A PROBABILISTIC FRAMEWORK FOR SEGMENTATION AND TRACKING OF MULTIPLE NON RIGID OBJECTS FOR VIDEO SURVEILLANCE

Aleksandar Ivanović and Thomas S. Huang

Beckman Institute for Advanced Science and Technology
University of Illinois at Urbana-Champaign
405 North Mathews Ave. Urbana, IL 61801 USA
{ivanovic, huang}@ifp.uiuc.edu

ABSTRACT

This paper presents a probabilistic framework for segmenting and tracking multiple non rigid foreground objects for video surveillance, using a static monocular camera. The algorithm combines information in a probabilistic sense and poses the problem of matching the segmented foreground objects with blobs in the next frame as a non bipartite matching problem. To solve this problem, probability is calculated for each possible matching. Initialization of new objects is also treated in a probabilistic manner. The new framework is shown to be able to handle a greater set of difficult situations and to improve performance significantly.

1. INTRODUCTION

In video surveillance, reliable segmentation of moving objects is essential for successful event recognition. Object segmentation can be looked upon as a local discrimination problem of two classes: foreground and background.

The segmentation can be done on pixel, blob and object level as described in Park and Aggarwal [1]. A simple body model (head, upper body, lower body) is used and Markov Random Fields (MRF) lead to segmentation at the blob level. Over segmented blobs are matched to the one from the previous frame, where the matching is posed as a weighted bipartite matching problem and the non-matched blobs are dealt using heuristics. Their method works for fairly simple and contrasted sequences.

Elgammal and Davis [2] also assume a similar model for a human that consists of a head, an upper body and a lower body. However, the assumption of a specific model for the body is too restrictive for a video surveillance system, whereas the system should be able to track people in general (e.g. even small children, people with carts etc.).

The approach of Gomila and Meyer [3] is also to over segment the video and then use relaxation labeling for matching of blobs. Their approach performs video segmentation by matching blobs but has been used only on a sequence containing one object.

Kenna et al. in [4], combine pixel RGB, chromacity and gradient features for segmenting foreground and background. Their algorithm for tracking is basic and if two or more objects get close, they get grouped together.

Wang et al. in [5] label each pixel as background, foreground, or shadow. They claim that the edge information can be used to solve the camouflage problem successfully. However, this model boils down to MAP-MRF optimization and cannot run in real time.

Withagen et al. in [6] use similar features as our model. They use Expectation Maximization (EM) to model background and color histograms for tracking. The new concept in their work is that objects have core ("entire object except for a layer of certain thickness"). However, the foreground object in general may not have a core (e.g. humans with carts). Also, they do not consider tracking of multiple objects or occlusion, and for object detection they compare size to a threshold.

Capellades et al. in [7] introduced the correlogram that can be used instead of color histograms to handle occlusions and improve tracking.

In this paper, we introduce a novel probabilistic framework for the pixel segmentation and for matching of foreground objects to blobs. Our approach also accounts for the grouping of objects. Finally, our method is particularly robust to initialization, which is a very common problem for many tracking algorithms.

2. PIXEL PROBABILITY MODEL/ ESTIMATION OF PROBABILITY OF FOREGROUND (P_F)

In this section we present a novel approach for the estimation of the probability that a frame pixel belongs to the foreground. We initially look at a simple model that is based on the Mahalanobis distance $M_b(x, y)$. Then we de-

This work was supported in part by ARDA under the VACE II Program, Contract No. MDA904-03-C-1787.

0-7803-8554-3/04/$20.00 ©2004 IEEE.

scribe a better way to compute the probability of foreground $P_f(x, y)$.

2.1. Simple background model

Each pixel, for particular dimension of Lu*v* space, can be modeled with a single Gaussian. The Lu*v* space is chosen because it is perceptually more uniform than the RGB color space [8]. We initialize the background pixel model by computing the statistics over the training sequence, if it is free of foreground objects. Otherwise, it is possible to use a bootstrapping algorithm (e.g. [9]).

A simple way to achieve segmentation [10] makes use of the Mahalanobis distance $M_b(x, y)$, which corresponds to the probability that pixel $p(x, y)$ belongs to the background. The Mahalanobis distance $M_b(x, y)$ between each pixel and the corresponding background pixel in Lu*v* color space is defined as:

$$\begin{aligned} M_b(x,y) &= \frac{|L(x,y) - L_{mean}(x,y)|}{\sqrt{L_{var}(x,y)}} \\ &+ \frac{|u^*(x,y) - u^*_{mean}(x,y)|}{\sqrt{u^*_{var}(x,y)}} \\ &+ \frac{|v^*(x,y) - v^*_{mean}(x,y)|}{\sqrt{v^*_{var}(x,y)}} \end{aligned} \quad (1)$$

where $L_{mean}(x, y)$, $u^*_{mean}(x, y)$, and $v^*_{mean}(x, y)$ are the means and $L_{var}(x, y)$, $u^*_{var}(x, y)$, and $v^*_{var}(x, y)$ are the variances of the pixel $p(x, y)$ in Lu*v* color space. This method is very simple and, consequently, does not give satisfactory results.

2.2. Foreground probability model $P_f(x, y)$

A more intelligent approach is to associate more features with every pixel $p(x, y)$, and then assign label $F(x, y) = 0$ for background or $F(x, y) = 1$ for foreground. For each pixel $p(x, y)$, we have a feature vector $A(x, y) = [M_b(x, y), D(x, y), P_h(x, y)]$ and a label $F(x, y)$. $D(x, y)$ is the absolute distance of the current pixel to the background pixel in RGB color space:

$$\begin{aligned} D(x,y) &= |R(x,y) - R_{mean}(x,y)| \\ &+ |G(x,y) - G_{mean}(x,y)| \\ &+ |B(x,y) - B_{mean}(x,y)|, \end{aligned} \quad (2)$$

where the means over the video frames $R_{mean}(x, y)$, $G_{mean}(x, y)$, and $B_{mean}(x, y)$ are calculated in RGB space. $P_h(x, y)$ is a color similarity measure , which is computed from the cumulative histogram of **all tracked objects**, as the number of pixels in the bin that contains $p(x, y)$, divided by the number of the pixels in the histogram. Note that the histogram has 16 bins for each dimension in RGB space.

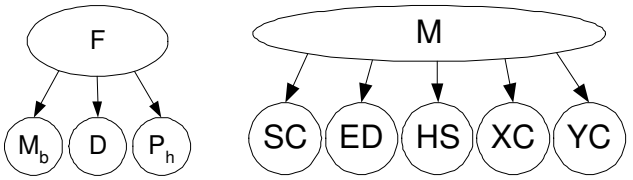

Fig. 1. (a) a Bayesian Network for each pixel (b) a Bayesian Network for matching foreground objects to connected components.

We employ a Bayesian Network (BN) [11] to model the relationship of pixel label $F(x, y)$ with feature vector $A(x, y)$, as depicted in Fig. 1a. We model $P(A|F = 0)$ and $P(A|F = 1)$ using a Gaussian mixture model (GMM) using $K = 5$ Gaussians, whose covariance matrix is diagonal for i.i.d. (independent identically distributed) features. One BN is trained for all the pixels. To obtain the parameters of the GMM, the EM algorithm is applied, where all prior probabilities are assumed equal.

Finally, using Bayes rule, we combine the information from all these features to compute the probability that a pixel belongs to the foreground $P_f(x, y)$ as:

$$P_f(x,y) \propto \frac{P(A|F=1)P(F=1)}{P(A|F=0)P(F=0) + P(A|F=1)P(F=1)}$$

This pdf (probability density function) is used in the following section to successfully separate the foreground and background.

3. CONNECTED COMPONENTS MATCHING

Now that we have found the probability of the foreground $P_f(x, y)$, we want to use it to find the objects in the new frame. First, $P_f(x, y)$ is binarized using an adaptive threshold and a simple union find algorithm is used to find its 8-connected components that correspond to foreground objects.

The connected components can now be matched to the foreground objects in the previous frame. In the ideal case (Fig. 2a), one foreground object is matched to one connected component. If the object has disappeared from the scene, we match the foreground object to a dummy node. When two or more objects are near one other or one is occluded by the other (Fig. 2b), they match to a single connected component. If the background is very similar in color to a part of an object, that part can be mistakenly classified as background and therefore that object will be matched to two or more connected components (Fig. 2c).

This matching can be formulated as a non-bipartite matching problem, and solved using some of the standard algo-

rithms [12]. To solve these problems, probabilistic matching is used where we introduce features for each tracked foreground object and each connected component. Let us consider an ordered pair $(f(i), c(j))$ where $f(i)$ is a foreground object and $c(j)$ is a connected component. Foreground object $f(i)$ and connected component $c(j)$ can be separately described with a feature vector:

$$k(t) = [x_s(t), y_s(t), S(t), H(t), x_c(t), y_c(t)]. \quad (3)$$

Here, ($x_s(t)$, $y_s(t)$) is the horizontal and vertical size of the bounding box, $S(t)$ its size in pixels, $H(t)$ is the color histogram of object/component t, ($x_c(t)$, $y_c(t)$) is the centroid of all the pixels of an object/connected component, and t is index (i.e. $t = i$ or $t = j$). From these features, we derive information for matching a foreground object $f(i)$ to a connected component $c(j)$:

$$m(i,j) = [SC(i,j), ED(i,j), HS(i,j), XC(i,j)YC(i,j)]$$

where $SC(i,j)$ is the size change defined as $SC(i,j) = S(j)/S(i)$, $ED(i,j)$ is the Euclidean distance between ($x_c(i)$, $y_c(i)$) and ($x_c(j)$, $y_c(j)$), $HS(i,j)$ is the similarity between $H(i)$ and $H(j)$, $XC(i,j) = x_s(j)/x_s(i)$ is the horizontal, and $YC(i,j) = y_s(j)/y_s(i)$ is the vertical change.

If an object $f(i)$ and a connected component $c(j)$ should be matched we label ordered pair $(f(i), c(j))$ with $M = 1$, or else with $M = 0$. To compute the probability of matching $P(M = 1|m(i,j))$, we employ the BN (Fig. 1b), which can be trained on a simple case (e.g. tracking of a single object for few seconds).

The probability that the foreground object $f(i)$ corresponds to connected component $c(j)$ is :

$$P(i \rightarrow j) = \frac{P(M = 1|m(i,j))}{P(M = 0|m(i,j)) + P(M = 1|m(i,j))}.$$

The probability that foreground objects $f(i_1)$ and $f(i_2)$ are grouped together and correspond to connected component $c(j)$ is :

$$P(i_1{+}i_2 \rightarrow j) = \frac{P(M = 1|m(i_c,j)}{P(M = 0|m(i_c,j) + P(M = 1|m(i_c,j))},$$

where $f(i_c)$ is the foreground object made by combining objects $f(i_1)$ and $f(i_2)$.

The probability that the foreground object $f(i)$ corresponds to the connected components $c(j_1)$ and $c(j_2)$ is :

$$P(i \rightarrow j_1{+}j_2) = \frac{P(M = 1|m(i,j_c)}{P(M = 0|m(i,j_c) + P(M = 1|m(i,j_c))},$$

where $c(j_c)$ is the connected component resulting from combining the components $c(j_1)$ and $c(j_2)$. $P(j_0 \rightarrow d)$ is the probability foreground object $c(j_0)$ disappeared, where d is dummy node.

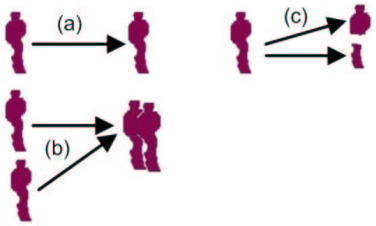

Fig. 2. Matching example: foreground objects are left of the arrows, connected component candidates are right of the arrows: (a) ideal case, one foreground object is matched to one connected component (b) occlusion: two foreground objects are matched to one connected component (c) one foreground object is matched to two connected components.

3.1. Object Detection

The connected components not matched to any foreground object are considered to become new objects. The most simple and common way to decide if a candidate should become new object is to calculate the size of candidate and compare it to a fixed threshold. However, that approach is not very robust and small objects (kids) or objects that are similar to the background may not be detected. To tackle that, we define a set of features $T = [S, LC, SH, CS]$, where S is the size of the connected component, LC is the distance to the nearest location of an appearance of a foreground object, SH is a simple shape feature frequently used for characterizing objects [13] defined as $SH = \frac{area}{perimeter^2}$, and CS is color similarity of object candidate to the average foreground object. Label $N = 0$ is assigned to candidates that are not new objects (e.g. that correspond to a shadow or lighting change), while label $N = 1$ is assigned to candidates that are appearing objects. Fig. 3 depicts the BN for object detection.

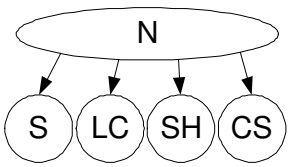

Fig. 3. A Bayesian Network for object detection

4. EXPERIMENTAL RESULTS

Our approach was tested on a 55 minute long indoor sequence. In Fig. 4 tracked objects present in frames 571 and 1885 have been displayed, with the segmentation using the simple approach (Sec. 2.1) and the new approach.

(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 4. Segmentation example: (a)(f) foreground objects (frame 571 and 1885) (b) probability based only on background model (c) P_f probability of foreground (d)(g) segmented objects using only background model (e)(h) segmented objects using P_f probability of foreground.

Our implementation runs at about 0.5 seconds per 720×480 frame and 0.2 seconds per 352×280 using a 2.8 MHz Pentium 4 computer, and the resulting video can be found at `http://www.ifp.uiuc.edu/~ivanovic`

5. CONCLUSION

The contributions of this paper are: (a) a new probabilistic framework for pixel segmentation and for matching of objects to blobs, (b) a framework that can account for grouping of objects, and (c) a method robust to initialization, a common problem for other tracking algorithms.

We conclude that our non bipartite matching formulation is better able to model multi-object tracking and gives more reliable segmentation results.

6. REFERENCES

[1] S. Park, J.K. Aggarwal, "Segmentation and tracking of interacting human body parts under occlusion and shadowing", *Proc. on Workshop on Motion and Video Computing* 2002, pp. 105–111, 5–6 Dec. 2002

[2] A. Elgammal and L. S. Davis, "Probabilistic framework for segmenting people under occlusion," *Proc. IEEE 8th Int. Conf. Computer Vision*, vol. 2, pp. 145–152, 2001.

[3] C. Gomila, F. Meyer, "Graph-based object tracking," *IEEE Int. Conf. on Image Processing*, vol.2 , pp. 41–44, 2003.

[4] S.J. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler. "Tracking Groups of People". *Computer Vision and Image Understanding*, vol. 80 pp. 42–56, 2000.

[5] Y. Wang, T. Tan, and K.-F. Loe, "A probabilistic method for foreground and shadow segmentation," *Proc. IEEE Int. Conf. on Image Processing*, vol. 3, pp. 937–940, 2003.

[6] P. Withagen, K. Schutte and F. Groen , "Object Detection and Tracking Using a Likelihood Based Approach", *Proc. IEEE Int. Conf. on Image Processing*, vol. 1, pp. 589–592, 2003.

[7] M.B. Capellades, D. Doermann, D. DeMenthon, and R. Chellappa, "An appearance based approach for human and object tracking," *Proc. IEEE Int. Conf. on Image Processing*, vol. 2, pp. 85–88, 2003.

[8] C. Poynton, "Frequently Asked Questions about Color", `http://www.poynton.com/ColorFAQ.html`, 2002.

[9] D. Gutchess, M. Trajkovic, E. Cohen-Solal, D. Lyons, and A.K. Jain, "A background model initialization algorithm for video surveillance," Proc. IEEE International Conf. on Computer Vision, vol. 1, pp.733-740, 2001.

[10] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers. "Wallflower: Principles and practice of background maintenance," *Proc. International Conference on Computer Vision*, vol 1, pp. 255–261, 1999.

[11] D. Heckerman, "A tutorial on learning with bayesian networks," Tech. Rep. MSR-TR-95-06, *Microsoft Research, Advanced Technology Division,* March 1995.

[12] H.N. Gabow, Z.Galil, T.H. Spencer , "Efficient implementation of graph algorithms using contraction", *Journal of the ACM* (JACM) vol. 36, Issue 3, pp. 540 – 572, 1989

[13] M.D. Beynon, D.J. Van Hook, M. Seibert, A. Peacock, and D. Dudgeon, "Detecting Abandoned Packages in a Multi-Camera Video Surveillance System," *Proc. Advanced Video and Signal Based Surveillance*, pp. 221–228, 2003