Real-time Face Detection and Tracking of Animals

Tilo Burghardt

Department of Computer Science
University of Bristol
United Kingdom

Email: burghardt@cs.bris.ac.uk

Janko Čalić

Department of Computer Science
University of Bristol
United Kingdom

Email: janko@cs.bris.ac.uk

Abstract—This paper presents a real-time method for extracting information about the locomotive activity of animals in wildlife videos by detecting and tracking the animals' faces. As an example application, the system is trained on lions. The underlying detection strategy is based on the concepts used in the Viola-Jones detector [1], an algorithm that was originally used for human face detection utilising Haar-like features and AdaBoost classifiers. Smooth and accurate tracking is achieved by integrating the detection algorithm with a low-level feature tracker. A specific coherence model that dynamically estimates the likelihood of the actual presence of an animal based on temporal confidence accumulation is employed to ensure a reliable and temporally continuous detection/tracking capability. The information generated by the tracker can be used to automatically classify and annotate basic locomotive behaviours in wildlife video repositories.

I. 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. [2] utilise saliency maps to minimize multi-agent tracking of low-contrast translucent targets in underwater footage. Haering et al. [3] 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 [4] develop a periodic model of animal motion and exploit conditional density propagation to track flocks of birds. An interesting approach, by Ramanan and Forsyth [5], takes into account the temporal coherency and builds appearance models of animals. Though dealing only with human faces, the algorithm by Everingham et al. [6] 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. [7] and Hannuna et al. [8] 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. The detection algorithm is an adapted version of a human face detection method that exploits Haar-like features and the AdaBoost classification algorithm [1]. 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. The results show that the tracking information can be 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 presence of the animal, together with its locomotive behaviour, creates a strong prior 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 within the ICBR project [9], [10], that focuses on the computer vision research challenges in the domain of wildlife documentary production. Therefore the information on the existence and behaviour of a specific animal is vital to the process of video media reuse from an large digital video repository.

This paper is organised as follows. In Section II, a method for animal face detection that uses Haar-like features and AdaBoost classifiers is presented. Section III describes the algorithm that combines detection with tracking in a joint interest model. The evaluation results are presented in Section IV, while the final conclusions are given in Section V.

II. ANIMAL FACE DETECTION

In the domain of content-based retrieval of wildlife videos for automatic reuse and repurposing in a digital media pro-

Fig. 1. (*top row*) One of the sample patches, its normalised luminance map, its normalised (green-red) colour-opponent map and the down-sampled map as used for detector training (*bottom row*) The three most characteristic features superimposed on the sample opponent-map and the accumulated overlap density of all 250 features.

Fig. 2. (*left*) Pool of Haar-like feature kernels used to quantify a local contrast configuration. (*right*) Receiver Operating Curve (ROC) of the detector for lion faces.

duction chain, the information on the presence of a particular animal species and the way the animals behave in a given video sequence is essential. In order to achieve the classification of the animal locomotive behaviour, a specific method that initially locates and consequently tracks the animal faces in a given video clip has been developed.

Firstly, in order to measure the image support for the presence of an animal face we have utilised an algorithm for detection and recognition of human faces introduced by Viola and Jones [1]. This algorithm exploits the local contrast configurations of the luminance channel in order to detect the image regions with human faces. The detector is driven by a classifier that is generated by utilising the AdaBoost algorithm. We have applied a modified version of this algorithm, where the contrast features are extracted from a colour opponent map calculated as the difference of the red and green colour channels. This change of the primary input space, as exemplified in Figure 1, strongly improves the robustness of the signal to shadow and illumination changes in *natural scenes* as highlighted by Troscianko et al. [11].

The procedure employs a pool of Haar-like characteristics as

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, a feature f consists of N rectangular components $r(x, y, w, h)$. Each component contributes to f with its average pixel value $S(r)$ weighted by v :

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

Viola and Jones [1] show that each $S(r)$ can be computed in a highly efficient way by using only four accesses to the integral image (II), as defined in Equation 2.

$$S(r) = II(x-1, y-1) + II(x+w-1, y+h-1) - II(x+w-1, y-1) - II(x-1, y+h-1) \quad (2)$$

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

$$II(x, y) = \sum_{j=0}^{x-1} \sum_{i=0}^{y-1} I(i, j) \quad (3)$$

$$II(-1, y) = 0 \wedge II(x, y) = II(x-1, y) + H(x, y), \\ H(x, -1) = 0 \wedge H(x, y) = H(x, y-1) + I(x, y).$$

In order to detect regions with animal faces, the gentle AdaBoost [12] learning algorithm is utilized to compose a classifier from the most characteristic set of the Haar-like features. The classifier is trained on a hand-labelled set of positive and negative sample image patches. As shown in the second row of Figure 1, the strongest classifying features picked by the algorithm are located around the distinctive facial areas, e.g. the nose, the eyes and the jaw. The final "lion face" classifier combines 250 features. The training is done using 680 positive image regions and 1000 negative ones. In the working area (see Figure 2) the false alarm rate of 1 in 10,000 is very low given a hit rate of still 93%.

In comparison to the classifier driven by contrast configurations of the luminance channel, the presented

Fig. 3. Comparison of the detector performance with growing detector complexity using single channel features and shadow-robust colour difference features.

Fig. 4. Examples of lion face detections: images show scale invariance, slight posture invariance and multiple detections of the algorithm

Fig. 5. Examples of feature points stipulated in the center of the detected region for low-level tracking

shadow-robust approach that utilises colour difference features achieves a ten times lower false alarm rate at the same hit rate, as depicted in Figure 3.

An exemplar set of frontal face detections is presented in Figure 4. In order to achieve invariance to different face postures and thus a smooth and robust tracking at later stages, multiple detectors are constructed for the side and profile views following the same approach as for the frontal one.

III. COMBINING DETECTION WITH TRACKING

The Haar-detector performs scale invariant recognition in real-time but nevertheless it is limited to a finite number of trained face postures. The motion prediction for frames having the interim postures using a motion model is very difficult since most animal movements follow a dynamically changing non-linear pattern. Therefore, an algorithm for reconstruction of the animal face trajectory through frames of non-detections has been developed. It is based on tracking of the low-level image features and it is independent from the success of the detection algorithm.

A specific rectangular interest model m_i is established at the image locations where lion faces are first detected by the algorithm described in Section II. The central area of an interest model is then populated with a sparse set of points having the strongest value of the image gradient in both image

directions (see Figure 5), following a suggestion by Shi and Tomasi [13]. The stipulated points are tracked utilizing a pyramidal implementation of the Kanade-Lucas-Tomasi [14] tracker. The single feature points within the cloud of tracked points 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 $\Delta c_i(t) = c_i(t) - c_i(t-1)$ gives the displacement estimate of the interest model m_i . In the case of a nearby face detection $d_i(t)$, the newly estimated position of m_i is corrected towards the location of the detection to avoid model drifting. This process is summarized by Equation 4 and graphically illustrated in Figure 6.

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

The parameter τ controls the adjustment of the interest model m_i and balances the influence between new detections and the tracked feature points. As depicted in Figure 7, the optimal value of the parameter τ is determined experimentally by minimizing the error between the resulting tracked trajectory and the hand labelled ground-truth. Note that the model drifts strongly for very low values of τ , since it relies on the low-level feature tracking only. In contrast, for very high τ , the trajectory is rendered erroneously because of the spatial instability of the Haar-detector that causes jumps of the estimated model position.

The dynamic updating of the feature points prevents the interest model from drifting and achieves robustness to abrupt camera movement, fast animal motion, partial occlusions and insufficient image gradient of particular tracking points. In Figure 10 at the end of the paper, examples of the robustness of the algorithm to partial occlusion, scale and illumination changes as well as slight change of posture are exemplified.

Fig. 6. Visualization of the calculation performed to estimate a subsequent model position by combining the results of detection and tracking.

Fig. 7. Optimization of the mixing parameter τ by experimentally minimising the tracking inaccuracy relative to the ground truth. The selected parameter of $\tau = 4$ defines an optimal, domain specific trade-off between the minimization of model drift and the reduction of spatial jumps.

A. Confidence accumulation

In order to generate continuous and coherent information about the trajectory of the interest model a tracking confidence parameter is proposed. The temporal density of the detections along with the extracted trajectory gives an estimate for the likelihood of the actual presence of an animal face. Decisions about the animal's appearance and disappearance are made based on this likelihood. To calculate the temporal density a *confidence parameter* gets assigned to each interest model m_i .

The confidence parameter is initialized with the first detection. It decreases over time if there are no detections and increases if detections are present. Once the confidence parameter value exceeds the initial acceptance threshold the model is accepted as a valid instance of an animal face. The current frame as well as the adjacent previously tracked frames are labelled positive (see Figure 8). For the accepted models, parts of the trajectory tracked without detections (e.g. caused by occlusions) are validated by the next occurring detector response, as for the examples given in Figure 10.

In comparison to the existing face tracking algorithms, the

proposed technique introduces a specific method of combining two independent mechanisms in order to tackle challenges of such a heterogeneous and broad domain as the wildlife video. The main objective here was to develop a generic algorithm to infer smooth and continuous trajectory of an animal in order to characterise its locomotive behaviour. Therefore, the features utilised in the detector and tracker modules are chosen independently to achieve stronger synergy, i.e. the detector exploits a set of most classifiable features and their spatial distribution, while the tracker uses the best trackable features without any spatial distribution model. The majority of existing tracking algorithms fuse these two paradigms of detection and tracking into a single statistical model applying constraints on temporal and spatial distribution of tracked features [4]. In that case, a manual initialisation is often necessary, which is infeasible in the context of large wildlife archives. The independence of detection and tracking is advantageous in an analysis of wild animals, since their behaviour often results in intervals of distorted feature configurations, hindering detection in that way. Therefore, the independent tracking algorithm relies on temporal continuity of the distorted low-level features to compensate for the missing detections.

However, unlike other algorithms that track deformable objects [6], [15], there is no underlying tracking model of the object in the presented approach. In the context of generic animal tracking for content-based retrieval, it is difficult to generate underlying models for every single species. Therefore, the proposed method doesn't assume any kind of motion model, leaving the possibility of scalable analysis of different species opened.

Fig. 8. The decision on the appearance and disappearance of the animal's face is made a posteriori, based upon the density of the adjacent detections

Fig. 9. FFT power spectrum of vertical head motion taken from a subset of 25 sample clips containing mid distance shots of lions.

IV. RESULTS

In order to evaluate the results of the presented algorithm, the results of animal's locomotive classification, that exploit extracted detection and tracking information, are compared to a manually labelled ground truth. The dataset comprises 12,000 hours of wildlife footage, semantically labelled with information such as the presence of an animal, its species and its locomotive behaviour. Using lions as an example, the following locomotive concepts were used in the evaluation process: walking, trotting and standing. This evaluation method follows the scenario where the presented detection and tracking methods are used to automatically index wildlife footage with the presence and behaviour of particular animal species. The tracking algorithms generates sufficient information to distinguish between different classes of an animal's mode of locomotion. Since the tracked data belongs to a specific, merely rigid body part, e.g. the head of the animal, the generated trajectory is exceptionally accurate and does not interfere with other internal motion components at different parts of the animal's body. Therefore, a method that exploits the head trajectory information only has been developed in order to differentiate between locomotion classes.

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 the 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 9. Clips containing trotting lions generate 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 the investigated clips. The confusion matrix of the classification results is presented in Table I.

The presented algorithm is able to generate semantics such as: existence of the particular animal in the shot; its locomotive behaviour; detection of multiple animals and their interrelations. This information boosts the priors in the learning process of an animal model or a classifier. In such a complex domain as wildlife video, the paradigm

of unsupervised model learning is not feasible. Therefore, the presented methodology offers a reliable solution to the problem of the automatic annotation of raw wildlife footage by supervised learning. By utilising the presented detection and tracking algorithms, the detection of an animal specimen and the classification of its basic locomotive behavior becomes feasible, narrowing the context of semantic annotation in such an incomprehensible scenario.

V. CONCLUSIONS

In this paper, we have presented an algorithm for real-time detection and tracking of animals in wildlife video footage. The technique is based upon a face detection algorithm combined with a tracker and uses a novel interest model that enables continuous and smooth tracking of animals. The face detection method utilises a set of Haar-like features in the AdaBoost classification algorithm. Once detected, face regions are tracked by applying an interest model that tracks low-level features using the Kanade-Lucas-Tomasi algorithm. By continuous monitoring of detections and model parameters, the interest model is updated and repositioned to achieve smooth and accurate animal face tracking.

In order to evaluate the results, the tracking information is utilised in the classification module to assign a locomotive behaviour to the tracked animal. The focus of our future work will be on generalisation of the presented algorithm to a wider context of automatic wildlife video annotation. 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. In addition to the activities in the area of individual animal identification [16], the next stage of our research will be focused towards a hierarchy of detectors that would enable tracking of different animal species driven by a predefined semantic ontology on wildlife.

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 <http://www.3cresearch.co.uk>.

	head-frequency based classification		
	detection <i>still</i>	detection <i>walk</i>	detection <i>trot</i>
<i>still</i>	10	0	0
<i>walking</i>	0	9	0
<i>trotting</i>	0	1	5

TABLE I

EXPERIMENTAL CLASSIFICATION RESULTS GIVEN AS CONFUSION MATRIX: ANNOTATION LABELS ON THE LEFT REPRESENT THE GROUND TRUTH

Fig. 10. Illustrative example of successful tracking with partial occlusion, differences in lighting and slight change of pose.

REFERENCES

- [1] P. Viola and M. Jones. Robust real-time object detection. *Second International Workshop on Statistical and Computational Theories of Vision*, 2001.
- [2] 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.
- [3] 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.
- [4] D. Tweed and A. Calway. Tracking multiple animals in wildlife footage. *16th International Conference on Pattern Recognition*, 2:24–27, 2002.
- [5] D. Ramanan and D. A. Forsyth. Using temporal coherence to build models of animals. *9th International Conference on Computer Vision*, 1:338–345, 2003.
- [6] M. R. Everingham and A. Zisserman. Automated person identification in video. *3rd International Conference on Image and Video Retrieval*, 1:289–298, 2004.
- [7] 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.
- [8] 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.
- [9] J. Calic, N. Campbell, M. Mirmehdi, B. Thomas, R. Laborde, S. Porter, and N. Canagarajah. Icbv - 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.
- [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.
- [14] C. Tomasi and T. Kanade. Detection and tracking of point features. Technical Report CMU-CS-91-132, Carnegie Mellon University, April 1991.
- [15] D. DeCarlo and D. Metaxas. The integration of optical flow and deformable models with applications to human face shape and motion estimation. In *Computer Vision and Pattern Recognition, 1996. Proceedings CVPR '96, 1996 IEEE Computer Society Conference on*, page 231, 1996.
- [16] T. Burghardt, B. T. Thomas, P. J. Barham, and J. Calic. Automated visual recognition of individual african penguins. *5th International Penguin Conference, Ushuaia, Argentina*, September 2004.