OBJECT TRACKING BY ADAPTIVE FEATURE EXTRACTION

Bohyung Han and Larry Davis

Dept. of Computer Science
University of Maryland
College Park, MD 20742, USA
{bhhan, lsd}@cs.umd.edu

ABSTRACT

Tracking objects in the high-dimensional feature space is not only computationally expensive and but also functionally inefficient. Selecting a low-dimensional discriminative feature set is a critical step to improve tracker performance. A good feature set for tracking can differ from frame to frame due to the changes in the background against which the tracked object is viewed, and an on-line algorithm to adaptively determine a distinctive feature set would be advantageous. In this paper, multiple heterogeneous features are assembled, and likelihood images are constructed for various subspaces of the combined feature space. Then, the most discriminative feature is extracted by Principal Component Analysis (PCA) based on those likelihood images. This idea is applied to the mean-shift tracking algorithm [1], and we demonstrate its effectiveness through various experiments.

1. INTRODUCTION

There has been intensive research on tracking algorithms, and several fundamental frameworks have been investigated. The well-known deterministic tracking algorithm based on mean-shift [1] searches for a local maximum of the appearance similarity. The Kalman filter and its extensions [2, 3], and particle filters [4, 5, 6] are employed for probabilistic tracking to mitigate the weakness of deterministic algorithms. However, only limited research has been performed on adaptively finding discriminative feature sets, even though trackers rely on the "contrast" between the appearances of the target and its surrounding. Therefore, we suggest a feature extraction method to increase robustness of tracking algorithms.

Tracking can be performed in a multitude of different feature spaces: color, texture, motion, etc. However, the high-dimensional space is not appropriate for real-time applications, and its effectiveness is reduced by redundancy across the dimensions. The performance – accuracy and speed – of trackers can be improved by adopting a discriminative feature space and reducing the dimensionality. Sometimes, the aggregation of several heterogeneous visual features produces better tracking results, but the combination procedure is not straightforward. Likelihood images created by comparing foreground (tracked object) and background histograms in each feature can be used to choose useful features for tracking.

Previously, Swain and Ballard [7] address the target localization problem with the backprojection algorithm, and Ennesser and Medioni [8] modify this algorithm by using local histogram matching. However, those methods deal only with still images and have not been extended to video.

For object tracking, most algorithms exploit only preselected features, and do not change them during the tracking process. There have been only a few attempts to adapt tracking features on-line. Stern and Efros [9] improve tracking performance by choosing the best from 5 feature spaces and switching amongst them in each frame. In [10], a ranking system is proposed to select the best feature space among 49 candidates acquired by linear combination of 3 likelihood images in R, G, and B space. However, the generation of 49 likelihood images is time-consuming and it is very difficult to extend this technique to higher dimensional feature spaces. On the other hand, the feature selection and extraction method described in [11] proposes a feature value weighting scheme based on the background color information and focuses on salient target parts from the representation of target and candidate model. However, it ignores problems caused by insufficient samples in the high-dimensional feature space.

In this paper, two different color spaces – RGB and normalized RGB (*rgb*) – are combined in a single tracker by means of likelihood images, and likelihood images for various subspaces are constructed to alleviate the information loss associated with analyzing only 1D projections of these color spaces. The *curse of dimensionality* problem is avoided by equalizing the number of bins in the histogram of every subspace. Feature extraction is performed by PCA, and the number of dimensions is determined by the proportion of eigenvalues. Also, weights based on the property of the likelihood image are assigned to each target pixel, and the mean-

0-7803-8554-3/04/$20.00 ©2004 IEEE.

shift algorithm is utilized for tracking.

The rest of this paper is organized as follows. We explain how to generate the likelihood images in section 2, and the feature extraction and the weight assignment are addressed in section 3. The mean-shift tracking algorithm and experimental results are presented in section 4.

2. LIKELIHOOD IMAGES

Log-likelihood ratios are obtained from histograms of foreground and background pixels with respect to a given feature, and likelihood images are constructed based on the log-likelihood ratio. Then, the salient region in the target (foreground) can be detected by identifying high likelihood ratios.

Suppose the foreground is given and the background is regarded as the rectangular region surrounding the foreground. For the given feature space, let $\phi_{fg}(i)$ and $\phi_{bg}(i)$ be the frequency of pixels with value i in the foreground and the background, respectively. The log-likelihood ratio for a feature value i is given by

$$L(i) = \max\left(-1, \min\left(1, \log\frac{\max\left(\phi_{fg}(i), \delta\right)}{\max\left(\phi_{bg}(i), \delta\right)}\right)\right) \quad (1)$$

where δ is a very small number. The likelihood image for each feature is created by backprojecting the ratio into each pixel in the image.

Here, we use every subset of the RGB and *rgb* color channels as a feature set, so that 14 different likelihood images are generated for feature extraction.

Likelihood images for one- and multi-dimensional subspaces are created, which allows us to find more discriminative feature spaces since more basis images are provided for feature extraction. However, the high dimensional subspace might suffer from the *curse of dimensionality*. This problem is alleviated by maintaining the same number of bins for every subspace, so that each bin of each histogram occupies the same volume in the feature space. Specifically, there are 64 bins for every subspace (1×64 in 1D, 8×8 in 2D, and $4 \times 4 \times 4$ in 3D), and each bin equally divides the feature space. With this method, the log-likelihood ratio becomes more robust, and every likelihood image is independent of every other. Figure 1 shows that the likelihood images derived from multi-dimensional subspaces are more distinctive than those from 1D subspaces. Note that every likelihood image is normalized to the gray scale ($0 \sim 255$) for display. Figure 2 shows the benefit of using *rgb* as well as RGB. The likelihood images from the *rgb* subspaces are much more distinctive, even though the tracked object in the original *rgb* image is hardly recognizable to the human eye.

Since the most discriminative feature set can change from frame to frame, tracking in only one feature space might suddenly lose the object. Our method can select the feature set adaptively since the likelihood image provides the methodology to combine feature spaces and select better features.

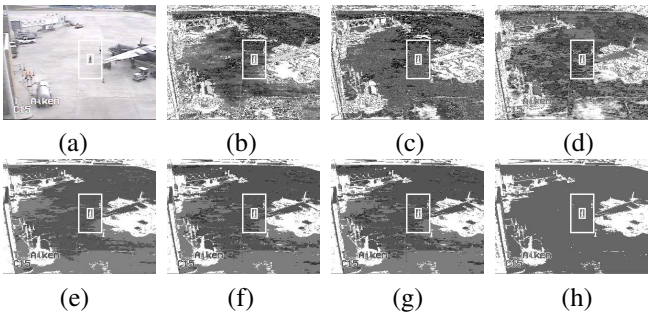


Fig. 1. Comparison between likelihood images for 1D subspaces and multi-dimensional subspaces. (a) original image (b) R (c) G (d) B (e) RG (f) GB (g) BR (h) RGB

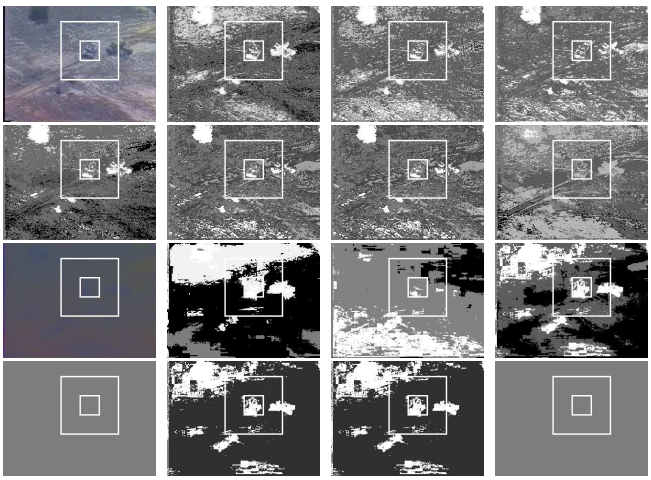


Fig. 2. Likelihood images for 1D subspaces of RGB and *rgb* space. Original image in RGB space, likelihood images for R, G, B, RG, GB, BR, RGB, original image in *rgb* space, likelihood images for *r, g, b, rg, gb, br, rgb* (in order of left to right and top to bottom)

3. FEATURE ANALYSIS

3.1. Feature Extraction

Our objective is to identify the most discriminative likelihood image with the lowest dimensionality in order to reduce the computational complexity and improve the tracking accuracy. In previous research [10], the most discriminative feature is determined by evaluating a small set of pre-defined linear combinations of likelihood images. In traditional pattern recognition, Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) are widely used to reduce the dimensionality.

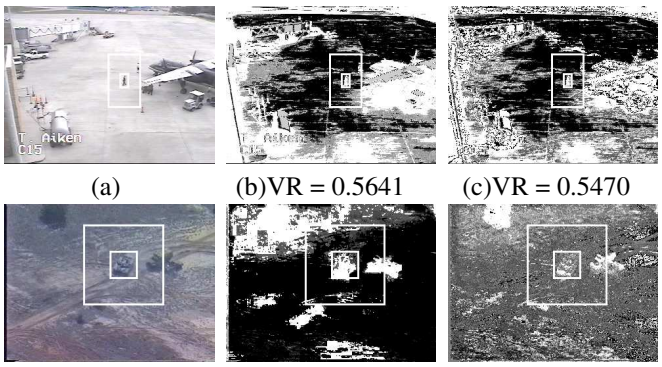
In the original color image, since the histograms of the

foreground and the background region are multi-modal, the linear discriminant method may not be suitable. However, notice that pixels in the foreground region mostly have positive values while the background is mainly composed of negative value pixels in the likelihood image. Even though the foreground and the background cannot be perfectly separated by the linear hyper-plane, we can expect that most pixels would be classified correctly by it. Also, the linear methods are much faster than their non-linear counterparts such as Kernel LDA [12] and Kernel PCA [13].

In this paper, we perform feature extraction with PCA. Suppose that S_{fg} and S_{bg} are the set of n-dimensional vectors sampled from the foreground and background area of n likelihood images, and that C is the $n \times n$ covariance matrix of these vectors. Let $\mathbf{e}_i$ $(i = 1, \ldots, n)$ be the eigenvectors associated with the eigenvalues λ_i which are sorted in non-increasing order. The eigensubspaces composed of k-basis vectors $\mathbf{e}_1, \cdots, \mathbf{e}_k$ are taken for tracking, where the k is the smallest number satisfying the following inequality,

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{n} \lambda_i} > th \tag{2}$$

and th is a pre-defined threshold value. As a result, k extracted likelihood images are used for tracking; Figure 3 shows results of feature extraction for two given images.



(a) (b)VR = 0.5641 (c)VR = 0.5470

(d) (e)VR = 0.6722 (f)VR = 0.5439

Fig. 3. Comparison between two feature extraction methods (2 examples) (a)(d) Original images (b)(d) PCA-based feature extraction (likelihood images associated with the highest eigenvalue), (c)(e) feature extraction by [10],

As suggested in [10], the degree of the salience for the foreground in a likelihood image can be measured by the variance ratio

$$VR = \frac{\text{Var}(S_{fg}^{\mathbf{e}_i} \bigcup S_{bg}^{\mathbf{e}_i})}{\text{Var}(S_{fg}^{\mathbf{e}_i}) + \text{Var}(S_{bg}^{\mathbf{e}_i})} \tag{3}$$

where $S_{fg}^{\mathbf{e}_i}$ and $S_{bg}^{\mathbf{e}_i}$ are the sets of values projected into 1D space spanning the eigenvector $\mathbf{e}_i$.

According to our experiment with this measure, PCA-based feature extraction is equivalent to or better than the method suggested in [10] in most cases as illustrated in Figure 3. Therefore, the subspace spanning $(\mathbf{e}_1, \cdots, \mathbf{e}_k)$ is practically a very good discriminative feature set.

3.2. Feature Value Weighting

Comaniciu [11] proposed assigning a weight term to each pixel in the target in proportion to the confidence that it belongs to the foreground. In this section, we explain how to compute such a eight based on the likelihood image.

Typically in the likelihood image, the target in the original image is transformed into a bright homogeneous region. The large bright region in the target area should have large weight because many foreground pixels are spatially condensed in that area. On the other hand, it is not desirable to give a large weight to small bright areas since they are likely to be noise.

The likelihood image associated with the highest eigenvalue is used for weighting each pixel. In order to remove noise in that likelihood image and obtain more robust spatial information, a Gaussian filter is applied to the image. If $v(\mathbf{x}_i)$ is the feature value of pixel $\mathbf{x}_i$ after Gaussian filtering, the weight is given by using the well-known *sigmoid M*-estimator

$$\tau(\mathbf{x}_i) = \frac{1}{1 + \exp\left(-c \cdot v(\mathbf{x}_i)\right)} \tag{4}$$

where c is a constant. Note that the weight is determined with respect to the spatial information as well as the intensity of each pixel because small bright regions will disappear while pixels in the large bright areas will still have large weights by Gaussian filtering.

4. OBJECT TRACKING

The method suggested in section 3 is embedded in the mean-shift tracking algorithm. Let $\{\mathbf{x}_i^\star\}_{i=1,\ldots,n}$ be the pixel location of the target model centered at $\mathbf{0}$, and $b(\mathbf{x}_i^\star)$ the color index of the sample $\mathbf{x}_i^\star$. The density for the color u in the target histogram is given by

$$\hat{q}_u = C \sum_{i=1}^{n} k((||\mathbf{x}_i^\star||)^2)\delta[b(\mathbf{x}_i^\star) - u]\tau(\mathbf{x}_i^\star) \tag{5}$$

where $k(\cdot)$ is a profile function and δ is the Kronecker delta function. Also the normalization constant C is given by

$$C = \frac{1}{\sum_{i=1}^{n} k((||\mathbf{x}_i^\star||)^2)\tau(\mathbf{x}_i^\star)} \tag{6}$$

Similarly, the new candidate can be represented by

$$\hat{p}_u(\mathbf{y}) = C_h \sum_{i=1}^{n_h} k((||\frac{\mathbf{y} - \mathbf{x}_i}{h}||)^2)\delta[b(\mathbf{x}_i) - u]\tau(\mathbf{x}_i) \tag{7}$$

where the normalization constant C_h is

$$C_h = \frac{1}{\sum_{i=1}^{n_h} k((||\frac{\mathbf{y}-\mathbf{x}_i}{h}||)^2)\tau(\mathbf{x}_i)} \quad (8)$$

After finding the most discriminative feature in the current frame, we apply the same transformation to the next frame and track the object in the converted image. The target model in the original space should be transformed according to the extracted feature space in each step, and the mean-shift algorithm is employed to track the object. When the threshold th is 0.7, the dimensionality of the feature space used in the tracking is not more than 3 in most cases in our experiments.

Two different sequences are tested, and tracking results are presented in Figure 4.

(a) airport sequence (frame 50, 100, 150, 200, 250, 300)

(b) tank sequence (frame 30, 60, 90, 120, 150, 180)

Fig. 4. Tracking results

5. DISCUSSION

In this paper, we proposed the adaptive feature extraction method based on likelihood images in various feature subspaces. This idea makes it straightforward to combine multiple heterogeneous visual features and find a better feature space for tracking. Also, the weight is assigned to each pixel considering the likelihood ratio and the spatial information of the pixel. We showed this can improve tracker performance, especially in low resolution video and in a very noisy environment.

However, there are currently several limitations of this approach. First of all, some visual information in the original image can be lost in the likelihood image during the projection, even if this problem is mitigated by various projections as proposed in section 2. The variance ratio is used to evaluate the distinctiveness between foreground and background, but a more thorough investigation of its validity is required.

Currently, only color information is used; the performance of the tracker could be much better if other visual features such as motion and texture could be integrated into this framework.

6. REFERENCE

[1] D. Comaniciu, V. Ramesh, and P. Meer, "Real-time tracking of non-rigid objects using mean shift," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition,* Hilton Head, SC, June 2000, vol. II, pp. 142–149.

[2] C. Stauffer and W.E.L. Grimson, "Learning patterns of activity using real-time learning," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, no. 8, pp. 747–757, 2000.

[3] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "Pfinder: Real-time tracking of the human body," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, pp. 780–785, 1997.

[4] M. Isard and A. Blake, "Condensation - Conditional density propagation for visual tracking," *Intl. J. of Computer Vision*, vol. 29, no. 1, 1998.

[5] P. Perez, C. Hue, J. Vermaak, and M. Gangnet, "Color-based probabilistic tracking," in *Proc. European Conf. on Computer Vision,* Copenhagen, Denmark, 2002, vol. I, pp. 661–675.

[6] Y. Rui and Y. Chen, "Better proposal distributions: Object tracking using unscented particle filter," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition,* Kauai, Hawaii, 2001, vol. II, pp. 786–793.

[7] M. Swain and D. Ballard, "Color indexing," *Intl. J. of Computer Vision*, vol. 7, no. 1, pp. 11–32, 1991.

[8] F. Ennesser and G. Medioni, "Finding Waldo, or focus of attention using local color information," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 17, no. 8, pp. 805–809, 1995.

[9] H. Stern and B. Efros, "Adaptive color space switching for face tracking in multi-colored lighting environment," in *Proc. IEEE Int'l Conf. on Automatic Face and Gester Recognition* Washington DC, USA, 2002, pp. 249–254.

[10] R. Collins and Y. Liu, "On-line selection of discriminative tracking features," in *Proceedings of the 2003 International Conference of Computer Vision (ICCV '03)*, October 2003.

[11] D. Comaniciu, "Kernel-based object tracking," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 25, no. 5, pp. 564–577, 2003.

[12] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller, "Fisher discriminant analysis with kernels," in *Neural Networks for Signal Processing IX*. 1999, pp. 41–48, IEEE.

[13] B. Schölkophand, A. Smola, and KR. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, pp. 1299–1319, 1998.