Efficient Visual Tracking by Probabilistic Fusion of Multiple Cues

Hanzi Wang and David Suter

Institute for Vision Systems Engineering
Department of Electrical and Computer Systems Engineering
Monash University, Clayton Vic. 3800, Australia
{hanzi.wang, d.suter}@eng.monash.edu.au

Abstract

It has been shown that integrating multiple cues will increase the reliability and robustness of a vision system in situations that no single cue is reliable. In this paper, we propose a method by fusing multiple cues (i.e., the color cue and the edge cue). In contrast to previous work, we propose a novel shape similarity measure which includes the spatial distribution of, the number of, and the gradient intensity of the edge points. We integrate this shape similarity measure with our recently proposed SMOG-based color similarity measure in the framework of particle filter (PF). Experimental results demonstrate the high robustness and effectiveness of our method in handling appearance changes, cluttered background, moving camera, and occlusions.

1. Introduction

The task of visually tracking objects is to use single or multiple cues of the objects to infer the hidden states (such as 2D positions, scales, poses, etc.) of the objects. Color and edge points are two of the most important visual cues and have been widely used for the task of visual tracking. However, most of the previous algorithms are based on a single cue. For example, in [1, 2, 3], the edge cue (or contour) was used to track objects; whereas the authors of [4, 5, 6, 7] utilized the color cue for tracking.

Although impressive results have been shown in the above-mentioned literature, it is obvious that no single cue is reliable in all situations. As multiple cues could provide complementary information, the fusion of several different cues will lead to an increased reliability and robustness.

There is some work that uses multiple cues to detect and track objects (e.g., [8, 9, 10, 11, 12]). Birchfield [8] used gradient intensity and a color histogram of the target for robust head tracking. Based on a factorized graphical model, Wu [9] presented a co-inference approach to integrate multiple cues for tracking. Shen [10] combined both color and shape information for object tracking. The authors of [11] fused the color cue with the sound cue for teleconferencing, and they fused color with the motion cue for video surveillance with a fixed camera. In [12], the authors used color and edge orientation histogram features to track object.

In this paper, we propose a tracking method in which the color cue and the edge cue are fused. In contrast to previous related work [8, 9, 10, 11, 12, 13], we propose a new shape similarity measure which incorporates the spatial distribution of edge points, the gradient intensity and the number of edge points (Section 3.1). We fuse this shape similarity measure with our newly proposed SMOG (Spatial-color Mixture of Gaussians) based color similarity measure [7] in the framework of the sequential Monte Carlo method [2].

2. Particle filter and the color cue

Our method is based on the framework of the particle filter (PF). For completeness, we summarize the main idea behind the particle filter in the next subsection.

2.1. Particle filter

Denoting by X_t and Y_t respectively the hidden state and the observation at time t. The goal of the particle filter is to estimate the posterior probability density function (pdf) $p(X_t|Y_{1:t})$ of the target state. If posterior pdf $p(X_{t-1}|Y_{1:t-1})$ is known at time t-1, current posterior pdf can be obtained by the following two steps:

Prediction Step:

$$p(X_t|Y_{1:t-1})=\int p(X_t|X_{t-1})p(X_{t-1}|Y_{1:t-1})dX_{t-1} \quad (1)$$

Update Step:

$$p(X_t|Y_{1:t}) \propto L(Y_t \mid X_t)p(X_t|Y_{1:t-1}) \quad (2)$$

In the prediction step, the prior $p(X_{t-1}|Y_{1:t-1})$ is propagated to $p(X_t|Y_{1:t-1})$ through a system dynamical model $p(X_t|X_{t-1})$. The predicted state is then corrected by an observation likelihood function $L(Y_t|X_t)$ in the update step.

In order to avoid the integral in Equation (1), the particle filter approximates the posterior pdf by a set of weighted particles $\{X_t^j, W_t^j\}_{j=1}^M$, (where $\sum_{j=1}^M W_t^j = 1$).

If we sample the particles from an importance density $X_t^j \sim q(X_t^j \mid X_{t-1}^j, Y_{1:t})$, the weight of each new particle becomes:

$$W_t^j \propto \frac{L(Y_t \mid X_t^j)p(X_t^j \mid X_{t-1}^j)}{q(X_t^j \mid X_{t-1}^j, Y_{1:t})} \quad (3)$$

The 18th International Conference on Pattern Recognition (ICPR'06)
0-7695-2521-0/06 $20.00 © 2006 IEEE

For simplicity, the authors of [14] set $q(\mathrm{X}_t^j \mid \mathrm{X}_{t-1}^j, \mathrm{Y}_{1:t}) = \mathrm{p}(\mathrm{X}_t^j \mid \mathrm{X}_{t-1}^j)$ and thus $W_t^j \propto L(\mathrm{Y}_t \mid \mathrm{X}_t^j)$.

The observation likelihood function plays a very important role in the particle filter because it determines the weights of particles in Equation (3), which affects the way particles are re-sampled. Therefore, a good observation likelihood function can significantly improve the performance of tracking.

2.2. SMOG-based color similarity measure

In [7], we proposed a Spatial-color Mixture of Gaussians (we call SMOG) based color similarity measure. SMOG represents richer information than the general color histogram because it considers not only the colors in the region but also their spatial layout.

We model the appearance of a target object O_t by SMOG with k modes $\{\omega_{t,l}^{O_t}, \mu_{t,l}^{\mathrm{S},O_t}, \mu_{t,l}^{\mathrm{C},O_t}, \Sigma_{t,l}^{\mathrm{S},O_t}, \Sigma_{t,l}^{\mathrm{C},O_t}\}_{l=1,\ldots,k}$, where $\omega_{t,l}^{O_t}, \mu_{t,l}^{\mathrm{S},O_t}, \mu_{t,l}^{\mathrm{C},O_t}, \Sigma_{t,l}^{\mathrm{S},O_t}, \Sigma_{t,l}^{\mathrm{C},O_t}$ respectively represent the weight of, the spatial mean of, the color mean of, the spatial covariance of and the color covariance of the lth mode. In calculating the spatial mean and the spatial covariance, one should normalize the coordinate space first.

Let $\Lambda_{t,l}^{\mathrm{C}}$ and $\Lambda_{t,l}^{\mathrm{S}}$ be respectively the color and the spatial similarity measure between the lth mode of the target candidate O_v and the lth mode of the target object O_t at time t. We have:

$$\Lambda_{t,l}^{\mathrm{C}} = \min(\omega_{t,l}^{O_t}, \omega_{t,l}^{O_v})$$

$$\Lambda_{t,l}^{\mathrm{S}} = \exp\left\{-\frac{1}{2}(\mu_{t,l}^{\mathrm{S},O_v} - \mu_{t,l}^{\mathrm{S},O_t})^T (\hat{\Sigma}_{t,l}^{\mathrm{S}})^{-1} (\mu_{t,l}^{\mathrm{S},O_v} - \mu_{t,l}^{\mathrm{S},O_t})\right\} \quad (4)$$

where $(\hat{\Sigma}_{t,l}^{\mathrm{S}})^{-1} = (\Sigma_{t,l}^{\mathrm{S},O_v})^{-1} + (\Sigma_{t,l}^{\mathrm{S},O_t})^{-1}$.

The SMOG-based color similarity measure in a joint spatial-color space can be written as:

$$\Lambda(O_t, O_v) = \sum_{l=1}^{k} \Lambda_{t,l}^{\mathrm{S}} \Lambda_{t,l}^{\mathrm{C}} \quad (5)$$

The color likelihood function is given by:

$$L_{color}(\mathrm{Y}_{t,color} \mid \mathrm{X}_t) \propto \exp\left\{-\frac{1}{2\sigma_{color}^2}(1 - \Lambda(O_t, O_v))\right\} \quad (6)$$

The SMOG-based color similarity measure is more discriminative than the color histogram based similarity measure and thus leads to better tracking results.

3. Fusion of the color cue and the edge cue

Although the SMOG-based color cue works robustly for most situations, we found it may work poorly when the appearance model greatly changes (e.g., a person rotates his/her head). As the edge cue may provide complementary information to the color cue, we will provide a new shape similarity measure and then we fuse it with SMOG-based color similarity measure in the framework of particle filter.

3.1. A new shape similarity measure

An accurate shape observation model is crucial in tracking. To represent the shape of an object, we first detect edge points. Denote by x_i^* the ith pixel location. The gradient intensities $\{G(\mathrm{x}_i^*)\}_{i=1,\ldots,N_0}$ of a set of pixels along a hypothesized contour are calculated. The pixels whose values of G(.) are larger than a threshold $\mathrm{T_G}$ (we set $\mathrm{T_G}$=12) are treated as edge points. The rest of the pixels are treated as noise and ignored at the later stage.

In contrast to [9, 10], which used only a set of selected contour points, we consider all edge points around the perimeter of the contour. Moreover, in contrast to [1, 8] which used only the gradient intensity as the feature of the edge cue, and [12] which used the edge orientation as the feature, we employ three features in our method: (a) the spatial distributions of the edge points; (b) the corresponding number of the edge points; and (c) the gradient intensities at the edge points. We consider (b) and (c) because we do not expect the size of an object and the gradient intensity of edge points will change dramatically at neighboring frames.

In order to represent the spatial distributions of the object shape, a spatial histogram along the hypothesized contour is used. The contour of the object shape is separated into U bins along the anticlockwise direction. Let b: $\mathrm{R}^2 \rightarrow \{1,\ldots,N\}$ be the function which associates to the pixel at location x_i^* a number $b(\mathrm{x}_i^*)$ corresponding to the index of the histogram bin, where N is the number of edge points around the object contour. The number of the edge points falling into the u'th bin can be written as:

$$h_u = \sum_{i=1}^{N} \delta[b(\mathrm{x}_i^*) - u] \quad (7)$$

where δ is the Kronecker delta function.

The probability of the edge points falling into the u'th bin is:

$$\hat{h}_u = h_u / \sum_{u=1}^{U} h_u \quad (8)$$

where we have $\sum_{u=1}^{U} \hat{h}_u = 1$.

The mean gradient intensity of the edge points in the u'th bin is:

$$G_u = (\sum_{i=1}^{N} G(\mathrm{x}_i^*)\delta[b(\mathrm{x}_i^*) - u]) / h_u \quad (9)$$

For a target object O_t and a target candidate O_v, we write the similarity of the number of the edge points between the u'th bin of O_t and O_v as:

$$S_{u,N} = \min(h_u(\mathrm{O}_t), h_u(\mathrm{O}_v)) / \max(h_u(\mathrm{O}_t), h_u(\mathrm{O}_v)) \quad (10)$$

For gradient intensity, similarly, we write the similarity between the u'th bin of O_t and O_v as:

$$S_{u,G} = \min(G_u(O_t), G_u(O_v)) / \max(G_u(O_t), G_u(O_v)) \quad (11)$$

The shape similarity between O_t and O_v can be formulated as:

$$\Gamma(O_t, O_v) = \sum_{u=1}^{U} \left[\min\left(\hat{h}_u(O_t), \hat{h}_v(O_t)\right) S_{u,N} S_{u,G} \right] \quad (12)$$

The likelihood function of shape can be written as:

$$L_{shape}(\mathrm{Y}_{t,shape} \mid \mathrm{X}_t) \propto \exp\left\{ -\frac{1}{2\sigma^2_{shape}} (1 - \Gamma(O_t, O_v)) \right\} \quad (13)$$

Although we consider the number of the edge points and the gradient intensity in the shape similarity in Equation (13), the values of $\{h_u\}_{u=1,\ldots,U}$ and $\{G_u\}_{u=1,\ldots,U}$ in the target template are adaptive and are updated at each frame. As shown in Section 4, our method is robust to scaling and changes of gradient intensity.

3.2. Framework to fuse the color cue and the edge cue

The particle filter has two advantages: (a) it can track multiple hypotheses simultaneously; (b) information from different measurement sources can be easily fused in a principled manner. We fuse the color cue and the edge cue in the framework of the particle filter. The detailed procedures are as follows:

(0) **Initialize** the color model (with k_0 Gaussians) and shape model of the target. We use only k ($\leq k_0$) key Gaussians where $\sum_{l=1}^{k} \omega_{t=0,l}^{O_t} > T_c$.

(1) **Predict** the region R, which includes all particles, in the 2D image (as shown in Figure 1 (b)). We employ a random walk dynamic model for the particle filter.

(2) **Extract** the foreground pixel (*FP*) map within the region R. A pixel is labeled as 1 if its Mahalanobis distances to any of the k Gaussiansis less than a constant (i.e., 2.5). The *FP* map is shown in Figure 1 (c).

(3) **Generate** the edge image within R and filter out the edge pixels whose labels in the *FP* map are not equal to 1. We remove the pixels whose neighboring pixel number is small. The final edge image is illustrated in Figure 1 (e).

(4) **Calculate** the color similarity $\{\Lambda_i\}_{i=1,\ldots,M}$ in Equation (5) and shape similarity $\{\Gamma_i\}_{i=1,\ldots,M}$ in Equation (12) for each particle. Normalize the color and shape similarity measure by:

$$\bar{\Lambda}_i = \Lambda_i / \max(\{\Lambda_i\}_{i=1,\ldots,M}) \qquad \bar{\Gamma}_i = \Gamma_i / \max(\{\Gamma_i\}_{i=1,\ldots,M}) \quad (14)$$

(5) **Measure and weight** each particle. Here, we use the normalized $\bar{\Lambda}_i$ and $\bar{\Gamma}_i$ for the observation likelihood function by Equation (6) and (12):

$$L(\mathrm{Y}_t \mid \mathrm{X}_t) = L_{color}(\mathrm{Y}_{t,color} \mid \mathrm{X}_t) L_{shape}(\mathrm{Y}_{t,shape} \mid \mathrm{X}_t) \quad (15)$$

Normalize the weight.

(6) **Output** the estimated mean state:

$$E[\mathrm{X}_t] = \sum_{j=1}^{M} W_t^j \mathrm{X}_t^j \quad (16)$$

(7) **Update** the parameter values of the target by an exponentially forgetting scheme similar to [7].

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

Figure 1. Some steps of the proposed procedures. (a) the original image with the initialized window; (b) the predicted region R; (c) the extracted *FP* map; (d) the gradient intensity image; (e) the gradient intensity image after removing pixels in step (3).

4. Experiments

In this section, we test the performance of our method. We use the test video sequences from http://vision.stanford.edu/~birch/headtracker/seq/. We implement our method in the normalized [r, g, I] color space (similar to [7]).

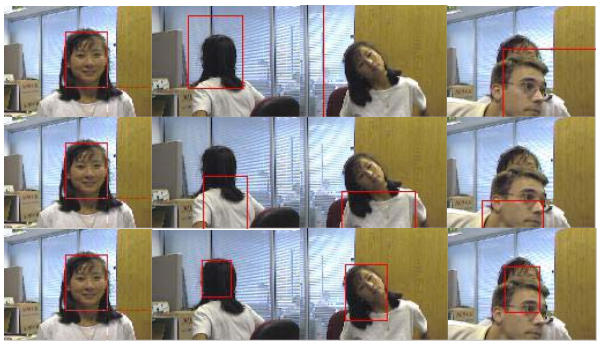

Figure 2. Tracking results with the color cue only (top), the edge cue only (middle) and multiple cues (bottom).

In the first example, we test a video sequence and some resulting frames are shown in Figure 2. This sequence is difficult because the background contains heavily cluttered scene and similar color to the girl's face. The girl's head rotated, skewed and scaled. There was also an occlusion at the end where two heads, with very close colors, occluded each other. The sequence contains 500 frames. Thus it can test the effectiveness of update in our method. Similar to [9] and [10], we

test the tracking performance of both single cue and multiple cues.

As shown in Figure 2, when we use the color cue alone, the method begins to break down with the rotation of the girl's head. The method using the edge cue (alone) fails to track the girl's head as the background contains many clutter edges. Only our proposed method using multiple cues successfully tracks the head throughout the sequence.

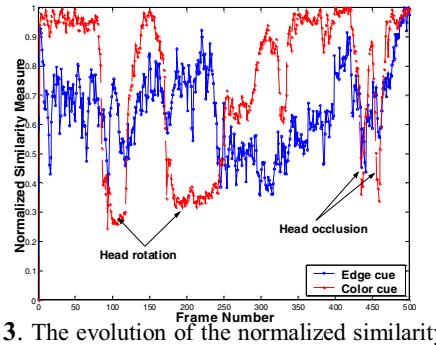

Figure 3. The evolution of the normalized similarity scores of each cue.

Figure 3 illustrates the evolution of the normalized similarity scores of the edge cue and the color cue for the whole video sequence. We can see that when the head rotates, the color-based similarity scores drop dramatically while the edge-based similarity measure is affected relatively less. When occlusions occur, both of the cues are influenced greatly.

Figure 4. More tracking results by the proposed method.

Figure 4 shows more tracking results. We can see our method using multiple cues has achieved attractive results and it is highly robust to cluttered backgrounds, color distractors, and occlusions.

5. Conclusions

In this paper, we provide a method that combines the color cue with the edge information. The provided method achieves better tracking performance in difficult situations. By doing so, we propose a new shape similarity measure considering the spatial distribution of, the number of, and the corresponding gradient intensities of the edge points. Experiments show very promising results in handling cluttered background, moving camera, appearance changes, occlusions, etc.

Acknowledgements

We thank Dr. Chunhua Shen for his valuable comments, and the ARC for support (grant DP0452416).

References

1.Birchfield, S. *An Elliptical Head Tracker*. in *Asilomar Conference on Signals, Systems, and Computers*. 1997. p. 1710-1714.

2.Isard, M. and A. Blake, *Condensation-Conditional Density Propagation for Visual Tracking*. IJCV, 1998. **29**(1): p. 5-28.

3.A.Blake and M. Isard, *Active Contours*. 1998, London, U.K.: Springer-Verlag.

4.Nummiaroa, K., E. Koller-Meierb, and L.V. Gool, *An Adaptive Color-Based Particle Filter*. IVC, 2003. **21**: p. 99-110.

5.Perez, P., et al. *Color-Based Probabilistic Tracking*. *ECCV*. 2002. p. 661-675.

6.Comaniciu, D., V. Ramesh, and P. Meer, *Kernel-based Object Tracking*. PAMI, 2003. **25**(5): p. 564 - 577.

7.Wang, H., D. Suter and K. Schindler, *Effective Appearance Model and Similarity Measure for Particle Filtering and Visual Tracking*. *ECCV*, 2006. p.606-618.

8 Birchfield, S. *Elliptical Head Tracking Using Intensity Gradients and Color Histograms*. CVPR. 1998. p. 232-237.

9.Wu, Y. and T.S. Huang, *Robust Visual Tracking by Integrating Multiple Cues Based on Co-Inference Learning*. IJCV, 2004. **58**(1): p. 55-71.

10.Shen, C., A.v.d. Hengel, and A. Dick. *Probabilistic Multiple Cue Integration for Particle Filter Based Tracking*. DICTA. 2003. p. 309-408.

11.Pérez, P., J. Vermaak, and A. Blake, *Data Fusion for Visual Tracking with Particles*. Proceedings of the IEEE, 2004. **92**(3): p. 495-513.

12.Yang, C., R. Duraiswami, and L. Davis. *Fast Multiple Object Tracking via a Hierarchical Particle Filter*. *ICCV*. 2005. p. 212-219.

13.Spengler, M. and B. Schiele, *Towards Robust Multi-cue Integration for Visual Tracking*. MVA, 2003. **14**: p. 50–58.

14.Doucet, A., S. Godsill, and C. Andrieu, *On Sequential Monte Carlo Sampling Methods for Bayesian Filtering*. Statistics and Computing, 2000. **10**(3): p. 197-208.