Autonomous Monitoring of Cliff Nesting Seabirds using Computer Vision

Patrick Dickinson[1], Robin Freeman[2], Sam Patrick[3] and Shaun Lawson[1]

1 Dept. of Computing and Informatics
University of Lincoln
Lincoln
UK
[pdickinson, slawson]@lincoln.ac.uk

2 Computational Ecology
and Environmental Science Group
Microsoft Research
Cambridge
UK

3 Dept. of Animal and Plant Science
University of Sheffield
Sheffield
UK

Abstract

In this paper we describe a proposed system for automatic visual monitoring of seabird populations. Image sequences of cliff face nesting sites are captured using time-lapse digital photography. We are developing image processing software which is designed to automatically interpret these images, determine the number of birds present, and monitor activity. We focus primarily on the the development of low-level image processing techniques to support this goal. We first describe our existing work in video processing, and show how it is suitable for this problem domain. Image samples from a particular nest site are presented, and used to describe the associated challenges. We conclude by showing how we intend to develop our work to construct a distributed system capable of simultaneously monitoring a number of sites in the same locality.

There is recent evidence that the UK has seen a sustained decline in the numbers of its breeding seabirds. Indeed, the summer of 2008 has seen widespread speculation that the breeding cycle of large numbers of cliff nesting seabirds such as Guillemots, Razorbills and Kittiwakes has failed, with many pairs of birds abandoning nests for long periods of time. In the wider context, any variation in seabird populations is seen as an important biodiversity indicator. Accurate monitoring of seabird populations and aggregate behaviours of birds is therefore of crucial importance to gaining an understanding of the changing environmental picture. The monitoring of seabird populations using manual means (e.g. counting birds on cliff-faces) can be very labour intensive and often impractical: seabirds generally breed in harsh, remote environments which are exposed to the elements, resulting in inconsistent and infrequent data sampling as well as human error.

In this paper we consider the use of computer vision as a method of automatically monitoring seabird populations. Our aim is to develop a system in which the output from a camera can be interpreted by a software system capable of automatically estimating the number of breeding birds on a cliff face, and determining whether an individual breeding bird is present at the nest site at the time of the image capture. This latter inference can be made since individual Guillemots, as

is typical of cliff nesting species, do not stray more than a short distance from their own nest site when present on the cliff face.

The use of computer vision for automated surveillance has attracted significant attention from researchers in recent years, motivated largely by concerns over public safety, and the widespread installation of CCTV cameras in city centres and transportation infrastructure. Much of this work has been concerned with developing low-level image processing techniques which can identify and track moving objects. However, most of these techniques have been developed for use in urban and man-made environments and do not transpose well to video data collected in unconstrained natural environments.

The focus of our work, then, is on the development of processing techniques which can reliably identify and tracking moving birds in natural environments. We start by reviewing some existing techniques, and show how they can be confounded by image artefacts which occur in natural scenes. We then describe some of our own previous work, in which we have sought to address some of these issues. We conclude by reviewing some recently acquired video data from a Guillemot nesting site in Skomer Island, off the coast of Wales. We consider what specific challenges are presented by data collected from this site, and how our work may be further developed to address them.

1 Vision based Surveillance

Data captured from a static video camera comprises a sequence of digitised images, captured at a specific frame-rate: each image consists of a 2D array of intensity or colour values (pixels) sampled from the projection plane of the camera. The ultimate objective of an automated surveillance system is to take this raw image sequence and infer some kind of semantic information about an observed scene. This procedure normally involves several levels or layers of processing, such that each level produces a more abstract and informative representation of the scene. Almost all surveillance systems work in this way, and whilst there is variation in individual processing architectures, the majority follow a common pattern.

1.1 Foreground Segmentation

The first processing stage is often referred to as "background subtraction", or "foreground segmentation", and involves identifying which pixels in an image correspond to moving objects of interest. In this respect, the term "foreground" is used to refer to moving objects (the part of the scene which is changing from one image to the next), and "background" refers to the invariant part of the scene. Naturally, it is the scene foreground which is usually of interest: once identified, foreground pixels are usually clustered to form object hypotheses, which may then be tracked from one image to the next. High level processes, such as behavioural modelling, may also be applied.

The correct classification of foreground pixels is critical to system performance, since any errors will adversely impact on higher level processing. Whilst many methods have been proposed, robust foreground segmentation remains an elusive goal, and has attracted much attention from computer vision researchers.

Most existing approaches to foreground segmentation involve independent modelling and classification of colour values recorded for each image pixel. A background model for a pixel is developed during an initial training phase. During this phase it is normally assumed that the majority of observed values for a pixel are generated by the scene background. Since pixels are classified independently, these are often referred to as "per-pixel" models.

In the simplest cases, exemplified by Hu *et al.* [4], a mean background colour value is learnt for each pixel from the training frames. A subsequent observed pixel value is classified as foreground if it is more than some threshold distance T from the mean. More sophisticated statistical pixel models have been proposed. For example, Park and Aggarwal's scheme [7] uses a Gaussian distribution in HSV colour space to model the background process for each pixel. Elgammal *et al.* [3] used a non-parametric kernel density estimate to describe each pixel background.

1.2 Stauffer and Grimson's Model

The most widely adopted per-pixel model is that proposed by Stauffer and Grimson [8]. This parametric representation of a multi-modal pixel process is more compact, using an adaptive Mixture of Gaussian (MoG) to model observations of each pixel's process in RGB colour space. Under this MoG model the probability of observing a new colour value $\mathbf{x}_{i,t}$ at pixel i and time t is given by:

$$p\left(\mathbf{x}_{i,t}|\mathbf{\Theta}_{i,t}\right) = \sum_{k=1}^{K} \omega_{i,t}^{k} \, \mathcal{N}\left(\boldsymbol{\mu}_{i,t}^{k}, \boldsymbol{\Sigma}_{i,t}^{k}\right) \tag{1}$$

where $\mathcal{N}$ is the multivariate Gaussian probability distribution function, $\omega_{i,t}^{k}$ is the component weight, $\boldsymbol{\mu}_{i,t}^{k}$ the mean, and $\boldsymbol{\Sigma}_{i,t}^{k}$ the covariance matrix of the 3-dimensional colour component k at time t. The standard conditions for MoG models apply, such that the component weights for each pixel sum to one. The number of components is fixed, and typically between 3 and 5 are used.

Components with the highest relative weightings are taken to represent the background. New observed pixel values are compared against the existing components, to determine if the new value is more likely to have been generated by the background or by a foreground object. The model is adaptive, such that gradual changes in lighting can be learnt.

However, a systemic weakness of per-pixel models, including Stauffer and Grimson's, is that independent modelling and classification represents a significant simplification of the actual scene structure. In real-world environments this can lead to frequent misclassifications caused by a variety of artefacts, such as:

1 Colour similarity of overlapping foreground and background objects

2 Sudden lighting changes

3 Object shadows

4 Small movements of background objects

5 Camera movements: the camera is assumed static, but may be moved by wind or other processes

In outdoor urban environments, some of these problems are more prevalent than others. For example, the issues of shadow detection and lighting changes has been well studied [10, 9].

1.3 A Spatially Coherent Model

Some authors, such as Migdal and Grimson [6], have attempted to reduce pixel misclassifications using Markov Random Field (MRF) based background models. This type of model places dependencies between adjacent pixels, such that a pixel is more likely to be classified as background if

its neighbouring pixels are also background. Conversely, a foreground pixel is more likely to be found among other foreground pixels.

MRFs are probabilistic graphical models in which each node represents a random variable and undirected edges between nodes represent dependencies. In the case of foreground segmentation schemes, nodes represent pixel labellings $l_i \in \{foreground, background\}$ and each node is connected to its immediate 4 or 8 pixel neighbourhood. The edges express the Markovian nature of the local node dependency: given its neighbours, each node is conditionally independent of the rest of the field. From a Bayesian perspective, given an observed image $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, and suitable observation likelihood functions, the aim is to estimate the image labelling $\mathbf{l} = \{l_1, \ldots, l_N\}$ which maximises the posterior probability:

$$p(\mathbf{l}|\mathbf{X}) \propto p(\mathbf{X}|\mathbf{l})\, p(\mathbf{l}) \tag{2}$$

in which $p(\mathbf{l})$ represents a prior probability for a given labelling configuration, and is defined by the dependencies between adjacent pixels. $p(\mathbf{X}|\mathbf{l})$ is the likelihood of the labelling determined independently for each pixel: Migdal and Grimson [6] simply use Stauffer and Grimson's per-pixel model to estimate the background likelihood.

Maximum *a posteriori* (MAP) MRF labelling techniques seek to estimate a globally optimum image labelling by maximising the posterior of Equation (2). This may be formulated as an energy minimisation problem, and solved using Gibbs Sampling techniques. The advantage of the random field approach is that spatial dependencies can be easily expressed at pixel level and the MAP estimation results in a segmentation which is globally optimal. Whilst MRF based approaches are demonstrably superior to per-pixel models [2], their robustness is still challenged when large clusters of background pixels appear which have low likelihood under the per-pixel background model. This can arise where background objects move significantly, and may cause additional localised misclassifications.

1.4 Challenges in Natural Environments

As we have mentioned previously, most research in vision based surveillance has been focussed on monitoring people and vehicles in urban environments. Consequently, work has been directed at optimising performance under these conditions. However, relatively little work has sought to apply such systems to the monitoring of wildlife in natural environments, where a quite different range of problems are presented. Application of per-pixel and MRF based techniques to video data captured from typical bird habitats reveals that pixel misclassifications are often caused by natural processes which act to modify the spatial structure of the scene background. For example:

1 Movement of trees and plants caused by wind

2 Movement of water

3 Camera movements caused by wind

In each case, frequent false positive foreground classifications tend to obscure the true target. Currently, little work exists which addresses these problems, though these has been some recent interest in the elimination errors caused by moving foliage [1]. There has also been some recent interest in the use of automated surveillance for the monitoring of birds [5, 11], but the above issues have not yet been robustly addressed.

2 Our Approach to Segmentation

We have recently proposed a method of foreground segmentation which is effective in reducing misclassifications caused by moving trees and water [2]. Rather use a pixel-based background model, we represent an observed scene as an adaptive mixture of Gaussians (MoG) in 5-dimensional feature space. The features used are the 2D spatial coordinates and 3 colour coordinates in Y'UV format. The MoG is constructed with the intention that each component represents a homogeneous region, corresponding to a set of image pixels with similar spatial and colour characteristics.

A pixel observation at time t is represented by a feature vector $\mathbf{x}_t = [x, y, Y, U, V]^T$. The probability distribution function for a model component j is given by:

$$p(\mathbf{x}_t|j) = \frac{e^{-\frac{1}{2}(\mathbf{x}_t - \boldsymbol{\mu}_{j,t})^T(\boldsymbol{\Sigma}_{j,t})^{-1}(\mathbf{x}_t - \boldsymbol{\mu}_{j,t})}}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}_{j,t}|}} \tag{3}$$

where the parameters $\boldsymbol{\mu}_{j,t}$ and $\boldsymbol{\Sigma}_{j,t}$ are the mean and covariance matrix of the j^{th} component at time t, and the dimensionality D is 5. For K components the general mixture model conditions hold, such that the component weights sum to unity. The number of components, K, required to model the whole scene is estimated automatically from frame to frame, but normally we expect several hundred.

Our premise is that the components of our model represent different background processes, such as different objects, or parts of objects in the scene. Each such process generates a subset of the pixels in an observed image. Thus, given a new observed image and a set of model parameters, we may determine which component is most likely to have generated a particular pixel $\mathbf{x}_t$. We thus seek to assign an observed pixel value to the component with the highest log likelihoods:

$$C_{map} = argmax_j \{log(p(\mathbf{x}_t|j)) + log(\omega_{j,t})\} \tag{4}$$

Model components represent both background and foreground regions of the scene, and each is explicitly labeled as $l_c \in \{foreground, background\}$. Pixels are implicitly labeled according to the component to which they were assigned using Equation (4). Thus a pixel's classification is determined by the global characteristic of the component to which it is assigned, rather than it's own local characteristic. Within this framework we are able to specify the notions of scene foreground and background, and differentiate between them, at object level rather than pixel level.

All model components are updated by the statistics of their assigned pixels. In general, background components are updated more slowly than foreground components, reflecting the expectation that foreground will exhibit more dynamic behaviour. The initialisation, assignments, and update procedures are described in more detail in the remainder of this section.

Our approach expresses the spatial structure of the observed scene in the background model, and helps to reduce some types of misclassification by generating a spatially coherent segmentation. Further details are given in [2].

3 Segmentation of Water Birds

Our experiments [2] have shown that our background model is robust in scenes characterised by moving foliage and water. In contrast, we have also shown that the performance of both per-pixel and MRF-based models deteriorates drastically under the same conditions. As an example, we have compared our algorithm with Stauffer and Grimson's model, and Migdal and Grimson's model, on

a sequence showing a duck approaching water: example segmentations for a single image in this sequence are shown in Figure 1). The output from our algorithm is consistently better, showing considerably less false positive foreground pixels than either of the other two algorithms. Both of the other algorithms show high numbers of incorrectly labelled foreground pixels, likely to result in the incorrect identification of foreground objects. From these results, and others, we conclude that our algorithm is better adapted for use in natural bird habitats, where such processes are likely to occur.

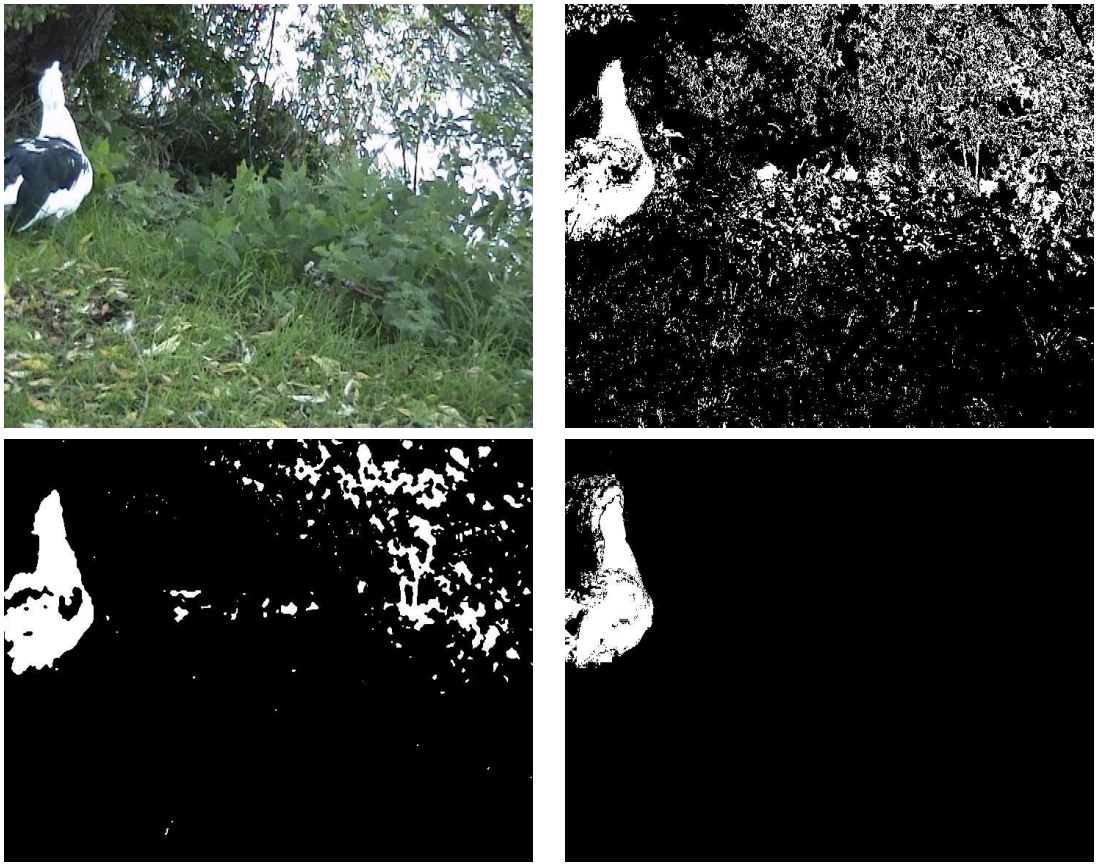

Figure 1: Foreground pixels classified in an image sequence showing a duck. From top to bottom, left to right: original image, using a per-pixel model, using an MRF model, using our model.

4 Planned Development of Our Work: The Skomer Island Colony

Our current interest is in video foreground segmentation to automatically analyse image sequences captured from a Guillemot nesting site on Skomer Island, North wales. Our intention is to automatically determine the number of birds on a series of cliff faces, and to monitor the condition of specific nest sites. We are intending to utilise existing hardware, which has been set up to facilitate manual monitoring of the site. This includes digital cameras, and a wireless network.

As a starting point, we have acquired image sequences of a particular cliff area, captured using a digital camera operated in a time-lapse capture mode. This sequence presents a significantly

different data profile than that normally encountered when using video capture equipment: firstly, the images are captured at 60-75 seconds apart, and secondly they are of a relatively high resolution (3072 x 2304 pixels). Example images are shown in Figure 2. It should be noted that since these images are captured at a high resolution, low-level details are not visible in Figure 2. An enlarged section of an image is shown in Figure 3.

Figure 2: Sample images captures captured from the Skomer breeding site.

The images are also characterised by two important artefacts: significant camera shake, and the presence of moving water, both of which are likely to confound standard segmentation techniques. Our method (described in section 2) is more suitable for processing under these conditions; however, capture rate, resolution, and other features of the data necessitate some further development. We now describe modifications that we intend to make to our algorithm, and a proposed architecture by which we might collectively monitor several sites on the island.

4.1 Proposed Modifications to Our Algorithm

Our algorithm (like Migdal and Grimson's) is computationally intensive, which makes real-time procesing at video frames rates (e.g. 25Hz) difficult, even at lower image resolution. To some extent this is exacerbated by the increased resolution of the Skomer images: we anticipate that segmentation of an individual image could take up to one minute on typical PC hardware. However, unlike video data, the time-lapse data is captured much more infrequently. Not only does this

Figure 3: Enlarged section of an image from the Skomer breeding site.

facilitate our projected processing requirements, but potentially allows us in incorporate additional functionality to further improve quality. We could, for example, capture images every 5 minutes.

The time between image frames does introduce one additional problem, however. Our segmentation method uses models of both the foreground and background to classify pixels. The foreground model relies on a relatively high frame rate, and assumes that the movement of individual objects is relatively small between successive images. We can no longer make this assumption with images captured minutes apart. Instead we propose to dispense with the foreground model (instead assuming a uniform colour distribution across the foreground), and develop a more sophisticated background model.

In our experiments ([2]) we have identified that our model is effective at eliminating pixel mis-classifications caused by small background movements. However, the MRF model is also effective in enforcing spatial coherence. We thus propose to develop our model by combining our background representation (spatially distributed Gaussians) with an MRF framework. Each image pixel will then be classified by assignment to one of a number of possible labels: either the foreground, or to a component which is spatially local. The observation likelihood, $p(\mathbf{X}|\mathbf{l})$ in Equation 2, for background labels will then be given by the colour distribution of those local Gaussian components. We believe that this will facilitate an effective segmentation using images which are captured minutes apart, and retain the robustness of our model to variations in the background.

4.2 Initialisation

A remaining problem is that of model initialisation. Surveillance systems are often initialised under controlled conditions; however, this is not possible at the Skomer site. Consequently, we intend to construct a background model over a longer period of time, perhaps several hours or days, before

monitoring becomes active. Further investigation is required to determine a suitable initialisation period, which will depend largely on bird mobility. Lighting changes will also need to be accounted for when using an extended initialisation period.

4.3 Monitoring Functions

Our proposed system is intended to provide two functions: estimation of bird populations over a cliff face area, and the monitoring of parent behaviour at specific nest sites. We anticipate that global monitoring can be effected completely autonomously. However, monitoring of individual nest sites may require the user to identify those sites manually in the image scene. It may be possible to infer the position of nest sites automatically from parent behaviour; however, this remains an open question.

4.4 Hardware Architecture

Our current system is developed as a single stand-alone monitoring device, comprising a single camera. However, we wish to simultaneously monitor several cliff face scenes at the Skomer site, using several cameras, and provide an aggregated set of results in real-time. To this end, our aim is to develop a distributed system comprising a set of cameras connected connected across a local network, and single PC at which the results are made available. There is an existing wireless network at the Skomer site which we can use to faciltate this.

The high bandwidth associated with transmitting video from multiple sources over a local network is often prohibitive. For this reason, a common approach is to operate low-level processing on-camera (using a "smart camera"), and to transmit the results only. In this case, however, each camera will generate an image only every few minutes. This means that, with relatively low bandwidth, it is feasible to transmit complete images and perform all processing centrally. This significantly simplifies implementation, subject to the processing resources available to the central processing unit. Based on our estimates of one minute per frame, and a frame capture interval of 5 minutes, we might expect to be able to process 5 camera outputs on a single central machine in real-time.

4.5 Summary

We have described a proposed method for using visual surveillance techniques to automatically a monitor a seabird population. In particular, we have considered the application of such techniques to a specific site on Skomer island, populated by Guillemots. We have described our previous work in low-level video image processing, and shown how we have successfully addressed specific problems associated with surveillance in natural environments.

Examination of images collected from Skomer island have revealed a number of challenges associated with both image features, and the existing installed hardware. We have described how we intend to develop our existing work to support a distributed surveillance system capable of providing automated monitoring at this site.

References

[1] G. Dalley, J. Migdal, and W. Grimson. Background subtraction for temporally irregular dynamic textures. In *Proc. of IEEE Workshop on Applications of Computer Vision*, Copper Mountain, CO, USA, 2008.

[2] P. Dickinson. *Monitoring the Vulnerable using Automated Visual Surveillance*. PhD thesis, University of Lincoln, UK, 2008.

[3] A. Elgammal, R. Duraiswami, D. Harwood, and L. Davis. Background and foreground modeling using nonparametric kernel density estimation for visual surveillance. *Proceedings of the IEEE*, 90(7):1151–1163, July 2002.

[4] W. Hu, X. Xiao, Z. Fu, D. Xie, T. Tan, and S. Maybank. A system for learning statistical motion patterns. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 28(9):1450–1464, 2006.

[5] A. Kembhavi, R. Farrell, Y. Lou, D. Jacobs, R. Duraiswami, and L. Davis. Traking down under: following the Satin Bowerbird. In *Proc. of IEEE Workshop on Applications of Computer Vision*, Copper Mountain, CO, USA, 2008.

[6] J. Migdal and W. Grimson. Background subtraction using Markov thresholds. In *Proc. of IEEE Workshop on Applications of Computer Vision / IEEE Workshop on Motion and Video Computing*, volume 2, pages 58–65, Breckenridge, CO, USA, January 2005.

[7] S. Park and J. Aggarwal. Segmentation and tracking of interacting human body parts under occlusion and shadowing. In *Proc. of IEEE Workshop on Motion and Video Computing*, pages 105–111, Orlando, FL, USA, 2002.

[8] C. Stauffer and W.Grimson. Adaptive background mixture models for real-time tracking. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 246–252, Fort Collins, CO, USA, 1999.

[9] Y. Tian, M. Lu, and A. Hampapur. Robust and efficient foreground analysis for real-time video surveillance. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 1182–1187, San Diego, CA, USA, 2005.

[10] M. Xu and T. Ellis. Illumination-invariant motion detection using colour mixture models. In *Proc. of British Machine Vision Conference*, pages 163–172, Manchester, UK, 2001.

[11] K. Younis, B. Aldawood, G. Abandah, S. Alsisan, and J. Igual. Automatic wireless monitoring of bird behavior. In *Proc. of IASTED International Conference on Signal and Image Processing*, pages 511–516, Hawaii, USA, 2008.