

Video Object Segmentation Based on Graph Cut with Dynamic Shape Prior Constraint

Peng Tang and Lin Gao

*Institute of Image and Graphics, Sichuan University, Chengdu, 610065, P. R. China
pengtnsc@gmail.com*

Abstract

In this work, we present a novel segmentation method for deformable objects in monocular videos. Firstly we introduce the dynamic shape to represent the prior knowledge about object shape deformation in a manner of auto-regressive model which treats the shape as a function of subspace shapes at previous time steps. Then both spatial-temporal image information and model prediction are fused in the framework of Markov random field energy, which can be effectively minimized by graph cut algorithm so as to achieve a global optimum segmentation. To capture model variations, both the orthogonal basis and the autoregressive model parameters are updated on-line using final segmentation results, thereby forming an effective closed loop system. Finally, promising experimental results demonstrate the potentials of the proposed segmentation method with respect to noise, clutter, and partial occlusions.

1. Introduction

Moving object subtraction is a fundamental and critical task in many machine vision applications. Traditional approaches to detect foreground objects are mainly based on background modeling[8, 7, 6] by collecting pixels in the current frame that deviate significantly from the model estimations. If the primary sources of uncertainty are measurement noise, Gaussian densities would be an adequate choice[8] to depict the background characteristics. However, the multimodality of practical outdoor background distribution usually caused its failure. More complex background models are proposed consequently, such as the Mixture of Gaussian[8], nonparametric model with kernel density estimation[7], and ARMA background model[6], etc. Even though many object subtraction algorithms have been proposed in the literature, the problem of

foreground object subtraction in complex environment is still far from being completely solved.

Practically, intensity based segmentation easily fails to subtract meaningful objects with weak edges, in clutter, or under occlusion. As opposed to disturbances and spurious noises, the interest targets tend to have an arbitrary and discriminate features, such as shape, color, texture, etc., which contain substantial evidences for the foreground subtraction in consequent frames. But shape modeling is still difficult since the deformable shapes tends to lie on complex and non-separable low dimensional manifolds in image space. In recent years, many works interest in shape prior aided segmentation, as priors give extra degrees of freedom and render the vision problem well-posed to cope for missing or misleading information in the input images due to noise, clutter, and occlusion[1, 3]. However most works use level set-based dynamic shape. Since graph cut guarantees a global optimum[5], our solution to this problem is to include shape priors in a graph cut based formulation. Freedman[3] has proposed a method to incorporate static shape priors into a graph cuts based approach, but due to the energy function restriction in graph cut[4], it is still difficult to cope with dynamic shapes.

In this work, we propose a dynamic shape under graph cut framework which uses PCA to learn a statistical model of relevant shapes, and models the deformation in the shape subspace. At each iteration the previous segmentation is used for prediction in shape space. These priors are then used to perform segmentation via the graph cut technique.

2. Segmentation By Energy Minimization

Object segmentation in video clips can be considered as pixel-wise classification problem with strong spatial-temporal coherence information. We propose to solve the foreground/background segmentation problem by minimizing the object energy function. Specifically, let \mathcal{P} be a set of all pixels and $p \in \mathcal{P}$ represents

a pixel. Denote the unknown label of each pixel as $f_{p,t}$ that is a binary variable, i.e., $f_{p,t} \in \{0, 1\}$ representing foreground and background respectively. Notice that the suffix t is omitted in this section for convenience. An image energy based objective function incorporated with shape prior can be formulated over the unknown labeling variable f_p of each pixel under the first-order Markov random field (MRF) framework[3].

$$\arg \min [E_{MRF}(f) + \gamma E_{shape}(f)] \quad (1)$$

where E_{shape} is an energy based on the shape prior, while the image energy E_{MRF} is designed as a combination of the data dependent term and the smoothness term, also regarded as a region based term and a boundary term, which is weighted by $0 \leq \lambda \leq 1$ for relative influence:

$$E_{MRF}(f) = \sum_{p \in \mathcal{P}} D_p + \lambda \sum_{\{p,q\} \in \mathcal{N}} V_{p,q} \quad (2)$$

where D_p represents the cost of assigning label f_p to pixel p . In image segmentation, the D_p usually evaluates the evidence for pixel labels based on color distributions in foreground and background.

$$D_p(f_p) = -\ln \Pr(I_p | f_p) \quad (3)$$

In this work, likelihoods were modeled in terms of Gaussian Mixture Models(GMM)[8] in RGB color space, where both foreground and background mixtures were learned via Expectation Maximization (EM).

$$\Pr(I_p | f_p) = \sum_{j=1}^{N_G} w_j \eta(I_p; \mu_{j,f}, \Sigma_{j,f}) \quad (4)$$

$V_{p,q}$ is the penalty of assigning the neighboring pixel p and q with similar pixel intensities to different regions.

$$V_{p,q}(f_p, f_q) = \frac{1}{d(p,q)} \exp\left(-\frac{1}{2h^2} \|I_p - I_q\|^2\right) \quad (5)$$

where $d(p,q)$ is the euclidian distance from point p to q and h is the bandwidth. Since the object intensity tends to be piece-wise, it imposes a tendency to spatial continuity of labels and favors the segmentation boundary along regions where strong edges are detected.

Then we may consider introducing shape priors knowledge E_{shape} , which are usually represented implicitly as silhouette images when no topology constraints are given. Current segmentation contour C and prior shape contour C_0 can easily be obtained from current segmentation $\Omega = \{p \in \mathcal{P} : f_p = 1\}$ and prior shape silhouette Ω_0 . Define a distance function $\Psi : \mathbb{R}^2 \rightarrow \mathbb{R}$ whose value at point p corresponds to the

nearest euclidian distant from the p to the prior shape contour, $\Psi(p, C_0) = \min\{d(p, s) : s \in C_0\}$. Then a distance metric in the shape space can be specified to assign a regular unsigned distance value $dist(C, C_0)$ to measure boundary change in the differential framework for $C \rightarrow C_0$. In order to avoid local differential computations[3], the energy of motion of a contour C is represented by integral measures of boundary change.

$$E_{shape}(C, C_0) = \sum_{p \in \mathcal{P}} \alpha_p \Psi_0(p, C_0) \quad (6)$$

where α_p is a binary parameter that is defined as:

$$\alpha_p(f_p) = \begin{cases} 1 & \text{if xor}(p \in \Omega_0, p \in \Omega) = \text{true} \\ 0 & \text{otherwise} \end{cases}$$

As the posterior energy of Markov random fields given in eq.1 equals to the network flow in corresponding graph $G = (V, E)$, graph cut techniques are implemented to minimize the objective energy function by setting the edge weights E [5], where the vertice set $V = \mathcal{P} \cup \{S, T\}$ is pixel nodes augmented by two special vertices: the source S and the sink T . The edges set $E = \mathcal{N} \cup \{(p, S), (p, T) : p \in \mathcal{P}\}$ consists of all clique pairs of pixels, along with edges between each pixel and the source or sink. Usually 1 or 2 order neighborhood is chosen.

3. Dynamic Shape Modeling

In this section we are going to present the model to character the shape deformation. Though the predicted shapes do not correspond to valid shapes in all instances, it can be used to constrain a segmentation process to favor familiar shape evolutions. Given a set of sequential m -by- n images $\{I_t\}_{t=1}^n$, we can form a training set of vectors $\{\mathbf{y}_t\}_{t=1}^n$ by lexicographic ordering of the pixel elements of each image I_t , where $\mathbf{y}_t \in \mathbb{R}^{d=m \times n}$. Inspired by the literature [2], we define a sequence I_t to be a *dynamic shape* if there exists a set of spatial filters $\phi : \mathbb{R}^n \rightarrow \mathbb{R}^d$, such that by defining $I_t = \phi(\mathbf{x}_t)$, we have $\mathbf{x}_t = \sum_{i=1}^k \mathbf{A}_i \mathbf{x}_{t-i} + \mathbf{v}$ for some choice of matrices $\mathbf{A}_i \in \mathbb{R}^{n \times n}$ and initial condition \mathbf{x}_0 , where $\mathbf{x}_t \in \mathbb{R}^n$ is the state variable and $\mathbf{v} \in \mathbb{R}^n$ is an IID realization from the stationary distribution $q(\cdot)$. Without loss of generality, we can assume $k = 1$ since we can redefine the state of the above model $\mathbf{x}(t)$ to be $[\mathbf{x}_t^T, \mathbf{x}_{t-1}^T, \dots, \mathbf{x}_{t-k}^T]^T$.

Since the temporal observations \mathbf{y}_t suffer from the curse of dimensionality, shape evolution can be analyzed in the low dimensional linear subspace containing meaningful variations. The principal orthogonal directions of maximum variation for $\mathbf{y}(t)$ are known as the

eigenvectors of measurement covariance matrix Σ_y [6]. Therefore, those eigenvectors is an appropriate choice for the filter bank. Thus, the dynamic shape can be considered as second order stochastic process that is formulated as a linear autoregressive model in the subspace learned using PCA.

$$\mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \mathbf{v}, \quad \mathbf{y}_t = \mathbf{C}\mathbf{x}_t + \bar{\mathbf{y}} + \mathbf{w} \quad (7)$$

where $\mathbf{A} \in \mathbb{R}^{n \times n}$ is a state transformation matrix and $\mathbf{C} \in \mathbb{R}^{d \times n}$ is a measurement matrix; $\bar{\mathbf{y}}$ is the mean of \mathbf{y}_i , $\bar{\mathbf{y}} = E(\{\mathbf{y}_i\})$; the noise \mathbf{v} and \mathbf{w} comply with Gaussian distributions, $\mathbf{v} \sim \eta(0, \mathbf{Q})$, and $\mathbf{w} \sim \eta(0, \mathbf{R})$.

Given l measurements $\{\mathbf{y}_t\}_{t=1}^l$, the \mathbf{A} and \mathbf{C} can be estimated by maximizing the likelihood of the dynamic shape learning problem, $\{\tilde{\mathbf{A}}, \tilde{\mathbf{C}}\} = \arg \max_{\mathbf{A}, \mathbf{C}} \Pr(\mathbf{y}_1, \dots, \mathbf{y}_l)$. We use batch PCA [6] to obtain the parameter initial values. Since the covariance matrix calculation may be inefficient, it can be approximated with the sample covariance matrix $\Sigma_y = \mathbf{Y}_t \mathbf{Y}_t^T$, where $\mathbf{Y}_t = [\mathbf{y}_1 - \bar{\mathbf{y}}, \dots, \mathbf{y}_l - \bar{\mathbf{y}}] \in \mathbb{R}^{d \times n}$ is a matrix formed by zero mean vectors. Let $\mathbf{X}_t = [\mathbf{x}_2, \dots, \mathbf{x}_n] \in \mathbb{R}^{l \times n}$, $\mathbf{X}_{t-1} = [\mathbf{x}_1, \dots, \mathbf{x}_{n-1}] \in \mathbb{R}^{l \times n}$. According to eq.7, we can get:

$$\mathbf{X}_t = \mathbf{A}\mathbf{X}_{t-1} + \mathbf{V}_t \quad \mathbf{Y}_t = \mathbf{C}\mathbf{X}_t + \mathbf{W}_t \quad (8)$$

Considering Frobenius norm, the optimized estimation of measuring matrix $\tilde{\mathbf{C}} = \arg \min_{\mathbf{C}} \|\mathbf{Y}_t - \mathbf{C}\mathbf{X}_t\|$ can be computed through the eigenvectors of $\mathbf{Y}_t \mathbf{Y}_t^T$ using singular value decomposition (SVD)[6, 1] of \mathbf{Y}_t , thus, $\mathbf{Y}_t = \mathbf{U}\Sigma\mathbf{V}^T$, where \mathbf{U} and \mathbf{V} are orthogonal matrix, $\mathbf{U} \in \mathbb{R}^{n \times l}$, $\mathbf{V} \in \mathbb{R}^{l \times l}$, and $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_n)$, while σ_i represents the square roots of the singular value. It follows immediately from the fixed rank approximation property of the SVD that the unique solution is given by:

$$\tilde{\mathbf{C}} = \mathbf{U}, \quad \tilde{\mathbf{X}} = \Sigma\mathbf{V}^T \quad (9)$$

The optimized estimation of system transition matrix \mathbf{A} is given as $\tilde{\mathbf{A}} = \arg \min_{\mathbf{A}} \|\mathbf{X}_t - \mathbf{A}\mathbf{X}_{t-1}\|$. Based on normal equation, the sum of error squares is minimized when \mathbf{A} is given as $\mathbf{A} = \mathbf{X}_t \mathbf{X}_{t-1}^T (\mathbf{X}_{t-1} \mathbf{X}_{t-1}^T)^{-1}$. Furthermore, let $\mathbf{M} = [0, 0; \mathbf{I}, 0]$, $\mathbf{N} = [\mathbf{I}, 0; 0, 0]$, so $\mathbf{X}_t = \Sigma\mathbf{V}^T \mathbf{M}$, $\mathbf{X}_{t-1} = \Sigma\mathbf{V}^T \mathbf{N}$, then we can get:

$$\tilde{\mathbf{A}} = \Sigma\mathbf{V}^T \mathbf{M} \mathbf{V} (\mathbf{V}^T \mathbf{N} \mathbf{V})^{-1} \Sigma^{-1} \quad (10)$$

As the shapes of interest objects evolve consistently, The model parameters and the orthogonal decomposition are updated on-line using new evidence to capture the variations outside the training sets. The incremental PCA is implemented in our experiments whose detail can be found in the literature[6].



Figure 1. Top row: Prediction results of proposed method. Bottom row: Ground truth data

4. Experimental Results

To validate the effectiveness of proposed algorithm, a number of experiments were carried out, which includes challenging scenes such as waving branches, shadows, and weak objects boundaries etc. We demonstrate the sufficiencies of linear filter to represent the shape deforming process. Our method is implemented in C++ codes under OpenCV environment, and runs in a 3GB Pentium IV CPU processor machine using 1024 MB of RAM. Optimizations may be performed to reduce hardware requirements in our future work so as to comply our algorithm with the need of real-time video surveillance.

By considering the shape prior of airplanes that is obtained previously and manually, the proposed algorithm is implemented in an airport surveillance application as shown in Fig 2. The setting of graph cut segmentation parameters are given as $\gamma = 0.3$ and $\lambda = 0.5$. It can be seen that due to the blurred boundaries of object in wide view-angle images, the performance of the graph cut segmentation with intensity alone can be considered unsatisfactory (2a). Our algorithm outperformed compared with the object subtraction algorithm. Because of the use of the shape prior, the analysis was performed in a hyper feature space which results in a much fine segmentation outcome even in a high noised condition (2b,c).

We validated the accuracy of the estimated dynamical models by comparing the input sequence to that of synthesized shapes, as given in Fig.1. 10 latent variables are used in the dynamic shapes trained from 24 sequential silhouette images. To verify the robustness of our algorithm when handling nonrigid moving targets with partial occlusion in cluttered background, the experiment of a pedestrian passing by a swaying tree is shown in Fig.3. The mixture of Gaussian[8] algorithm is chosen as the comparison, where the number of Gaussian is chosen as 5 for each pixel. As can be seen from the Fig.3(a), the MGM fails to segment the object accurately, because the it is a method based on

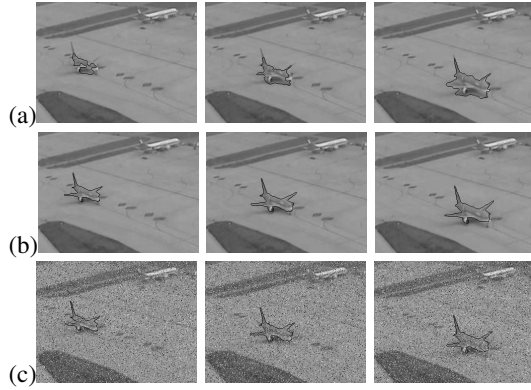


Figure 2. Scenario of a pedestrian occluded by branches. (a) Segmentation with intensity alone (b) Our Method result (c) Our Method result under strong noise

pixel level intensity statistics. The graph cut parameter settings are still $\gamma = 0.3$ and $\lambda = 0.5$. By considering shape regional information, the static shape prior[3] performs better but still can not adapt to the shape variation. The outperforming of our algorithm is because of introducing shape prior with dynamic evolution to model the foreground and background simultaneously so as to reduced the background noise effectively.

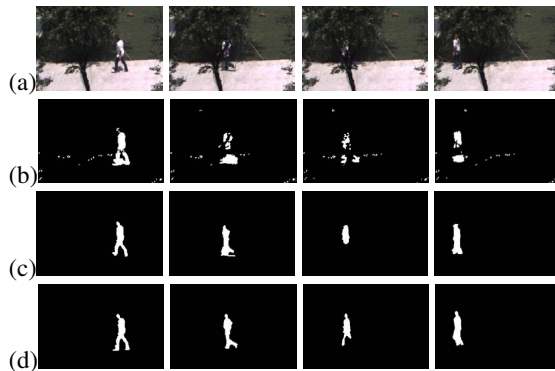


Figure 3. Scenario of a pedestrian occluded by branches. (a) Original image (b) Segmentation result of Mixture of Gaussian model (c) Segmentation result of static shape prior (d) Our Method result

5. Conclusions

In this work, an implicit representation of dynamical shape deformation is formulated using autoregressive models to capture the spatial-temporal correlations. The model prediction is then introduced as the prior guidance into object segmentation process. The resulting dynamical shape models therefore allow us to handle shapes of varying topology. Consequently the optimization is solved under Markov random fields framework by using the graph cut algorithm to obtain the global optimum solution; The final segmentation results are also implemented to update the model parameters iteratively. Experimental results demonstrate that the dynamical shape priors outperform static shape priors in the presence of large amounts of disturbances or noise.

Acknowledgments

The work described in this paper was funded by the NSFC Project under Grant 60736046.

References

- [1] D. Cremers. Dynamical statistical shape priors for level set-based tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(8):1262–1273, 2006.
- [2] G. Doretto, A. Chiuso, Y. N. Wu, and S. Soatto. Dynamic textures. *International Journal of Computer Vision*, 51(2):91–109, 2003.
- [3] D. Freedman and T. Zhang. Interactive graph cut based segmentation with shape priors. In *Proc. of the Conf. on Computer Vision and Pattern Recognition*, volume 1, pages 755–761, San Diego, CA, 2005.
- [4] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):147–159, 2004.
- [5] N. Komodakis and G. Tziritas. Approximate labeling via graph cuts based on linear programming. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(8):1436–1453, 2007.
- [6] A. Monnet, A. Mittal, N. Paragios, and V. Ramesh. Background modeling and subtraction of dynamic scenes. In *Proc. of the Conf. on Computer Vision*, pages 1305–1312, Madison, Wisconsin, 2003.
- [7] Y. Sheikh and M. Shah. Bayesian modeling of dynamic scenes for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(11):1778–1792, 2005.
- [8] C. Stauffer and W. E. L. Grimson. Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):747–757, 2000.