Learning of Multiple Objects in Images using Robust Statistics

Yung-Hsiang Chen

Instrument Technology Research Center, National Applied Research Laboratories

yschen@itrc.org.tw

***Abstract*—This paper is shown learning of multiple objects from images, that can process the extraction of the multiple objects in the scene. Our method is applied to raw image sequence data and extracts the objects one at a time. First, the moving background is learned, and moving objects are found at later stages. The algorithm recursively updates an appearance model of the tracked object so that possible occlusion of the object is taken into account which makes tracking stable. We apply this method to learn multiple objects in image sequences, and present results showing successful extraction of objects from real images.**

***Keywords*—Multiple objects; moving background.**

1. INTRODUCTION

We are given as input a set of images containing views of multiple objects, and wish to learn appearance-based models of each of the objects. Over the last decade or so a layer-based approach to this problem has become popular, where each object is modelled in terms of its appearance and region of support, e.g. Wang and Adelson [1], Irani et al. [2]. Jojic and Frey [3] provided a principled generative probabilistic framework for this task, where each image must be explained by instantiating a model for each of the objects present with the correct instantiation parameters.

A major problem with this formulation is that as the number of objects increases, there is a combinatorial explosion of the number of configurations that need to be considered. If there are J possible objects, and that there are possible values that the instantiation parameters of any one object can take on, then we will need to consider combinations to explain any image. Jojic and Frey [3] tackled this problem by using a variational inference scheme, searching over all instantiation parameters simultaneously. We developed a sequential approach to object discovery whereby each model is extracted in turn from the whole dataset using a robust statistical method. We denote this latter method learning of multiple objects from images. A fuller description of the framework is given in section 2.

Both the method of Jojic and Frey [3] and the GLOMO algorithm work on unordered sets of images. In this case training can be very slow as it is necessary to search over all possible instantiation parameters of at least a single object on every image. However, for video sequences we could considerably speed up the training by first tracking the objects before knowing their full structure. Tracking can approximate the underlying sequence of transformations of an object in the frames and thus learning can be carried out with a very focused search over the neighborhood of these transformations or without search at all when the approximation is accurate. We have developed an algorithm that works in conjunction with the GLOMO method so that the stage where the GLOMO algorithm learns an object (by assuming unordered images) is now speeded up by applying first tracking and then learning based on a focused search. First, the moving background is tracked and learned, and moving foreground objects are found at later stages. The tracking algorithm itself recursively updates an appearance model of the tracked object so that occlusion is taken into account and approximates the transformations by matching this model to the frames through the sequence. This provides accurate transformations, e.g. in the experiments in section 3, we learn the objects without search by using only the transformations found by the tracking algorithm and obtain good results. Updating a model of the tracked object in a robust way enables also the algorithm to recover when it loses the track e.g. due to occlusion of the object.

The structure of the remainder of the paper is as follows: In section 2 we describe the method of learning of multiple objects. The section 3 gives experimental results. We conclude with a discussion in section 4.

2. EXPERIMENTAL METHODS

2.1. Generative model for multiple objects

We first consider the case when there is only one foreground object with appearance f and mask π. Ignoring for a moment the effect of transformations we would have

$$p(x_i) = \pi_i p_f(x_i, f_i) + (1-\pi_i) p_b(x_i, b_i) \quad (1)$$

where $p_f(x_i, f_i) = N(x_i, f_i, \sigma_f^2)$ and $p_b(x_i, b_i) = N(x_i, b_i, \sigma_b^2)$, here $N(x, \mu, \sigma^2)$ denotes a Gaussian distribution x over with mean μ and variance σ^2. Now assume that the transformation space for the foreground object has been discretized to J_f values, indexed by j_f. Application of this transformation to the mask gives a transformed mask $T_{j_f}\pi$, and similarly for f. Similarly the background transformation can take on J_b values indexed by j_b. Thus

$$p(x_i \mid j_f, j_b) = \prod_{i=1}^{P} [(T_{jf}\pi)_i p_f(x_i, (T_{jf} f)_i) + (1-T_{jf}\pi)_{ip_b}(x_i, T_{jb} b)_i)] \quad (2)$$

where 1 denotes the vector with ones. The likelihood of an image $p(x) = \sum_{j_f=1}^{J_f} \sum_{jb=1}^{J_b} P_{jf} P_{jb} p(x \mid j_f, j_b)$ where P_{jf} and P_{jb} are uniform prior probabilities.

The above generative model assumes one foreground and one background layer, however it can be easily extended to include L foreground objects assigned to L depth layers.

2.2. Learning the background

At this stage we consider images containing a background and many foreground objects. However, we concentrate on learning only the background and regarding the foreground objects as outliers (at this stage). This goal can be achieved by robustifying the background model described above so that occlusion can be tolerated.

For a background pixel, the foreground objects are interposed between the camera and the background, thus perturbing the pixel value. This can be modelled with a mixture distribution as $p_b(x_i, b_i) = \alpha_b N(x_i, b_i, \sigma_b^2) + (1-\alpha_b) U(x_i)$, where α_b is the fraction of times a background pixel is not occluded and the robustifying component $U(x_i)$ is a uniform distribution common for all image pixels.

The background can be learned by maximizing the log likelihood $L_b = \sum_{n=1}^{N} \log \sum_{jb} P_{jb} p(x^n \mid j_b)$ where

$$p(x \mid j_b) = \prod_{i=1}^{P} p_b(x_i, (T_{jb} b)_i) \quad (3)$$

and $p_b(x_i, (T_{jb} b)_i)$ has been robustified as explained above. The maximization of the likelihood over (b, σ_b^2) can be achieved by using the EM algorithm and searching over J_b transformations of the background.

2.3. Learning the foreground objects

Imagine that the background b and the most probable transformations $\{j_b^n\}$ in all training set images are known. What we wish to do next is to learn the frst foreground object and ignore the rest objects.

Since multiple objects can exist in our images, a different object from the one being modelled may be interposed between the foreground object we model and the camera, so that we again have a mixture model $p_f(x_i, f_i) = \alpha_f N(x_i, f_i, \sigma_f^2) + (1-\alpha_f) U(x_i)$ where α_f is the fraction of times a foreground pixel is not occluded. Having made this robustification, the model described with one foreground object plus a background (which is already known) can be trained using EM and maximizing with respect to only this object.

A second foreground object is learned by removing those pixels explained by the first foreground object. On image x^n we infer transformation j_1^n, and at pixel i we obtain the posterior probability (or responsibility)

$$r_i^n(j_1^n) = \frac{\alpha_f N(x_i^n, (T_{j_1^n} f_1)_i, \sigma_1^2)}{\alpha_f N(x_i^n, (T_{j_1^n} f_1)_i, \sigma_1^2) + (1-\alpha_f) U(x_i^n)} \quad (4)$$

and compute $\rho_1^n = (T_{j_1^n}\pi_1) * r^n(j_1^n)$. ρ_1^n will roughly give values close to 1 only for the non-occluded object pixelsc of image x^n, and these are the pixels that we wish to remove from consideration. Thus we construct a re-weighted objective function involving ρ_1^n and run the robust learning again. This procedure is then

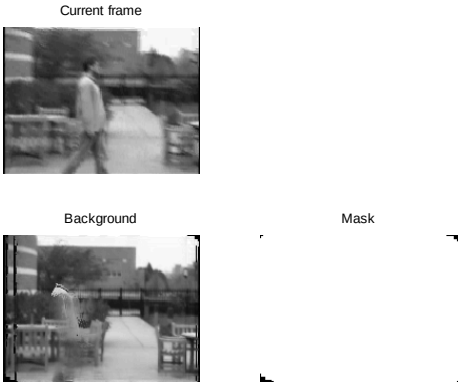
repeated, removing more pixels as more objects are learned.

3. EXPERIMENTS AND RESULTS

We demonstrate our method on two video sequences, is shown as fig. 1and fig. 2.

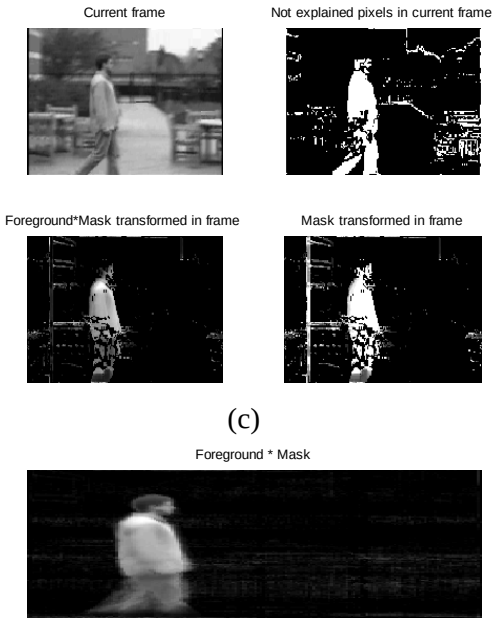
3.1. The experiments of learning the foreground one object

Fig. 1(a) is the tracking the background. Fig. 1(b) is the learning the background using EM, Fig. (c) is the tracking the foreground object #1. Fig. (d) is the learning the foreground object #1 using EM. Fig. 1(e) is the joint refinement of the parameters.

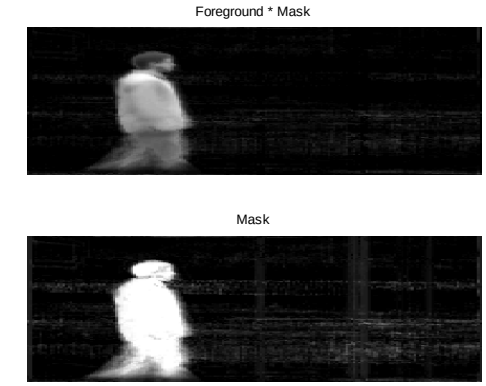


(a)

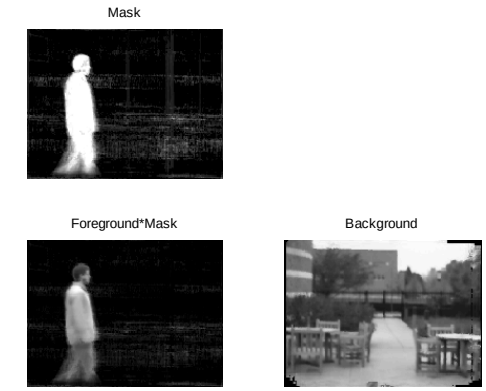
(b)



(c)



(d)



(e)

Fig. 1 The experiments of tracking the foreground one object.

3.2. The experiments of learning the foreground two objects

Fig. 2(a) is the tracking the background. Fig. 2(b) is the learning the background using EM. Fig. 2(c) is the tracking the foreground object #1. Fig. 2(d) is the learning the foreground object #1 using EM. Fig. 2(e) is the Tracking the foreground object #2. Fig. 2(f) is the learning the foreground object #2 using EM. Fig. 2(g) is the compute the occlusion order of the foreground layers and the Joint refinement of the parameters.

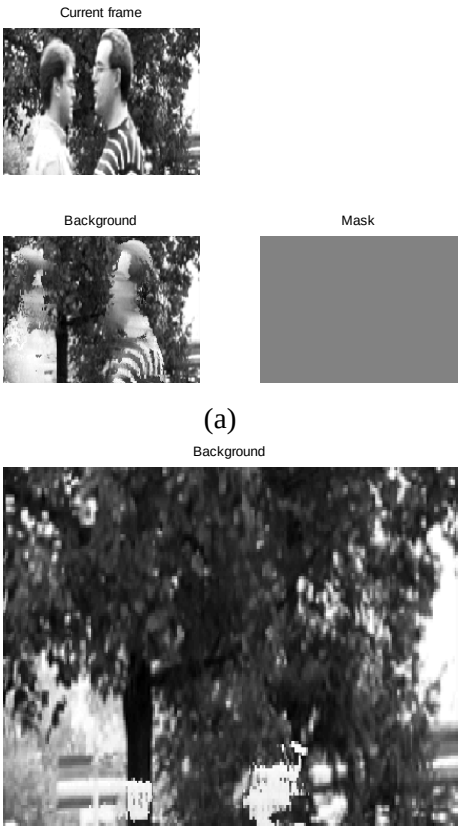


Fig. 2 The experiments of tracking the foreground two objects.

4. CONCLUSIONS

Above we have shown how to use a tracking method. We have also demonstrated that this method can identify articulated parts of objects. Some issues for further work include the identifying when a detected model is a part or an independent object (using mutual information), dealing with parts/objects that have internal variability and finding non-articulated parts.

REFERENCES

[1] J. Y. A. Wang, and E. H. Adelson, "Representing Moving Images with Layers," *IEEE Transactions on Image Processing*, 3(5):625-638, 1994.

[2] M. Irani, B. Rousso, and S. Peleg, "Computing Occluding and Transparent Motions," *International Journal of Computer Vision*, 12(1):5-16. 1994.

[3] N. Jojic, and B. J. Frey, "Learning Flexible Sprites in Video Layers," *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2001*, 199-206, 2001