maintaining the same level of detection precision will be investigated. Furthermore, on the basis of the tracked region information, a more general model of the detected species and their behavioural patterns will be examined.

Acknowledgements

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.

References

[1] D. Walther, D. R. Edgington, and C. Koch. Detection and tracking of objects in underwater video. *IEEE International Conference on Computer Vision and Pattern Recognition*, 1:544–549, 2004.

[2] N. Haering, R.J. Qian, and M.I. Sezan. A semantic event-detection approach and its application to detecting hunts in wildlife video. *IEEE Transactions on Circuits and Systems for Video Technology*, 10:857–868, 2000.

[3] D. Tweed and A. Calway. Tracking multiple animals in wildlife footage. *16th International Conference on Pattern Recognition*, 2:24–27, 2002.

[4] D. Ramanan and D. A. Forsyth. Using temporal coherence to build models of animals. *9th International Conference on Computer Vision*, 1:338–345, 2003.

[5] M. R. Everingham and A. Zisserman. Automated person identification in video. *3rd International Conference on Image and Video Retrieval*, 1:289–298, 2004.

[6] David Gibson, Neill Campbell, and Barry Thomas. Quadruped gait analysis using sparse motion information. In *International Conference on Image Processing*. IEEE Computer Society, September 2003.

[7] Sion L. Hannuna, Neill W. Campbell, and David P. Gibson. Segmenting quadruped gait patterns from wildlife video. *VIE 2005 - IEE Visual Information Engineering Conference*, 2005.

[8] J. Calic, N. Campbell, M. Mirmehdi, B. Thomas, R. Laborde, S. Porter, and N. Canagarajah. Icbr - multimedia management system for intelligent content based retrieval. In *International Conference on Image and Video Retrieval CIVR 2004*, pages 601–609. Springer LNCS 3115, July 2004.

[9] P. Viola and M. Jones. Robust real-time object detection. *Second International Workshop on Statistical and Computational Theories of Vision*, 2001.

[10] J. Calic, N. Campbell, A. Calway, M. Mirmehdi, T. Burghardt, S. Hannuna, C. Kong, S. Porter, N. Canagarajah, and D. Bull. Towards intelligent content based retrieval of wildlife videos. *WIAMIS 2005 - 6th International Workshop on Image Analysis for Multimedia Interactive Services*, 2005.

[11] T Troscianko, C A Parraga, P G Lovell, D Tolhurst, R Baddeley, and U Leonards. Natural illumination, shadows, and primate colour vision. *ECVP European Conference on Visual Perception*, 2004.

[12] R. Lienhart and J. Maydt. An extended set of haar-like features for rapid object detection. *IEEE International Conference on Image Processing*, 1(1):900–903, 2002.

[13] J. Shi and C. Tomasi. Good features to track. In *IEEE Conference on Computer Vision and Pattern Recognition*, Seattle, June 1994.