Fragments based Parametric tracking

Prakash C, Balamanohar Paluri, Nalin Pradeep S, and Hitesh Shah[1]

Sarnoff Innovative Technologies Pvt Ltd

Abstract. The paper proposes a parametric approach for color based tracking. The method fragments a multimodal color object into multiple homogeneous, unimodal, fragments. The fragmentation process consists of multi level thresholding of the object color space followed by an assembling. Each homogeneous region is then modelled using a single parametric distribution and the tracking is achieved by fusing the results of the multiple parametric distributions. The advantage of the method lies in tracking complex objects with partial occlusions and various deformations like non-rigid, orientation and scale changes. We evaluate the performance of the proposed approach on standard and challenging real world datasets.

1 Introduction

Two prominent components of a tracking system are: object descriptor and search mechanism. Object descriptor is the representation of the object to be tracked using a set of features that capture various properties of the object such as the appearance, shape, texture etc. Given an object descriptor, the search mechanism like [1, 2], locates the region in a new image that best matches the object description.

There are multiple methods suggested in the literature for object descriptors. Most of the successful methods for tracking employ non-parametric object descriptor like histogram [1, 3–7], as it faithfully captures the variability in the features of the object to be tracked. However, with the increase in number of objects to be tracked or the features to be considered, the histogram size grows exponentially which is an undesired behavior. To address this issue, we propose a parametric object descriptor for color based tracking.

An N-dimensional Gaussian distribution is employed as the object descriptor in the proposed approach. Such a descriptor can accurately model a unimodal object. But objects under consideration for tracking are generally multimodal in color space making N-d Gaussian descriptor insufficient. Hence, we need to convert multimodal objects to unimodal representation. Primarily, there are two ways to achieve this conversion:

- By projecting the multimodal object into a space where it becomes unimodal
- By representing each mode separately

An approach similar to Collins et al [5] can be used to find a linear transformation to project the multimodal color object into a unimodal space. Also,

non-linear transformation as suggested by Larry et al [7] can be used to the same effect. However, in both the cases, the search for finding such an optimal transformation is not exhaustive, as it is computationally expensive, over the entire space of possible transformations. Hence the obtained transformation is suboptimal. For representing a multimodal object, Gaussian mixture model (GMM) distributions can also be used. However, computation of GMM parameters are expensive and an aprior knowledge of number of modes is quintessential, thus rendering it is not applicable for object descriptors in tracking.

Therefore, in this paper we propose a method based on fragmenting multimodal objects into multiple homogeneous models using Discriminant Analysis. The fragmentation process finds the fragments online as opposed to fragmenting the object into fixed sizes as suggested in [4]. Each fragment is then modelled using a single N-dimensional Gaussian distribution and tracked separately. These parametric distributions are used to generate a probability density function termed as strength image. The maximum likelihood(ML) framework proposed in [2] is used to estimate the location(mean) and shape(covariance) of the best matching region in the subsequent frame.

The paper is organized as follows: Section 2 explains the proposed approach. Experimental results are presented in Section 3 to illustrate the performance of the tracker and Section 4 concludes the work by presenting the future work.

2 Proposed Method

The proposed tracking approach is color based, hence in modelling an object we use the color values of the pixel. Our initial step involves fragmenting based on the color values of the pixels. Prior work involved application of multi-level thresholding technique to segment an illumination/gray image [8]; but these techniques can't be applied directly in our case as the our objective is to group regions similar in color rather than illumination (gray). Hence, multi-level thresholding is done in color space. The input template is in the RGB space. Multi level thresholding on the histogram generated using all three channels is not possible due to the immense size of the histogram ($256 \times 256 \times 256$). So, given the color template of the object in the RGB space, we first transform the input to HSV space. Since Hue represents the color component alone, multi-level thresholding on Hue gives the desired results. The grouped regions similar in color are then modelled using single parametric distribution. The Uni/Multi modal classification and the fragmentation processing use the Hue image for processing.

2.1 Fragmentation

The given region of interest (ROI) is initially divided into uniform blocks of size $M \times N$. Each block(B) with mean(μ), variance(σ) and Hue histogram(H) is then classified as homogeneous if any of the following two criterions satisfy:

1. If the variance of the region is less than a certain pre-defined value. The variance of the region is given by

$$\sigma = \sum_{i \epsilon B} (P_i - \mu)^2 \tag{1}$$

where P_i is the hue value at i and the μ is the mean of the block.

2. If the block is divided into two classes C1 with values $[1, \ldots, t]$ and C2 with values $[t+1, \ldots, L]$ using the optimal threshold given by eqn 2 and if the Separability factor(SF) given by eqn 3 of the block is less than a certain pre-defined value.

$$\arg\max_t \sum_{i=1}^{N} W_i (\mu_i - \mu)^2 \tag{2}$$

where N is the number of classes, W_i is the total number of pixels in the class i, μ_i is the mean of the i^th class and μ is the mean of the block.

$$SF = \frac{BCV}{TV} \tag{3}$$

where TV is the total variance of the block given by 1 and BCV is the between class variance given by:

$$BCV = \sum_{i=1}^{N} W_i (\mu_i - \mu)^2 \tag{4}$$

The fragmentation process is applied only for non-homogeneous regions. The multi-level thresholding is carried until the SF of the block is less than a predefined valueTh_{SF}. The TV of the region is constant and is used for normalizing purposes. The BCV will be high when the fragments of similar color are grouped and dissimilar colors are separated.The multi level thresholding is done in the following way:

Each time the fragment/class with the maximum within class variance is selected(Initially, the entire block is started as one class), since high within class variance signifies that the class is non-homogeneous. The division is done by finding the Optimal threshold given by 2. The process is repeated till the SF of the block is less than Th_{SF}.

The class pool thus created needs to be assembled together based on the color similarity. The assembled regions will signify the multiple unimodal regions of the multimodal object.

The assembling process is started with a new region which includes the first class of the first region. This is followed by a merging process which finds out the classes similar to this class. The criterion for similarity is the difference of the mean values of the two classes. The class with the least difference is identified and if the difference of the means is less than a pre-defined value Th_{mean} the class is merged into the region. If none of the classes can be merged to any of the existing regions, a new region is created by picking up the class which has the largest difference in the mean value with the existing regions. Then the unclassified classes are again tried to merge to this new region. The process is

repeated until all the classes are merged to the regions. The regions obtained thus form the unimodal fragments of the multimodal object. An example of the fragmentation is shown in Figure 1.

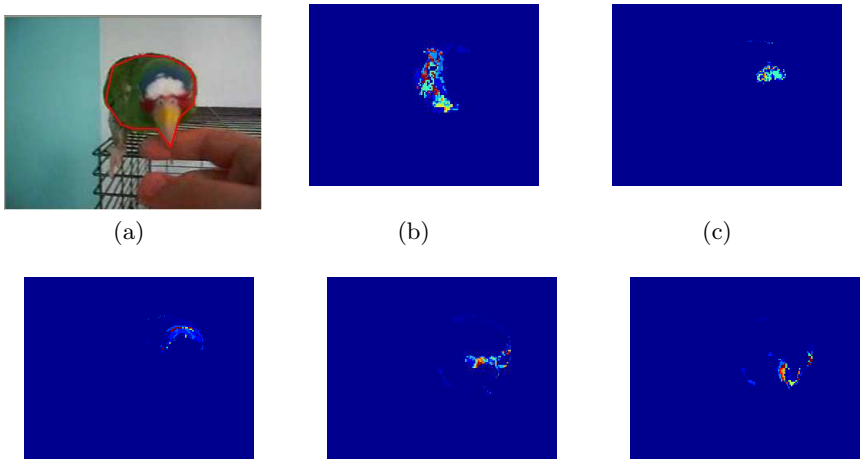


Fig. 1. Parrot sequence: The input image (a) is fragmented into five parts. (b) represents the body of the parrot(green), (c) represents the forehead(white), (d) represents the hair(blue), (e) represents the cheeks(red) and (f) represents the beak(yellow)

2.2 Modelling the Object

Each fragment obtained after the fragmentation process is modelled separately. Each region $\mathcal{R} \subset \mathcal{R}egions$ is described by the color values $\{R, G, B\}$, thus the feature descriptor at an image location $\mathbf{x} = \begin{bmatrix} x & y \end{bmatrix}^t$ is computed as $\mathbf{f}(\mathbf{x}) = \begin{bmatrix} R(\mathcal{I}, \mathbf{x}) & G(\mathcal{I}, \mathbf{x}) & B(\mathcal{I}, \mathbf{x}) \end{bmatrix}^t$. The *region covariance*, $\mathbf{C}$ of the feature descriptors in $\mathcal{R}$ is computed as

$$\mathbf{C} = \frac{1}{|\mathcal{R}|} \sum_{\mathbf{x} \in \mathcal{R}} (\mathbf{f}(\mathbf{x}) - \boldsymbol{\mu}\ )(\mathbf{f}(\mathbf{x}) - \boldsymbol{\mu}\ )^t \tag{5}$$

where $\boldsymbol{\mu} = \frac{1}{|\mathcal{R}|} \sum_{\mathbf{x} \in \mathcal{R}} \mathbf{f_x}$ is the mean feature descriptor in $\mathcal{R}$ and $|\mathcal{R}|$ is the number of pixels in the region $\mathcal{R}$. A simple covariance matrix computed with color features contains the information needed to capture the appearance of the object. An estimate of the color distribution in the target region is the Gaussian distribution. The ML estimates of the parameters of the Gaussian, $\boldsymbol{\Theta}$=($\boldsymbol{\mu}$, $\mathbf{C}$) is the target model. The probability density function (PDF), also termed as the

strength image, is computed over the new image. The value of each pixel in the strength image signifies the probability with which the pixel belongs to the target model. In the remainder of this paper we denote this value as $p(\mathbf{x}|\boldsymbol{\Theta})$ where $\mathbf{x}$ is the pixel location. The PDF in this case is computed as :

$$p(\mathbf{x}|\boldsymbol{\Theta}) \propto exp(-(\mathbf{f}(\mathbf{x}) - \boldsymbol{\mu})^t \mathbf{C}^{-1}(\mathbf{f}(\mathbf{x}) - \boldsymbol{\mu})) \tag{6}$$

The PDF is calculated for each of the unimodal regions of the object obtained through the fragmentation process. The PDF will have high values for pixels which belong to the particular parametric distribution and vice-versa. In the next section, we show how the PDF computed for the image can be used to track the region accurately in presence of various deformations.

2.3 ML Framework

The region to be tracked $\mathcal{R}_0$ is represented by an ellipse in our case. The position and shape of the object are described by the mean $\mathbf{M}_0$ and covariance $\mathbf{V}_0$ of the pixels in the region. Given the target model $\boldsymbol{\Theta}$, the objective of the search mechanism is to find a region $\mathcal{R}$ in the new frame described by mean and covariance $(\mathbf{M}, \mathbf{V})$ that maximize the function :

$$J(\mathbf{M}, \mathbf{V}) = \sum_{\mathbf{x} \in \mathcal{R}} p(\mathbf{x}|\boldsymbol{\Theta}) L(\mathbf{x}|\mathbf{M}, \mathbf{V}) \tag{7}$$

where the term

$$L(\mathbf{x}|\mathbf{M}, \mathbf{V}) \propto exp(-(\mathbf{x} - \mathbf{M})^t \mathbf{V}^{-1}(\mathbf{x} - \mathbf{M})) \tag{8}$$

prevents pixel locations that are farther from the original region from distracting the tracker. As a pixel's contribution falls off with the distance from the original region, this helps in both reducing the effect of outlier pixel on the search as well as preventing the tracker from *drifting* away from the object.

As shown in [2, 9, 10], the maximum-likelihood estimates of $\mathbf{M}$ and $\mathbf{V}$ can be obtained via an EM-like iterative procedure. The key to the method is to assume a set of hidden variables $w(\mathbf{x})$. Starting with an initial estimate $\mathbf{M}^0, \mathbf{V}^0$ of $\mathcal{R}$, the EM-iteration proceeds as below:

- **E-Step:** Given current estimates $\mathbf{M}^k$ and $\mathbf{V}^k$ of the mean and covariance of the region in k^{th} iteration, compute hidden variables $w^k(\mathbf{x})$:

$$w^k(\mathbf{x}) = \frac{p(\mathbf{x}|\boldsymbol{\Theta}) L(\mathbf{x}|\mathbf{M}^k, \mathbf{V}^k)}{\sum_{\mathbf{x}' \in \mathcal{R}} p(\mathbf{x}'|\boldsymbol{\Theta}) L(\mathbf{x}'|\mathbf{M}^k, \mathbf{V}^k)} \tag{9}$$

- **M-Step:** Using the hidden variables computed above, compute the next estimates of mean and covariance of the region, $\mathbf{M}^{k+1}$ and $\mathbf{V}^{k+1}$ of that maximize $J(.,.)$:

$$\mathbf{M}^{k+1} = \sum_{\mathbf{x} \in \mathcal{R}} w^k(\mathbf{x}) \mathbf{x} \tag{10}$$

$$\mathbf{V}^{k+1} = \sum_{\mathbf{x} \in \mathcal{R}} w^k(\mathbf{x}) (\mathbf{x} - \mathbf{M}^{k+1})(\mathbf{x} - \mathbf{M}^{k+1})^t \tag{11}$$

The optimal values for $\mathbf{M}$ and $\mathbf{V}$ are obtained by iterating the above steps until convergence. Our experimental results demonstrate that the search mechanism described above is both efficient and robust to a wide variety of changes in the shape of the object. In Algorithm 1, we explain the complete tracking algorithm.

Algorithm 1 Track(Video V, Region R_0)

$I_0 \leftarrow$Initial_Frame(V)
$(\mathbf{M}_0, \mathbf{V}_0) \leftarrow$ *Fit_Ellipse*($\mathcal{R}_0$)
$\boldsymbol{H} \leftarrow$ *HSV*($\mathcal{R}_0$)
Class_Pool $\leftarrow$ *Multi_Level_Thresholding*(H)
Fragments $\leftarrow$ *Assembling*(*Class_Pool*)
$\boldsymbol{\Theta} \leftarrow$ *Region_Covariance*($\mathcal{R}_0$, *Fragments*)
for *each frame* I_i *in* V ***do***
  $(\mathbf{M}_i, \mathbf{V}_i) \leftarrow (\mathbf{M}_{i-1}, \mathbf{V}_{i-1})$
  $S \leftarrow$ *Strength_Image*(I, $\boldsymbol{\Theta}$)
  $k \leftarrow 0$
  while *not converged* ***do***
    compute weight w^k *using equation 9*
    update estimates $(\mathbf{M}_i, \mathbf{V}_i)$ *using equations 10, 11*
    $k \leftarrow k + 1$
  end while
end for

3 Experimental Results

The tracking algorithm was tested on various challenging datasets [11]. It was also tested on few low contrast videos taken from Internet. The tracker performance was encouraging when tested for its ability to handle the following aspects:

Non-rigid deformations: Tracking non-rigid objects is a challenging problem in tracking. Couple of examples are highlighted in Figures 2, 4. In Cat sequence (Figure 2), the cat is tracked accurately under considerable deformations (sitting,jumping and running). Also, note that the contrast between the cat and the background is quite low. In case of Figure 4, the monkey is tracked successfully under extreme deformations. In both the cases, the tracked ellipse changes accurately to handle the non-rigid deformations of the object.

Orientation: Change in the orientation of objects is a common scenario in tracking. Tracking the object with accurate orientation is possible in our case since we track the object using an ellipse. The mean of the ellipse characterizes the location and the covariance signifies the scale and orientation. In Figure 4(b,c) ,the monkey undergoes considerable changes in orientation. The orientation of the ellipse changes according to the orientation of the object. Figure 3

is another example where the fish is tracked accurately in the presence of rapid orientation changes.

(a) (b) (c)

Fig. 2. Cat sequence: The cat is tracked successfully in presence of changes in scale and non-rigid deformations.

(a) (b) (c)

Fig. 3. Fish sequence: An example of low quality video containing partial occlusions and orientation changes. Note the other fishes in the tank with the similar color(but not pattern) to the object being tracked. Many existing trackers fail in such cases

Scale: Earlier trackers relied on techniques such as searching through exhaustive search space [12] or using templates of the object at different scales [13]. In our case, the EM-like algorithm enables efficient handling of scale changes by estimating the covariance of the tracked ellipse. Figure 2, shows how the tracking handles scale changes.

Partial Occlusion: Parital and full occlusions occur frequently in tracking scenearios and the tracker needs to handle these sucessfully. Even if an object is completely occluded for considerable time, the tracker should be able to track the object on reappearance. A scenario with complete and partial occlusions is shown in the Figures 3, 5. Figure 3(b) shows the cases where the fish is partially occluded. In Figure 3(c), the fish reappears after being completely occluded and the tracker was able to relocate the object. On a standard dataset as in Figure 5, the ellipse fits the partially visible person where major portion of the person is occluded by two other people.

(a) (b) (c)

Fig. 4. Monkey sequence: Inspite of changes in orientation and non-rigid deformations, the monkey is tracked precisely.

(a) (b) (c)

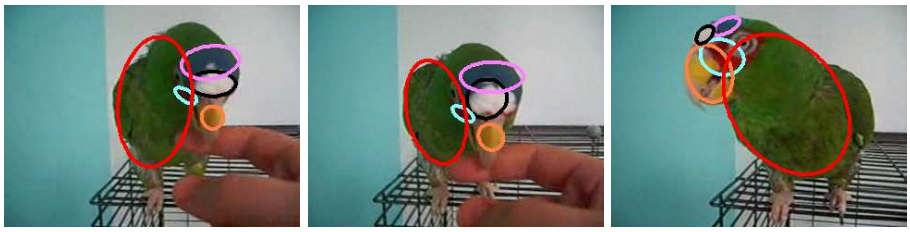
Fig. 5. Caviar sequence: The sequence shows the handling of partial occlusion of the person (blue ellipse) when he crosses two other people.

Handling Multimodal: Many of the datasets to be tracked has multimodal objects. We handle multimodal objects by fusing information from each unimodal. Figures 2, 3, 7 show tracking results on multimodal objects. For instance, Figure 7 shows the example of parrot, where each homogeneous region is extracted and modelled seperately as explained previously and Figure 6 shows the tracking of each unimodal region.

Videos with low quality and contrast: The quality of multimedia data available on the web varies significantly owing to various compression and transmission techniques. Several tests using videos with low quality and contrast were carried out to test our technique. In case of Figures 3,4 taken from googlevideos, the quality is poor owing to compression. In these videos, background color merges more with the color of the object. The tracker performance is very good and insensitive to these variations in video.

4 Conclusion

We have proposed a fragment based tracking approach in which the multimodal objects were fragmented into homegenous regions based on hue. These unimodal regions are then tracked using a single parametric distribution and these distributions are fused to form the final tracking result of the entire object. The tracker



(a) (b) (c)

Fig. 6. Parrot sequence: The multimodal object parrot is decomposed into multiple unimodals and tracked seperately

(a) (b) (c)

Fig. 7. Parrot sequence: The multimodal object is tracked sucessfully. Note that the ellipse completely fits the entire parrot enclosing all the homogenous regions.

proposed was also complimented with an efficient search mechanism to make the system robust to handle non-rigid deformations, occlusions, scale and orientation changes efficiently. For modelling the object more efficiently, the research is currently focused on combining other cues like motion,edge,texture with present color based tracker.

References

1. Comaniciu, D., Meer, P.: Mean shift analysis and applications. In: ICCV (2). (1999) 1197–1203
2. Z.Zivkovic, Krose, B.: An em-like algorithm for color-histogram-based object tracking. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition. (2004) 798–803
3. Comaniciu, D., Ramesh, V., Meer, P.: Real-time tracking of non-rigid objects using mean shift. In: IEEE Conf. on Comp. Vis. and Pat. (2000) 142–151
4. Adam, A., Rivlin, E., Shimshoni, I.: Robust fragments-based tracking using the integral histogram. In: CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, IEEE Computer Society (2006) 798–805
5. (Leordeanu, M., Collins, R.T., Liu, Y.)

6. Birchfield, S.T., Rangarajan, S.: Spatiograms versus histograms for region-based tracking. In: Proceedings of the Computer Vision and Pattern Recognition. Volume 2., Washington, DC, USA, IEEE Computer Society (2005) 1158–1163
7. Han, B., Davis, L.: Object tracking by adaptive feature extraction. In: In proceeding of International Conference on Image Processing. (2004)
8. Liao, P.S., Chen, T.S., Chung, P.C.: (A fast algorithm for multilevel thresholding)
9. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society. Series B (Methodological) **39**(1) (1977) 1–38
10. Neal, R.M., Hinton, G.E.: A new view of the EM algorithm that justifies incremental, sparse and other variants. In Jordan, M.I., ed.: Learning in Graphical Models. Kluwer Academic Publishers (1998) 355–368
11. project/IST 2001 37540, E.F.C.: found at url: http://homepages.inf.ed.ac.uk/rbf/caviar/ (2004)
12. Porikli, F., Tuzel, O.: Covariance tracking using model update based on means on riemannian manifolds. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition. (2006)
13. Birchfield, S.: Elliptical head tracking using intensity gradients and color histograms. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, IEEE Computer Society (1998) 232