http://dx.doi.org/10.1109/ICIP.2018.8451476

PROFILE HIDDEN MARKOV MODELS FOR FOREGROUND OBJECT MODELLING

Ioannis Kazantzidis, Francisco Florez-Revuelta, Jean-Christophe Nebel

Kingston University, London, UK
{ikazant, j.nebel}@kingston.ac.uk
University of Alicante, Alicante, Spain
francisco.florez@ua.es

ABSTRACT

Accurate background/foreground segmentation is a preliminary process essential to most visual surveillance applications. With the increasing use of freely moving cameras, strategies have been proposed to refine initial segmentation. In this paper, it is proposed to exploit the Vide-omics paradigm, and Profile Hidden Markov Models in particular, to create a new type of object descriptors relying on spatiotemporal information. Performance of the proposed methodology has been evaluated using a standard dataset of videos captured by moving cameras. Results show that usage of the proposed object descriptors allows better foreground extraction than standard approaches.

***Index Terms*—** Computer vision, Visual Surveillance, Foreground detection, Freely Moving Cameras, Vide-omics

1. INTRODUCTION

Visual surveillance often relies on a preliminary process which aims at extracting foreground objects from a video. Then, higher level computer vision tasks can be performed such as object recognition, pedestrian tracking or human action recognition. Accurate background/foreground segmentation requires tackling real life conditions including illumination changes, presence of shadows, image noise and camera jitter. Hundreds of solutions have already been offered for scenarios involving static cameras [5]; however, with the widespread use of action and smartphone cameras, approaches are particularly needed to deal with freely moving cameras. Due to the complexity of the task, foreground object modelling has been developed as a strategy to refine initial foreground extraction [18].

In order to address some of the challenges encountered by visual surveillance systems, including camera motion, a novel video analysis paradigm, ‘vide-omics’, has recently been proposed [14]. Inspired by the principles of genomics, this paradigm interprets videos as sets of temporal measurements of a scene in constant evolution without setting any constraint in terms of camera motion, object behaviour or scene structure. This puts variability at the core of every algorithm where the interpretation of scene mutations corresponds to video analysis.

Motivated by the potential of ‘vide-omics’ and its background/foreground segmentation implementation, it is proposed to enhance initial foreground extraction by generating foreground models using probabilistic models called Profile Hidden Markov Models (P-HMMs), which have proved extremely successful to annotate unknown biosequences, i.e. DNA, RNA or protein sequences [10].

Foreground extraction refinement methods. Foreground enhancement methods usually rely upon building foreground/background appearance models at a frame or video level which are then used to label pixels in a region of interest. Frame level methods address foreground refinement as an image segmentation problem, where seeds, often selected interactively, are exploited to either grow regions or initialise some graph-based energy minimisation techniques. GrowCut uses cellular automata to simulate the biological process of bacteria growth [22]. Growth occurs in predefined regions where pixels are labelled according to their neighbours’ values. Consequently, performance relies on accurate seed pixel labelling. To address this, [16] propagates initial labelling through regions with colour homogeneity estimated using [2]. Since initial labelling is sparse, long-distance label propagation may become challenging, thus, different sets of homogeneous regions are calculated, from coarse to fine, to allow labelling to propagate spatially and intra-level from finer to coarser levels. However, such propagation has limitations when dealing with small regions with insufficient annotations. Alternatively, graph-based methods operate on pixel networks or *cliques* where an energy function is minimised by rewarding labels matching an appearance model and encouraging similar labelling among neighbouring pixels. Markov Random Fields (MRFs) are exploited where an energy function is minimised using graph cuts [6]. However, since inference for a pixel in MRF depends on estimating the underlying distribution of cliques, modelling arbitrary pixel dependencies is difficult. This is addressed using Conditional Random Fields (CRFs). Since they do not rely on any underlying pixel distribution, any pixel dependency can be modelled, allowing more efficient image segmentation. Using fully-connected CRFs, all possible pairwise pixel dependencies in an image are considered [15]. A limitation of frame-based methods is that moving objects which are initially static are not identified as foreground. To deal with this, video-based methods were proposed taking advantage of information at video level. [17] extends the approach proposed by [16] by exploiting appearance models estimated using Gaussian Mixture

Models to propagate iteratively labels both in spatial and temporal neighbourhoods. Although producing better performance, interacting objects remain challenging. Moreover, foreground labelling may leak to background regions. To overcome these issues, [21] introduces a deep learning approach for learning a single generic appearance model or *visual-memory* module. It is represented by a Convolutional Gate Recurrent Unit [4], trained over a set of videos using a combination of features, i.e. high level scene representations [8] and motion likelihood maps. Enhanced foreground segmentation for a given frame is retrieved by parsing those features through the appearance model. Despite improved performance, this method cannot handle long video sequences due to encoding capacity of the *visual-memory* module.

Models for biosequence families. Similarly to the problem of identifying an object from some appearance model, one of the most important applications in genomics is the annotation of biosequences by comparing them to sequences of known functions. While there are efficient tools such as Blast [1] which offers fast database search by first finding identical fragments between two sequences and then extending them iteratively by allowing a variety of mutations maximising a similarity score, those tools lack sensitivity. A more powerful approach is to build sequence family models which are able to represent the diversity encountered within a family. HMMER is a database search method based on P-HMMs [9]. They produce sequence family *profiles* by identifying correspondences between the characters of all sequences of a given family. Then, each profile is used to define both the structure and the parameters of an HMM aiming at sequence alignment. HMMER outperforms Blast in terms of sensitivity, however at the cost of increased processing time. Since then, new versions of HMMER have been released using an acceleration pipeline which reduces the sequence search space [11]: while maintaining excellent sensitivity, speed dramatically improved and is now comparable to Blast's.

In this work, it is proposed to generate appearance models at video-level. Treating foreground information as pixel sequences and employing bioinformatics-based techniques allows not only the creation of rich spatio-temporal appearance models as well as the use of well-established techniques for sequence annotation.

2. METHODOLOGY

Using a foreground extraction approach tuned to produce a low rate of false positives (FP), the aim of a refinement algorithm is to increase the number of true positives while maintaining the FP number low. The proposed methodology is illustrated in Fig. 1. Following the 'vide-omics' paradigm, where images are processed at the scanline level, the initial foreground is divided into horizontal segments. They are then clustered to produce a set of models (P-HMM) representing different objects or object elements. Note that, in the rest of this paper, without loss of generality, the word 'object' is used instead of 'object or object element'. Finally, those models are applied to scan regions of interest in order to detect additional foreground segments.

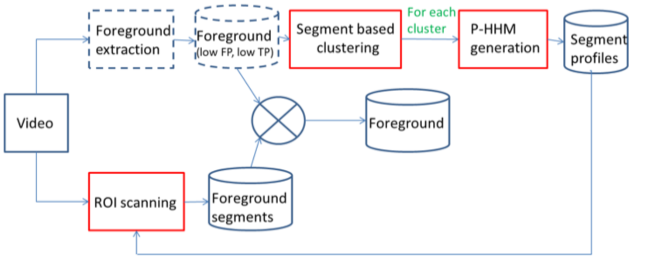

Fig. 1. Description of segmentation pipeline.

2.1. Profile hidden Markov models

In genomics, phylogenetic trees allow deducing evolutionary relationships among various biosequences which are assumed to descend from a common ancestor. They also make possible to infer the sequence of that common ancestor. By using it as a reference, all existing correspondences between characters from the different biosequences can be represented in a multiple alignment, highlighting the mutations that the common ancestor's sequence has undergone to produce each of those biosequences. That multiple alignment can be used to produce a statistical profile modelling the biosequence family. In visual surveillance, the picture of an object captured on a given frame can be interpreted as the product of mutations applied to a canonical object, its common ancestor. As a consequence, building a phylogenetic tree of all available object pictures would allow the generation of a model of that object family.

It is proposed to adapt the process suggested by [9] to generate P-HMMs to represent those object families. First, all foreground segments are clustered using sequence pairwise similarity scores to identify consistent groups of foreground segments. Each of them is used to generate object models. Note that a given object can be represented by several models, which may, for example, encode different views/configuration of that object; also segments may not be allocated to any group. The multiple alignment associated to each group corresponds to an object profile (Fig. 2), which provides the necessary information to generate the probability matrices associated to HMMs.

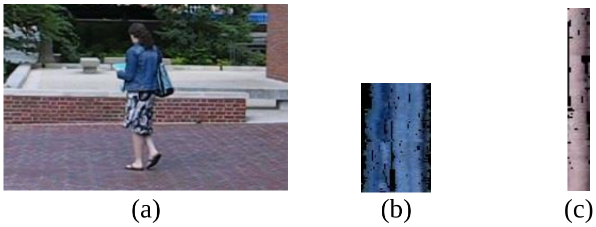

Fig. 2. Examples of object profiles. a) View of the moving individual in *people1*, b) denim jacket and c) leg profiles. Black pixels denote alignment gaps created by insertions.

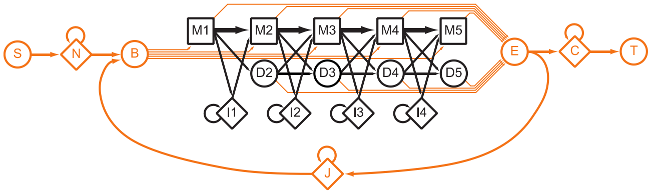

Fig. 3. P-HMM architecture [11], consisting of 5 match states. Diamonds, circles and squares represent the states of the HMM, while arrows indicate possible state transitions.

As proposed in HMMER2 [9], the "Plan 7" profile HMM architecture is used (Fig. 3) to infer the *hidden state path, π,* that corresponds to the observed sequence, *x*. It consists of a finite set of states *Q = {M, I, D, N, J, C, B, E, S, T}*, a matrix of *state transition probabilities* and a matrix of *emission probabilities* for every match and insertion state. While match and insertion states *(M, I)*, also known as emitting states, express the likelihood of a pixel to be aligned with that state, *D* states represent the deletion of a consensus state. Furthermore, special states are used to denote a) the start, *S*, or the termination, *T*, of *π*, b) the begin, *B*, and end, *E*, of a matching region with the profile, and c) non-aligned pixels with the P-HMM *(N, J, C)*. The number of *consensus* or *match* states, the emission probabilities for each *match* or *insertion* state and the transition probabilities denoted with black arrows (Fig. 3) are defined by the profile, while the transition probabilities denoted with orange arrows are algorithm-dependent [10].

2.2. Frame scanning

Once P-HMMs are available, they are used to scan video frames to detect object segments. When scanning a given image scanline, the probability for each of its pixels to be aligned with an insertion or match state of a given profile is calculated. Consecutive pixels with high probability are segmented and further evaluated. Eventually, if a hit is confirmed, pixels aligned with match or insertion states of the P-HMM are classified as foreground pixels.

In order to calculate the alignment probabilities for each pixel, posterior decoding is used because, unlike Viterbi, it allows the backward flow of information to influence the likelihood of each state at any position *i* [3]. First, each scanline, *x*, of length *n* is compared against each profile using the *forward/backward* algorithm [3] so that *posterior* probabilities, $P(\pi_i=k|x)$, are obtained for every pixel, *i*, and state, *k*. *Forward* probabilities, $f_k(i)$, encapsulate the total probability of observing the *i* first pixels of the sequence *x* being in state *k*. Similarly, *backward* probabilities, $b_k(i)$, capture the total probability of observing the *n-i* last pixels being in state *k*. Both *forward* and *backward* algorithms are based on dynamic programming where, for each state *k*, a matrix is filled. Each cell of the matrix can be calculated recursively using the formulas:

$$f_k(i) = e_k(x_i) \sum_l f_l(i-1)\, a_{lk} \tag{1}$$

$$b_k(i) = e_k(x_{i+1}) \sum_l f_l(i+1)\, a_{lk} \tag{2}$$

where *l*, e_k and a_{lk} are, respectively, a given state, the emission probability of state *k*, and the transition probability from state *l* to state *k*.

Finally, after calculating *p(x)*, i.e. the total probability of the observation sequence *x*, the *posterior* probabilities for each pixel *i* and state *k* are estimated as:

$$P(\pi_i = k|x) = \frac{f_k(i)b_k(i)}{p(x)} \tag{3}$$

Once *posterior* probabilities are available for every pixel and state, scanlines are scanned to identify parts that match with the profile. Since a scanline may contain multiple regions matching a given profile, it is scanned sequentially to identify these regions. Once the beginning of a region is found, i.e. a pixel with match posterior probability above a threshold, *t1*, that region is extended until a pixel is found with an end posterior probability decreasing below another threshold, *t2*. In case pixels of a region are matched several times by a given profile, it is further examined using *stochastic clustering* to divide that region in non-overlapping sub-regions [11]. Finally, each identified region whose forward score is positive is locally aligned with its matching profile: pixels corresponding to either match or insertion states are then labelled as foreground pixels.

3. EXPERIMENTS

3.1. Experimental setup

The proposed method was evaluated on the Berkeley Motion Segmentation Dataset (BMS-26) [7], which is a widely used benchmark for motion segmentation. Similarly to [14], twelve videos with moving cameras were selected: *people2*, *cars1-10* (PTZ motion) and *marple10* (freely moving camera). The foreground outputs produced by the vide-omics inspired algorithm for foreground extraction [14] were used as reference, since that approach has a low false positive rate. The performance of the proposed foreground enhancement method was compared using the F1 score against the GrowCut algorithm (GC) [22], a variational method (Ochs) [16] and a CRF based method for image segmentation (CRF) [15].

In order to create the P-HMMs, the UPGMA algorithm was selected to produce the phylogenetic trees [20], where the inter-segment distances were calculated using pair-wise sequence alignment [14] and then normalised [12]. Note that to ensure foreground segments are discriminative enough, very short ones, i.e. shorter than 5% of a frame's width, were not considered. Groups, the intra-group similarity of which were above 35% and had more than 16 members, were judged as being suitable to create P-HMMs. Finally, profiles with high gap frequency, i.e. above 50%, were rejected. For the tested videos, the number of profiles varied between 12 for *car6* - 30 frames capturing a unique object with linear motion - and 275 for *marple10* - 460 frames showing 3 objects with complex motions.

During profile construction, the emission probability of a pixel i in the state k, $e_k(x_i)$, was modelled employing Kernel Density Estimation [19] using a multivariate Gaussian kernel K and a 3x3 bandwidth matrix, H.

$$e_k(x_i) = \frac{256^3}{n}\sum_{j=1}^{n}|H|^{-1/2}K\left(H^{-1/2}\left(x_i - C_j\right)\right) \tag{5}$$

where C is an array containing the RGB pixel values of a match or insertion state, and n is the number of pixels in C. Diagonal elements of H are set to 1 as it is the integer maximising the F1 score within the range [1,10] in the *people1* video [7].

In order to reduce both scanning costs and the number of false positives, profiles were only applied to regions of interests at proximity of the initially extracted foreground. Regions of interests were defined for each frame as the initial foreground mask dilated by a disk structure element of size 1% of the frame's width. During region identification, a value of 25% posterior probability is used for threshold *t1* and a reduction of at least 10% of the posterior probability is set for *t2*. Those horizontal regions were locally aligned employing maximum expected accuracy (MEA) algorithm [13]. Eventually, foreground pixels identified by P-HMM search were added to the initial foreground mask. Finally, as in [14], some post-processing was applied: small regions (area lower than the square of 1% of the frame's width) were removed.

3.2. Results

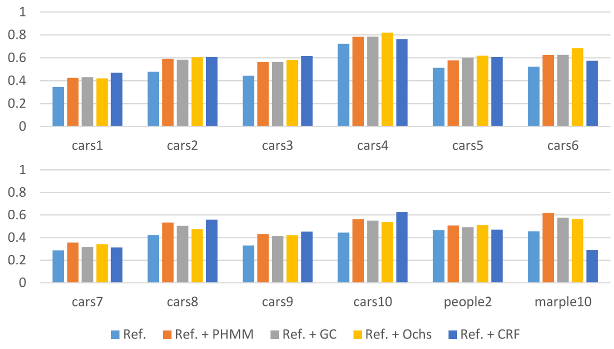

Fig. 4. F1 scores calculated for all videos and methods.

The proposed method has been evaluated using F1 scores on 12 sequences and compared with 3 other refinement approaches (see Fig. 4). Since sequences have different numbers of ground truth frames, varying between 3 and 15, F1 scores are presented as means of both video and frame F1 scores (see Table 1). As expected, refinement methods generally outperform significantly (~+20%) the approach used as reference. Considering results on a video basis, both the proposed method and Ochs exhibit better performance (54.8%) than enhancements produced by either GC (53.7%) or CRF (52.9%). However, when performance is evaluated on a frame basis, the proposed method is significantly better, 59.1%, than all other approaches. Example of refinement by the proposed method is shown in Fig. 5.

The very competitive results obtained with the proposed foreground enhancement method shows that appearance models built at a video level provide richer and more complete object representations. Furthermore, segment based models offers higher specificity. Finally, this work supports the value of the 'vide-omics' paradigm [14].

Table 1. F1 scores of foreground refinement methods. They are calculated as a mean of either video or frame F1 scores.

	Mean F1 (video based)	**Mean F1 (frame based)**
Reference	0.453	0.461
Reference + Proposed	**0.548**	**0.591**
Reference + GC	0.537	0.564
Reference + Ochs	**0.548**	0.563
Reference + CRF	0.529	0.409

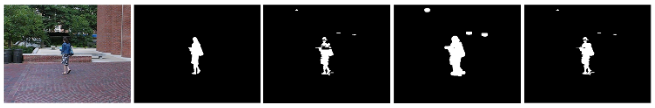
(a) (b) (c) (d) (e)

Fig. 5. Example of foreground refinement by the proposed method for the video *people1*: a) frame, b) ground truth, c) initial, d) dilated mask and e) refined foregrounds.

4. CONCLUSIONS

This paper introduces a new method to enhance foreground object detection by exploiting the Vide-omics paradigm. Its main contribution is a pipeline for generating novel object descriptors and detecting associated objects within a video. These descriptors, which encapsulate spatiotemporal information, are constructed automatically by, first, creating object profiles based on phylogenetic and biosequence analysis, and, second, generating HMMs defined by those profiles. Evaluation performed on a standard video dataset comprising a variety of scenes and camera motions demonstrates the added value of the proposed methodology. Future work will be focused on using those object descriptors to directly detect foreground objects in unseen frames. Such step will require developing a statistical framework able to estimate the significance of matches.

5. REFERENCES

[1] Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology,* vol. 215, no. 3, pp. 403-410, 1990.

[2] Arbelaez, P., M. Maire, C. Fowlkes, and J. Malik, "Contour Detection and Hierarchical Image Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 33, no. 5, pp. 898-916, 2011.

[3] Bahl, L., J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate," *IEEE Transactions on Information Theory,* vol. 20, no. 2, pp. 284-287, 1974.

[4] Ballas, N., L. Yao, C. Pal, and A. C. Courville, "Delving Deeper into Convolutional Networks for Learning Video Representations," *International Conference on Learning Representations,* 2015.

[5] Bouwmans, T., "Recent Advanced Statistical Background Modeling for Foreground Detection - A Systematic Survey," *Recent Patents on Computer Science,* vol. 4, no. 3, pp. 147-176, 2011.

[6] Boykov, Y. and G. Funka-Lea, "Graph Cuts and Efficient N-D Image Segmentation," *International Journal of Computer Vision,* vol. 70, no. 2, pp. 109-131, 2006.

[7] Brox, T. and J. Malik, "Object Segmentation by Long Term Analysis of Point Trajectories," in *European Conference on Computer Vision*, 2010, pp. 282-295.

[8] Chen, L.-C., G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs," *International Conference on Learning Representations,* 2014.

[9] Eddy, S. R., "Profile hidden Markov models," *Bioinformatics,* vol. 14, no. 9, pp. 755-763, 1998.

[10] Eddy, S. R., "A Probabilistic Model of Local Sequence Alignment That Simplifies Statistical Significance Estimation," *PLOS Computational Biology,* vol. 4, no. 5, pp. 1-14, 2008.

[11] Eddy, S. R., "Accelerated Profile HMM Searches," *PLOS Computational Biology,* vol. 7, no. 10, pp. 1-16, 2011.

[12] Feng, D.-F. and R. F. Doolittle, "Progressive sequence alignment as a prerequisitetto correct phylogenetic trees," *Journal of Molecular Evolution,* vol. 25, no. 4, pp. 351-360, 1987.

[13] Käll, L., A. Krogh, and E. L. L. Sonnhammer, "An HMM posterior decoder for sequence feature prediction that includes homology information," *Bioinformatics,* vol. 21 (Suppl. 1), pp. 251-257, 2005.

[14] Kazantzidis, I., F. Florez-Revuelta, M. Dequidt, N. Hill, and J. C. Nebel, "Vide-omics: A genomics-inspired paradigm for video analysis," *Computer Vision and Image Understanding,* vol. 166, pp. 28-40, 2018.

[15] Krähenbühl, P. and V. Koltun, "Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials," presented at the NIPS - Neural Information Processing Systems, 2011.

[16] Ochs, P. and T. Brox, "Object segmentation in video: A hierarchical variational approach for turning point trajectories into dense regions," in *IEEE International Conference on Computer Vision*, 2011, pp. 1583-1590.

[17] Papazoglou, A. and V. Ferrari, "Fast Object Segmentation in Unconstrained Video," in *IEEE International Conference on Computer Vision*, 2013, pp. 1777-1784.

[18] Rother, C., V. Kolmogorov, and A. Blake, "'GrabCut': interactive foreground extraction using iterated graph cuts," *ACM Transactions on Graphics,* vol. 23, no. 3, pp. 309-314, 2004.

[19] Silverman, B. W., *Density estimation for statistics and data analysis*. London; New York, 1986.

[20] Sokal, R. R. and C. D. Michener, "A statistical method for evaluating systematic relationships," *University of Kansas Science Bulletin,* vol. 38, pp. 1409-1438, 1958.

[21] Tokmakov, P., K. Alahari, and C. Schmid, "Learning Video Object Segmentation with Visual Memory," presented at the IEEE International Conference on Computer Vision, 2017.

[22] Vezhnevets, V. and V. Konushin, "'GrowCut' - Interactive Multi-Label N-D Image Segmentation By Cellular Automata," in *GraphiCon*, 2005.