PATCH-BASED ADAPTIVE TRACKING USING SPATIAL AND APPEARANCE INFORMATION

Junqiu WANG and Yasushi YAGI

The Institute of Scientific and Industrial Research, Osaka University
8-1 Mihogaoka, Ibaraki, Osaka, 560-0047

ABSTRACT

We present a patch-based tracking algorithm in which both appearance and spatial information are taken into account for target localization. We decompose a target into several patches based on appearance similarity and spatial distribution. Each patch has its distinctive appearance and spatial distribution. Appearance information is described by kernels which are non-parametric; while spatial information is represented by spatial Gaussians. The overall motion is estimated by mean shift algorithm. The motion is refined based on the likelihood images computed using pixel classification. The proposed tracker provides better position and likelihood images.

Index Terms— Visual tracking, patch-based, feature selection, spatial and appearance information.

1. INTRODUCTION AND RELATED WORK

Visual tracking is still an unsolved problem after the intensive investigation over the years. Adaptive tracking [5] can improve the performance of a tracker by adaptively selecting features that make the target discriminative against its background. Unfortunately, adaptive tracking may fail when model drifts [5, 8]. Pixels in the background can be mistakenly labeled as foreground and incorporated into the target model. Thus the target representation deviates from the true appearance. In this work, we aims at improving pixel classification and target localization by explicitly considering appearance and spatial information.

Representation and localization are the main issues to be tackled in designing a robust tracker. An ideal target representation can capture the essence of the target that is invariant to certain changes and discriminative enough to distinguish the target out. Moreover, it should be flexible enough to accommodate object variability due to different lighting conditions or viewpoints. Histograms and other non-parametric forms such as kernel density estimation have been widely used in visual tracking. The robustness of histograms to change in pose and shape has been employed in various tracking algorithms. The advantage of this method is achieved by discarding all spatial information. Spatial information of image patches, however, is important for discriminating the target and its background in many cases.

In order to make use of appearance and spatial information, we decompose a target into several patches based on the appearance similarity and spatial distribution. Both appearance and spatial information are employed in the proposed tracker: appearance information is described by kernels which are non-parametric; spatial information is represented by spatial Gaussians. The overall motion is estimated by the mean shift algorithm. Likelihood images are computed based on pixel classification using the patch-based representation. The motion can be refined based on the likelihood images.

Multi-cue plays a powerful role in human visual perception. While using multiple cues in detection or tracking, two sets of features should be complementary. Although the color cue is very important in detection and tracking, other types of cues such as shape can be very helpful especially when they are used jointly with colors. In this work, we select the best features from the shape and color cues. The shape cue is represented by gradient orientation histograms and the color cue is described by color histograms. We calculate color histograms in the RGB, the HSV spaces and the normalized rg space. The discriminative features are selected by evaluating the discriminative ability of each feature.

1.1. Related Work

Spatial information has been addressed in previous works with different representations. Birchfield and Rangarajan [3] propose a spatiogram-based tracking algorithm to make use of spatial information. They model a target using a histogram in which each bin is spatially weighted by the mean and covariance of the location of the pixels that contribute to that bin. The experimental results in [3] is not satisfying because the spatial information is not well described. Wang et al. [12] model the appearance of objects based on mixture of Gaussians in a joint spatial-color space. To initialize the tracking, they adopt EM approach which is time consuming. The normalized color rg and intensity are employed to describe the appearance of the target. Although rg are robust to illumination changes, they are not discriminative in many cases. In contrast, we use a feature selection procedure to find those

features discriminative against the background.

The importance of feature selection has been noticed by Collins et al. [5]. Yin and Collins [15] extend [5] and propose a spatial divide and conquer approach which arbitrarily subdivides foreground and background into smaller regions. Different features are selected based on the separability between the target and one of its sub-background regions. One of the drawbacks of their work is the assumption that a target can be represented by a uni-modal distribution.

Our work also relates to Avidan's work [1] in which an ensemble of simple weak classifiers is used for the binary foreground/background appearance model maintenance and tracking. Each weak classifier is trained online from a specific frame, and the ensemble is collected from a predefined range of recent frames.

The paper is organized as follows. In Section 2, we introduce how to decompose and represent a target into patches. Section 3 describes feature selection for appearance modeling. We proposed the method for target localization and likelihood ratio image re-computation in Section 4. The performance of the proposed method is evaluated in Section 5. This paper is concluded in Section 6.

2. TARGET DECOMPOSITION AND REPRESENTATION

2.1. Decomposition

The initialization can be done by combining a detection algorithm and the GrabCut [10]. The detection algorithm provides a bounding box for the target. The GrabCut algorithm segments the target out using an iterative procedure [10].

The segmented target is represented by a collection of patches. These patches are generated using the k-means algorithm. The k-means algorithm is simple but effective in clustering different distributions. We have tried different color spaces such as RGB, HSV, rg in the k-mean clustering. Since the clustering based on color information is not satisfying enough, spatial information is also embedded into the k-means. The best results are achieved using HS-XY. The target decomposition results are illustrated in Fig. 1.

The number of patches has to be defined before the decomposition when using k-means. However, different targets might have different numbers of patches. In this work, the problem is dealt with by giving a relatively large number of patches in the very beginning (e.g., 8). The patches computed by running the k-means are refined based on their sizes. The pixels in these small patches (less than 5% of the target) are assigned to large patches according to their distance the centers of these patches.

The results of target decomposition are a collection of patches with appearance similarity and spatial adherence. We use $\mathcal{G}$ to denote a collection of patches sampled from the target region; and g denote a patch in the collection. The target

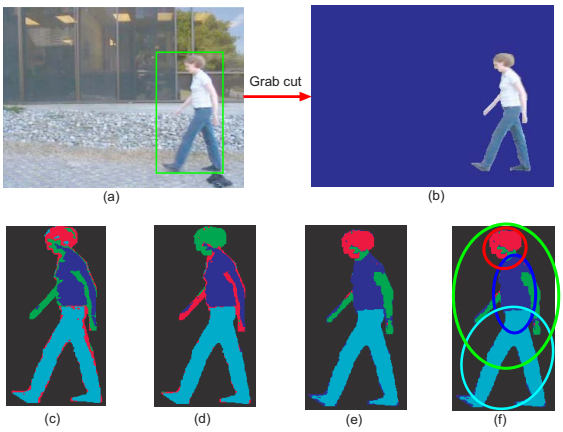

Fig. 1. (a) The input image. (b) Target segmentation using GrabCut [10]. (c) K-means-based target descomposition using RGB; (d) HSV; and (e) HS-XY. (f) Parametric representation of spatial information.

is described by its appearance which is a kernel [6, 13] in this work. In addition, each part is described separately with appearance and spatial information.

2.2. Representation

The target and its patches are described by their appearance and spatial information. Appearance information is described by kernels [6]. Spatial information in each part is represented by a spatial Gaussian. The spatial Gaussian of patch g is composed of its average μ_g and covariance matrix $\sum_g$. The spatial information of each part is also shown in Fig. 1.

3. FEATURE SELECTION

Feature selection finds the best subset from the features available for tracking or recognition. It plays an important role in tracking [5]. A feature can be selected or discarded based upon some predefined criteria such as principal component analysis [7], class separability measure [9], or variable ranking [5]. In this work, variance ratio is used as the discriminative ability measure.

3.1. Computing Likelihood Ratios

Variance ratio [5] measures the separability of a feature based on likelihood ratios. Likelihood ratios map raw feature values nonlinearly into a new feature space. Those values appearing more often on the target are projected to positive values; and values appearing more frequently on the background are projected to negative values. Log-likelihood ratios can be computed using the histograms of foreground and background with respect to a given feature. The frequency of the pixels appeared in a histogram bin is calculated as $\zeta_F^{(b_{in})} = \frac{p_F^{(b_{in})}}{n_F}$

and $\zeta_B^{(b_{in})} = \frac{p_B^{(b_{in})}}{n_B}$, where n_F is the pixel number of the target region and n_B the pixel number of the background.

The log-likelihood ratio for a feature value *i* is given by

$$L^{(b_{in})} = \max(-1, \min(1, \log \frac{\max(\zeta_F^{(b_{in})}, \delta_L)}{\max(\zeta_B^{(b_{in})}, \delta_L)})), \quad (1)$$

where δ_L is a very small number.

3.2. Selecting Discriminative Features

The variance ratio of the likelihood function is defined as [5]:

$$\nu_g = \frac{\text{var}(B \cup F)}{\text{var}(F) + \text{var}(B)} = \frac{\text{var}(L; (p_F + p_B)/2)}{\text{var}(L; p_F) + \text{var}(L; p_B)}. \quad (2)$$

We evaluate the discriminative ability of each feature by calculating the variance ratio. In the candidate feature set, the color cue includes 7 different features: color histograms of R, G, B, H, S, r, and g; and the shape cue includes gradient orientation histogram. These features are ranked according to the discriminative ability by comparing the variance ratios. The feature with the maximum variance ratio is taken as the most discriminative feature.

4. TARGET LOCALIZATION AND LIKELIHOOD RATIO IMAGE COMPUTATION

The global motion is estimated using the mean shift algorithm. The mean shift algorithm is an efficient peak mode seeking method which can approach the peak mode in adaptive steps. The efficiency is employed to estimate the global motion of a target. The global motion estimation is similar to that in [6]. However, we adopt the idea of [14] in which discriminative features are selected from multiple cues. Although mean shift tracking is efficient, it does not provide good localization results in many cases (Fig. 2). We try to classify pixels in the bounding box into foreground and background using the appearance and spatial information of each patch. The probabilities that a pixel belongs to foreground or background are combined together to make a better likelihood ratio image. We compute the location of the target based on the refined likelihood image, which is re-localization of the target.

4.1. Pixel Classification

We use superscripts F and B to denote foreground and background; A and S denote appearance and spatial information respectively. The pixel classification is formulated following the Bayesian approach. $r_{\mathbf{x}}^g$ is the likelihood ratio of a pixel at $\mathbf{x}$ in patch g, it is computed by:

$$r_{\mathbf{x}}^g = \frac{p_{\mathbf{x}}^g(F|S, A)}{p_{\mathbf{x}}^g(B|S, A)} \quad (3)$$

It is difficult to compute an exact solution to Eq. 3. However, it can be approximated as:

$$r_{\mathbf{x}}^g \approx \frac{p_{\mathbf{x}}^g(S, A|\theta_F) p_{\mathbf{x}}^g(F)}{p_{\mathbf{x}}^g(S, A|\theta_B) p_{\mathbf{x}}^g(B)}, \quad (4)$$

$$p_{\mathbf{x}}^g(S, A|\theta_B) = p_{\mathbf{x}}^g(A|S, \theta) p_{\mathbf{x}}^g(S|\theta) p_{\mathbf{x}}^g(\theta)), \quad (5)$$

where θ_F and θ_B are parameters of the foreground and background.

The probability resulted from appearance distribution is computed using histogram back-projection method [11, 14]. The probability resulted from spatial distribution is computed as:

$$p_{\mathbf{x}}^g(S|\theta) = \frac{\exp(-\frac{1}{2}(\mathbf{x} - \mu_g)^T (\Sigma_g^S)^{-1} (\mathbf{x} - \mu_g))}{2\pi |\Sigma_g^S|^{\frac{1}{2}}}. \quad (6)$$

We computed likelihood images for each patch in the patch collection. These likelihood images are combined together to obtain the foreground likelihood image.

4.2. Merging Likelihood Ratios

Each patch has been described by different discriminative features. These features are discriminative in certain areas that are described by the spatial Gaussians. The likelihood ratios obtained in the previous subsection have different contributions according to the discriminative abilities. The merging of multiple likelihood ratio images considers these discriminative abilities. It is computed using:

$$r(\mathbf{x}) = \sum \nu_g r^g(\mathbf{x}) p_{\mathbf{x}}^g \quad (7)$$

where ν_g is the variance ration score of patch g; $p_{\mathbf{x}}^g$ is computed using Eq. 6.

4.3. Target Re-localization

The merged likelihood ratio image is better than that provided by the global mean shift. We estimate the location and bounding box of the target based on the merged likelihood ratio image. The re-localization can give a better location of the target. In practice, this is achieved by running another mean shift on the merged likelihood ratio image [4].

5. EXPERIMENTAL RESULTS

The tracking algorithm was tested on image sequences with ground truth. The testing results of two sequences are shown here due to the space limitation. Both of them were captured by moving cameras. The images in these sequences have a size of 360 $\times$ 240 pixels. The low resolution of the images makes the likelihood images computation difficult. Other factors also contribute to the difficulty such as the articulated target structures, the deformable property of the targets and the dynamic backgrounds.

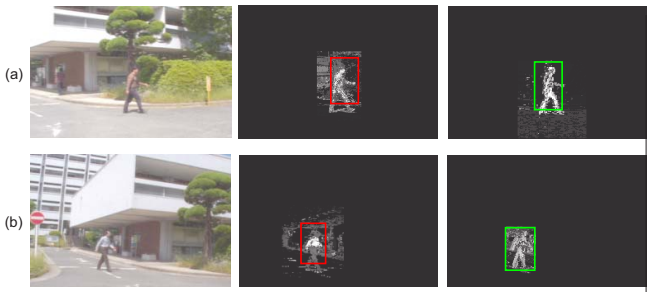

Fig. 2. The two sequences used in the experiment. The images in the first column are two frames in the input sequences. The second column shows the likelihood ratio images computed also using the global description of the target. The bounding boxes in the second column are estimated using mean shift with global description. The third column shows the merged likelihood ratio images and the bounding boxes given by re-localization.

The two sequences used in the experiments are shown in Fig. 2(a) and Fig. 2(b) respectively. In Fig. 2, the likelihood ratio images computed using the global description of the target are illustrated in the middle column. In Fig. 2(a), some pixels in the background have high probability of foreground. In Fig. 2(b), the upper body of the target is well described by the likelihood ratio image due to the distinctiveness of its color. Other parts of the target are not well reflected in the likelihood ratio image. The bounding boxes deviate away from the locations of the targets. We compute likelihood ratio images for each patch using their appearance and spatial information. Then these likelihood ratios are merged based on their distinctiveness. The merged likelihood ratio image are shown in the right column in Fig. 2. We shift the bounding box to a new position based on the merged likelihood ratio image. The bounding boxes estimated are better than those given by the overall description in the middle column in Fig. 2.

We evaluate the performance of the proposed algorithm quantitatively using the image sequences with ground truths. The ground truths are gotten by labeling the images manually into foreground and background. We threshold the likelihood ratio images and compare them with the ground truths. The comparison results are shown in Fig. 3. The error rates of the merged likelihood images are lower than that of the direct back-projection of the one histogram description in most cases. However, the merged likelihood images are worse than the direct back-projection in certain frames. We are investigating the reason of the poor performance in these frames.

6. CONCLUSION AND FUTURE WORK

We devise a patch-based adaptive tracking algorithm. The target decomposition is effective in improving the performance of the tracking. The proposed algorithm provides better target localization and likelihood images. These results lay a foundation for foreground segmentation using graph cuts or other methods. They are also useful in target model updating to avoid drifts.

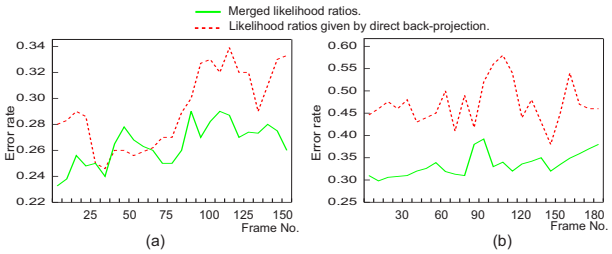

Fig. 3. Comparison of likelihood ratio images.

7. REFERENCES

[1] S. Avidan. Ensemble tracking. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(2):261–271, 2007.

[2] S. Birchfield and S. Rangarajan. Spatiograms versus histograms for region-based tracking. In *Proc. of CVPR*, pages 1158–1163, 2005.

[3] G. Bradski. Computer vision face tracking as a component of a perceptural user interface. In *Proc. of the IEEE Workshop Applications of Computer Vision*, pages 214–219, 1998.

[4] R. Collins, Y. Liu, and M. Leordeanu. On-line selection of discriminative tracking features. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(10):1631–1643, 2005.

[5] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(5):564–577, 2003.

[6] B. Han and L. Davis. Object tracking by adaptive feature extraction. In *Proc. of the IEEE Conf. on Image Processing*, pages 1501–1504, 2004.

[7] A. Jepson, D. J. Fleet, and T. EI-Maraghi. Robust online appearance models for visual tracking. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(10):1296–1311, 2003.

[8] H. Nguyen and A. Smeulders. Robust tracking using foreground-background texture discrimination. *Int'l Journal of Computer Vis.*, 69(3):277–293, 2006.

[9] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph. (TOG)*, 23(3):309–314, 2004.

[10] M. Swain and D. Ballard. Color indexing. *Int'l Journal of Computer Vis.*, 7:11–32, 1991.

[11] H. Wang, D. Suter, and K. Schindler. Effective appearance model and similarity measure for particle filtering and visual tracking. In *Proc. of ECCV*, pages 606–618, 2006.

[12] J. Wang and Y. Yagi. Discriminative Mean Shift Tracking with Auxiliary Particles. *Proc. 8th Asian Conference on Computer Vision*, pages 576–585, 2007.

[13] J. Wang and Y. Yagi. Integrating color and shape-texture features for adaptive real-time object tracking. *IEEE Trans. on Image Processing*, 17(1), 2008.

[14] Z. Yin and R. Collins. Spatial divide and conquer with motion cues for tracking through clutter. In *Proc. of CVPR*, pages 570–577, 2006.