MASTER

Multi-person tracking using a discriminative color appearance model

Chunduri, S.

Award date:
2012
MASTER THESIS

Multi-Person Tracking Using a Discriminative Color Appearance Model

Author
Siddharth Chunduri
(0754881) M.sc Embedded Systems
s.chunduri@student.tue.nl

Supervisor from Philips Research
Ir. Chris Damkat
chris.damkat@philips.com

Supervisor from TU/e
Prof. Dr. Ir. Gerard De Haan
G.d.Haan@tue.nl

August 15, 2012
Abstract

The focus of this thesis is computer vision related: it is to overcome the limitations of non-discriminative object trackers such as the CAMSHIFT algorithm that arise when tracking multiple objects. A new appearance model that improves the tracking of multiple objects by enhancing their discriminative color features has been developed. This model is formed by employing a supervised fuzzy c-means clustering in an effort to determine the differentiating color features. The tracking performance of the proposed method has been evaluated in comparison with the conventional histogram appearance model based CAMSHIFT algorithm for offline tracking of multiple targets. The discriminative color appearance model based tracker exhibits an improved performance in comparison with the CAMSHIFT algorithm, on being evaluated using an automated trajectory based metric.

Keywords: Video based object tracking, CAMSHIFT, Appearance models, Fuzzy C-Means Clustering.
I want to start by thanking Prof. Gerard de Haan for providing me the opportunity to work at Philips Research and facilitating me to work on a topic for which I possess a great deal of passion. I thank my supervisor and mentor at Philips Ir. Chris Damkat for showing extraordinary patience and faith in me. I am thankful to him for the freedom I had during my thesis and for always encouraging me to keep things simple. I could not have done it without his support. I thank my colleagues Erik Agterdenbos, Saurav Aran and Conil Aurilean for always keeping our office cheerful. I am greatly indebted to my best friend Puvvula Bhargava for always keeping me on my toes. Finally last but not the least I thank my Mom and Dad for their trust and love.
# Contents

1 Introduction ........................................... 1
   1.1 Problem description .................................. 1
   1.2 Related work ........................................ 2
   1.3 Report Organization .................................. 2

2 CAMSHIFT .............................................. 3
   2.1 CAMSHIFT components .................................. 3
      2.1.1 Object representation by use of an appearance model .... 3
      2.1.2 Appearance model backprojection .......................... 5
      2.1.3 Position Localization .................................. 5
   2.2 Adaptive appearance model ............................ 7
   2.3 Adaptive target size ................................... 7
   2.4 Algorithm ............................................ 7
   2.5 Implementation ....................................... 8
      2.5.1 Histogram appearance model ........................... 8
      2.5.2 Adaptivity of the appearance model and target area .... 9
      2.5.3 Extension to multi-person tracker ....................... 9
   2.6 Limitations of CAMSHIFT .............................. 9

3 Discriminative color appearance model based tracking .......... 11
   3.1 Description ........................................... 11
      3.1.1 Determination of the discriminative color appearance model .. 11
      3.1.2 Backprojection ...................................... 15
   3.2 Discriminatory color appearance model based tracking algorithm ... 16
   3.3 Extensions ............................................. 17
      3.3.1 Backprojection with modified responsibilities .......... 17
      3.3.2 Including background information in discriminatory color appearance model .. 20
      3.3.3 Extension to multiple centroids per object ............. 21
      3.3.4 Extension to multiple centroids per background .......... 21
      3.3.5 Implementation details ................................ 21

4 Performance evaluation .................................. 23
   4.1 Evaluation metric ..................................... 23
   4.2 Dataset .............................................. 24
   4.3 Evaluation ............................................ 26
      4.3.1 Backprojection evaluation ............................... 26
      4.3.2 Effect of Background model ............................ 28
      4.3.3 Effect of multiple centroids per object or background .... 29
      4.3.4 Performance comparison of discriminatory color appearance model based tracking with CAMSHIFT .......................... 31
   4.4 Computational complexity of the discriminative color appearance model based tracker compared to CAMSHIFT ......................... 32

5 Conclusions .......................................... 33
   5.1 Future Work .......................................... 34
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>An object being tracked and its corresponding color histogram appearance model.</td>
<td>4</td>
</tr>
<tr>
<td>2.2</td>
<td>An object and its corresponding search area’s histogram backprojection.</td>
<td>5</td>
</tr>
<tr>
<td>2.3</td>
<td>Localized object in backprojected search area and original image.</td>
<td>6</td>
</tr>
<tr>
<td>2.4</td>
<td>Block diagram of CAMSHIFT algorithm.</td>
<td>8</td>
</tr>
<tr>
<td>2.5</td>
<td>Limitation of CAMSHIFT in multi-person tracking.</td>
<td>9</td>
</tr>
<tr>
<td>3.1</td>
<td>Discriminatory color appearance model</td>
<td>13</td>
</tr>
<tr>
<td>3.2</td>
<td>Result of discriminatory color appearance model</td>
<td>14</td>
</tr>
<tr>
<td>3.3</td>
<td>Responsibility back projection</td>
<td>15</td>
</tr>
<tr>
<td>3.4</td>
<td>Block diagram of discriminatory appearance model based tracking algorithm</td>
<td>17</td>
</tr>
<tr>
<td>3.5</td>
<td>Objects being tracked.</td>
<td>18</td>
</tr>
<tr>
<td>3.6</td>
<td>Backprojection using original responsibilities.</td>
<td>19</td>
</tr>
<tr>
<td>3.7</td>
<td>Responsibility modification for improved backprojection</td>
<td>19</td>
</tr>
<tr>
<td>3.8</td>
<td>Responsibility backprojections without and with a background model</td>
<td>20</td>
</tr>
<tr>
<td>3.9</td>
<td>Hue color wheel.</td>
<td>22</td>
</tr>
<tr>
<td>4.1</td>
<td>Sample frame and objects being tracked from the self made video.</td>
<td>24</td>
</tr>
<tr>
<td>4.2</td>
<td>Sample frames and objects being tracked from the CAVIAR and i-Lids dataset.</td>
<td>25</td>
</tr>
<tr>
<td>4.3</td>
<td>Average of tracking accuracies for objects 1 and 2 over the entire dataset using different backprojection approaches.</td>
<td>27</td>
</tr>
<tr>
<td>4.4</td>
<td>Combined tracking accuracies using the discriminatory color appearance model with and without background information.</td>
<td>29</td>
</tr>
<tr>
<td>4.5</td>
<td>Combined tracking accuracies for extensions enabling multiple centers per object / background.</td>
<td>30</td>
</tr>
<tr>
<td>4.6</td>
<td>CAMSHIFT v/s Best performing configuration of the discriminatory color appearance model based tracker.</td>
<td>31</td>
</tr>
<tr>
<td>Table</td>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>-------</td>
<td>-----------------------------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>4.1</td>
<td>Test Videos</td>
<td>24</td>
</tr>
<tr>
<td>4.2</td>
<td>Evaluation of backprojection approaches</td>
<td>26</td>
</tr>
<tr>
<td>4.3</td>
<td>Effect of backprojection on tracking performance</td>
<td>28</td>
</tr>
<tr>
<td>4.4</td>
<td>Tracking accuracies for different extensions of the discriminatory color appearance model</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>based tracker</td>
<td></td>
</tr>
<tr>
<td>4.5</td>
<td>Combined tracking accuracies for different extensions of the discriminatory color appearance</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>model based tracker</td>
<td></td>
</tr>
<tr>
<td>4.6</td>
<td>CAMSHIFT v/s Discriminatory color appearance model based tracking</td>
<td>31</td>
</tr>
<tr>
<td>4.7</td>
<td>Computational speeds of the two algorithms</td>
<td>32</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

Computer vision is a branch of artificial intelligence that enables machines to interpret visual information as humans do. This young field has a wide range of applications ranging from security, logistics to medicine. A crucial component of this area of expertise is object tracking. Object tracking can be defined as the process of determining an object’s position in a video at every time instant. The objects may be either humans, rigid or nonrigid structures. In context of this work the term objects refers to people.

Many classifications of object tracking techniques have been proposed in the scientific community. Keeping in mind the relevance to this thesis, there is one meaningful way to classify object tracking. Based on the number of targets, tracking can be classified as single object and multiple object tracking. Commonly tracking is based on object features, hence called appearance model based tracking. In this a model representing an object of interest is created by using features like colors and shapes.

This work focuses on the demerits of a well know appearance model based, single object tracking algorithm called CAMSHIFT (Continuously adaptive meanshift) [6][5][4] in a multiple object scenario. A proposal for a new approach that can overcome its limitations in the specified scenario is made. This graduation assignment was carried out at Philips Research, Eindhoven under the framework of the VICOMO (Video context modeling) Project an ITEA2 (Information technology for European advancement) initiative. (http://www.vicomo.org).

1.1 Problem description

The CAMSHIFT algorithm is a well know tracking method. It was originally designed to track a single object using color features [6]. In short it creates a color based representation model of the object and tracks this model using a localization method. The color based representation of the object is a histogram of the color values describing the object. Localization means converging on to the object’s position in an image.

Due to its original design goals the CAMSHIFT has some inherent limitations when it comes to tracking multiple objects. There exist situations in which it gets confused by two similarly colored objects. This can result in the localization of an unintended object or even total loss of tracking performance. A complex background containing patches of color similar to the object of interest can also cause this confusion. CAMSHIFT serves as an example for a non-discriminatory tracker that illustrates this common limitation affecting all feature based trackers and is used as a starting point for this thesis.
The goal of this thesis is to create a new clustering based color representation model, that will distinguish two similarly colored objects by enhancing their discriminatory color features. Clustering is done explicitly in color space, in order to establish the colors unique to each object. A supervised version of fuzzy c-means clustering has been used [2]. This model has been designed to replace the existing color representation of an object in the CAMSHIFT algorithm.

1.2 Related work

In this section various relevant color and clustering based tracking techniques are summarized. The CAMSHIFT algorithm was first proposed by Bradski [6] for tracking faces in a perceptual user interface. It described a continuously adaptive color representation for the object being tracked. Only the hue channel in the HSV color space was utilized for the object (face) representation. The replacement of the static color model in the MEANSHIFT tracking [4] by an adaptive model, made tracking robust inspite of appearance changes in the object. This was observed in the context of the face tracking application. Allen et al [5] extended the application of CAMSHIFT from face tracking to object tracking. The appearance model was made to incorporate information from all the three channels of the HSV (hue, saturation, value) color space. A significant effort has been put in to model target appearances that lead to a performance gain. The object representation can also be created by accumulating the models (color histograms) representing the object in various views [10]. Israel [8] proposed a masked histogram approach that used seed growing algorithms to create a more robust object appearance model.

Several clustering based algorithms have also been proposed for object tracking. Clustering can be defined as the process of sorting raw data in to groups. Hua et al [9] proposed a 5-D feature space based clustering for object tracking. The five dimensions include the color components of Y,U,V and the two spatial dimensions of the image. Clustering has been used to group foreground pixels obtained after background subtraction, for tracking multiple people by Dubuisson et al [3]. Hua et al [11] proposed another clustering based tracker that combines both spatial and color components.

This thesis draws inspiration from the above mentioned research, combining the above two lines of study, clustering based tracking & appearance models, leading to a new clustering based discriminative appearance model with the aim of gaining performance while tracking multiple objects. The details of the clustering employed and the model itself will be explained in the later sections of this report.

1.3 Report Organization

This section describes the organization of the report. In chapter 2 the CAMSHIFT algorithm, which is a type of independent, stand-alone object tracker incapable of discriminating objects based on their colors, is explained. The details of the implementation created during the project and used for benchmarking have been specified. In chapter 3 all the details pertaining to the discriminative color appearance model based tracking, including the clustering mechanism have been described. The details of the implementation created have also been presented. In chapter 4 a comparative performance evaluation of the proposed discriminative color appearance tracker with the CAMSHIFT algorithm for tracking multiple objects is presented, along with the metric used in benchmarking. In chapter 5 the final conclusions and possible future work are discussed.
In this chapter the CAMSHIFT algorithm is described. It is a typical example of a non-discriminative color tracker where each object is tracked independently, giving rise to the possibility of confusion in case of similarly colored objects. In addition the details of the implementation used during the project are presented. An effort has also been made to understand the inherent limitations of this algorithm in context of multi-person tracking.

CAMSHIFT, an abbreviation for Continuously adaptive meanshift, is a color based tracking algorithm. As the name suggests it is a modified version of the meanshift tracking algorithm proposed by Comaniciu et al [4]. As already discussed in chapter 1, this modified meanshift i.e CAMSHIFT was proposed by Bradski et al [6] for face tracking in gaming peripherals. It uses the color statistics (i.e a color histogram model) to track an object. The CAMSHIFT algorithm is explained in detail throughout the subsequent sections of this chapter.

2.1 CAMSHIFT components

The CAMSHIFT algorithm can be represented by three functional modules, each performing a crucial role in the tracking process, they are:

- Object representation by use of an appearance model.
- Appearance model back projection.
- Position localization.

2.1.1 Object representation by use of an appearance model

In order to track an object in a video sequence, a model representing it has to be established. This is known as an appearance model. It can be constructed from various features describing an object like shapes, contours, texture, or color. One or many of these features can be utilized. The CAMSHIFT appearance model uses color features. Object tracking in a video means searching for an observation at each frame that best matches the original appearance model of the object created at the first frame or a reference frame.

As already mentioned, the CAMSHIFT tracker utilizes the statistics of color features for tracking objects. The appearance model for an object being tracked is a normalized histogram of the colors representing it. It is created as described in the following steps.

- In the reference frame of the video sequence, the object being tracked is specified using primitive shapes like an ellipse or a box. This area is called the target area. The target area can be thought of as the object position.
- A histogram of color values of the pixels within the target area is constructed as in equation 2.1. Each histogram bin represents an interval in the range of color values possible. And the bin value represents the number of pixels falling in the corresponding color interval.

\[
q_u = \sum_{k=1}^{n} \delta(c(I_k) - u) 
\]

(2.1)

where, \(q_u\) is the value of the histogram bin with index \(u\). \(\delta(p)\) is an impulse function such that \(\delta(p) = 1\) if \(p = 0\). \(I_k\) with \(k = 1...n\) is one of the \(n\) image pixels within the target area. \(c : I_k \rightarrow 1...m\) is a function that associates each pixel value to the index of the histogram bin in to which it falls. \(m\) is the total number of bins.

- The color histogram is normalized by dividing each bin value by the maximum of all bin values as shown in equation 2.2. Each bin value will now represent the normalized count of color values belonging to the object falling in to the bin. As obvious, the dominant color values (majority colors) will have higher values, thus highlighting these significant features.

\[
\hat{q}_u = \min\{\frac{q_u}{\max(q)}, 1\} 
\]

(2.2)

where, \(q_u\) represents the value of a histogram bin with index \(u\). \(\hat{q}_u\) represents the normalized value of the histogram bin with index \(u\). \(\max(q)\) is the maximum of all the bin values constituting the histogram \(q\) before normalization.

The appearance model for an object is depicted in figure 2.1. It is a normalized histogram representing the object’s color information in the hue channel of the HSV (hue, saturation, value) color space. However, in practice more complex 2-D or 3-D histogram models are created using the two or all three channels. As evident from figure below the bins representing the dominant colors of the object are given higher bin value after normalization.

Figure 2.1: An object being tracked and its corresponding color histogram appearance model.
2.1.2 Appearance model backprojection

Once the appearance model (color histogram) has been created, it is used to locate the color features representing the object in subsequent frames. This is called appearance backprojection, which in this case can be referred to as histogram backprojection. In this subsection the process of back projecting the histogram model to emphasize the object’s dominant color features in the image is described.

Before backprojection, the area of image where it will be performed has to be determined. This is known as the search area. The most likely position of the object in a subsequent frame is to be determined in this region based on its appearance model. Traditionally it is centered at the previously determined target area (subsection 2.1.1) and slightly bigger than it. The reason for this choice is due to the fact that the displacement of any object from one frame to another is typically very small. Therefore the object position in subsequent frames can be safely assumed to be within an area larger than the target area: the search area. In addition, backprojecting only in the search area reduces a lot of computational cost that would have been incurred in case it was done over the entire image.

The bins of the color histogram model under which each search area pixel falls is determined. The corresponding bin values are then assigned to each pixel as shown in equation 2.3. This is known as backprojection.

\[
\hat{I}_k = \hat{q}(I_k)
\]  

(2.3)

where \(I_k\) with \(k = 1\) to \(h\) is a pixel present in the search area and \(\hat{I}_k\) is the corresponding backprojected pixel. \(h\) is the number of pixels within the search area. \(\hat{q}(I_k)\) gives the value of the bin in the normalized histogram \(\hat{q}\) in to which the pixel \(I_k\) falls.

As a consequence, all the pixels that represent an object’s dominant colors are given higher values. A pixel representing a color not belonging to the object will end up being suppressed due to the lower normalized bin value assigned to it. Figure 2.2 depicts the effect of backprojection on the search area using the histogram model for the corresponding object depicted in figure 2.1. From the figure it is evident that the shirt color (red) is given a value of 1, while the trousers are given a lower value around 0.6 corresponding to the each color’s bin value in the histogram.

![Figure 2.2: An object and its corresponding search area’s histogram backprojection.](image)

2.1.3 Position Localization

Position localization refers to the process of converging on to the location of an object in its search area. It is done by employing the meanshift algorithm. The meanshift algorithm is generally used to find the peak of a data distribution. In this case it finds the peak of normalized bin values assigned in
the search area by backprojection, that signify the colors belonging to the object. The following steps describe the meanshift procedure.

- Step 1: For a given target area, its centroid within the search area is calculated using the zeroth and first order spatial moments as follows.

**Zeroth order moment:**

\[
M_{00} = \sum_{x=0}^{\text{width}} \sum_{y=0}^{\text{height}} I(x, y)
\]  

(2.4)

**First order moment for x:**

\[
M_{10} = \sum_{x=0}^{\text{width}} \sum_{y=0}^{\text{height}} x \cdot I(x, y)
\]  

(2.5)

**First order moment for y:**

\[
M_{01} = \sum_{x=0}^{\text{width}} \sum_{y=0}^{\text{height}} y \cdot I(x, y)
\]  

(2.6)

**Centroid:**

\[
x = \frac{M_{10}}{M_{00}}, \quad y = \frac{M_{01}}{M_{00}}
\]  

(2.7)

Where \(I(x, y)\) is the backprojected pixel value at spatial location \((x, y)\) in the image within the search window.

- Step 2: The target area is now centered at the new centroid.

- Step 3: Steps 1 and 2 are iterated till the centroid location of the target area does not change or moves a distance less than a preset threshold value. This is known as convergence.

- Step 4: The target area centered at the centroid obtained on convergence is the localized position of the object in the search area.

Figure 2.3 depicts the localized target area and its centroid, after the meanshift procedure.

![Figure 2.3: Localized object in backprojected search area and original image.](image-url)
2.2 Adaptive appearance model

So far in this chapter, all the basic functional modules of CAMSHIFT algorithm have been discussed. As mentioned in chapter 1 CAMSHIFT stands for continuously adaptive meanshift [6][5]. “continuously adaptive” means that the appearance model of the object is updated at each frame of the video. This is done to counter possible changes in object appearance over time caused by factors like lighting and different object appearance in different views. In such a case, using the original histogram appearance model leads to inaccurate tracking since it no longer represents the true appearance. Hence in order to overcome this challenge, in each frame at the site of localization a new histogram model is calculated.

2.3 Adaptive target size

As an object moves across the scene in a video, its size changes depending on its position. When it is farther away from the camera it has a smaller size and a larger size when closer to it. Thus in each frame the size of the target area is also updated to represent the scale change in the object’s size. This is calculated by using the zeroth order moment determined in the final iteration of meanshift during position localization. Zeroth order moment represents the distribution area of the backprojected pixels in the search area. Hence by using it the size of the object can be determined. It is done as shown in equation 2.8.

\[ s = 2 \times \sqrt{\frac{M_{00}}{\max(I_{\text{search}})}} \]  

(2.8)

where \( M_{00} \) is the zeroth order moment and \( I_{\text{search}} \) is the set of all pixel values within the search area. \( \max(I_{\text{search}}) \) is the maximum among the search area pixel values. \( s \) is new the dimension of the target area.

Depending on the type of objects being tracked the dimensions of the target area are set using \( s \). For instance when tracking people (special case of object tracking) the objects tend have a longer length. Hence in this case the breadth of the target area is set to \( s \), while the length is set to 1.5\( s \).

2.4 Algorithm

Now that the principle of CAMSHIFT algorithm has been discussed, in this section the algorithm and the process flow chart are presented. Given below is the algorithm of CAMSHIFT tracking used in the OpenCV library reference manual [1].

- Step 1: The target area at the object position in the reference frame is specified.
- Step 2: The histogram appearance model of the target area is calculated.
- Step 3: The appearance model histogram calculated in Step 2 is backprojected in a search area centered at the target area and slightly larger than it.
- Step 4: Position localization for the object using mean-shift algorithm in the backprojected search area is performed. The resulting spatial centroid and the zeroth moment are stored.
- Step 5: For the following frame the target area is centered at the spatial centroid found in Step 4. Its dimensions are set using the zeroth moment calculated during position localization in Step 4 as described in section 2.3.
- Step 6: Go to Step 2

Figure 2.4 depicts the process flow chart of the CAMSHIFT algorithm described above.
2.5 Implementation

In this section the implementation of the CAMSHIFT tracker created during the thesis is described. The tracker was implemented entirely in Matlab using the image processing and statistics toolbox. Details like the histogram appearance model, extension of the implementation to track multiple people, the search area size are mentioned here.

2.5.1 Histogram appearance model

For the implementation a 3-D histogram appearance model was chosen. The hue, saturation and value channels of the HSV color space formed the three input dimensions. 16 bins each were chosen for the hue and saturation channels. For the value channel (brightness) only five bins were used. The lower number of bins for brightness were selected to prevent any radical changes to the appearance model in a scene with significantly varying lighting conditions. Five bins for brightness was the generally reliable choice over the entire dataset. This design choice has been supported by Allen et al [5] in their effort make CAMSHIFT robust with multiple quantized feature spaces. In addition Sebastian et al [7] proved the improved performance of CAMSHIFT tracker when using color spaces where color and brightness information are separated (HSV). Thus the choice of HSV as the color space used for tracking instead of the well known RGB space. In total the implemented histogram model has a total of 16(hue)x16(saturation)x5(value) bins.
2.5.2 Adaptivity of the appearance model and target area

As discussed in section 2.2 the name CAMSHIFT refers to adaptivity in the histogram appearance model to make it robust for objects with radically different appearance in different views. However, in the implementation better tracking performance was obtained with a static appearance model i.e. $\alpha = 0$. This was also observed by Allen et al [5], who noticed that when using CAMSHIFT for a purpose other than face tracking, better performance was obtained with a static model. This is because color features not belonging to the object may be incorporated to the appearance histogram in case of improper localization of the object’s position.

The adaptivity of target area size has not been used in CAMSHIFT implemented during this thesis. Due to the adaptive target area size, pixels not belonging to the object may be included in to the histogram model calculated at each frame. For instance if the background in the search area has colors similar to the object, they are also assigned higher values. This leads to increase in the dimensions of the target area causing the background pixels to be included in the histogram update. The new histogram calculation at each frame is done with a fixed target area size specified at the reference frame. The search area used for backprojection is centered at the target area. Its dimensions are obtained after adding a constant offset to the target area dimensions. The target area dimensions are based of the interactive selection of the target at the reference frame.

2.5.3 Extension to multi-person tracker

The CAMSHIFT tracker was extended to multi-object tracking by assigning an instance of the single object tracker to each object. The implementation was extended to track two objects at a time. The backprojection values of each object were stored in separate two dimensional arrays to prevent overwriting and confusion in tracking.

2.6 Limitations of CAMSHIFT

The CAMSHIFT tracker was originally designed to track a single object. This focus has left the algorithm with some inherent limitations when used in multi-person tracking. When two objects represented by similar colors are close to each other the CAMSHIFT tracker can get confused, as the common color features on both the objects are highlighted during backprojection. Figure 2.5(b) depicts this scenario, where it can be seen that the color common to both objects 1 and 2 (figure 2.5(a)) is highlighted in the backprojection of object 1. This may cause confusion to the tracker instance following object 1.

![Figure 2.5: Limitation of CAMSHIFT in multi-person tracking.](image)
Any color based multi-object tracking algorithm depending on independent object appearance models is bound to face this issue. There is no mechanism or relation to look at the color features of multiple objects from a joint perspective. The CAMSHIFT algorithm is an instance of such algorithm where there is no provision for communication between multiple tracker instances. In this thesis a tracker based on a joint color appearance model, representing all the objects being tracked is proposed and discussed in chapter 3. It tries to find the discriminatory colors unique to each object and suppresses the shared color features thus avoiding confusion in tracking.
Chapter 3

Discriminative color appearance model based tracking

In the present chapter the discriminative color appearance model based tracking developed during this project is described. The need for such an appearance model can be attributed to the inherent shortcomings of the histogram appearance model while tracking multiple similarly colored objects. This has been clearly explained in section 2.6. In the proposed appearance model the discriminative colors corresponding to all the objects are determined by clustering their color data. The details of the discriminative appearance model, the clustering technique used and backprojection have been described in the following sections. In addition the challenges and extensions are also described.

3.1 Description

The discriminative color appearance model based tracking can be represented by three functional components. They are

- Determination of the discriminative color appearance model.
- Backprojection.
- Position localization.

It can be recalled from section 2.1 that the CAMSHIFT tracker has a very similar composition. The difference lies in the appearance model and backprojection. The position localization after the backprojection is same as in CAMSHIFT i.e. by use of meanshift algorithm. The first two components are described in the subsequent sections. For position localization subsection 2.1.3 can be referred.

3.1.1 Determination of the discriminative color appearance model

While tracking multiple objects using CAMSHIFT, each of them is represented by an individual appearance model based on a color histogram. The discriminative color appearance model on the other hand is a joint color space model dependant on every object being tracked. It is created by taking all the color values representing each object (from the target areas) and clustering them in a common color feature space.

Supervised fuzzy c-means clustering

Clustering is the process of partitioning raw data in to groups or sets. In this case it is utilized to determine the color values unique to each object. For this task a “supervised” version of the fuzzy c-means clustering algorithm [2] is employed. The term fuzzy refers to the possibility of a color value belonging to more than one object, instead of a hard partitioning where each color datum can belong to only one object. The extent to which a color value belongs to a particular object is called its responsibility. It normally lies between zero and one. Zero means no association, one represents a strong relation. The sum of the responsibilities of a point to all the clusters is always equal to one.
“Supervised” refers to the fact that the object to which each color value belongs is known throughout the clustering process.

**Clustering procedure**

The clustering is performed in the HSV color space. Every color sample being clustered is a 3 dimensional point, with each dimension representing one of the channels in the HSV color space. The creation of a joint color space discriminative appearance model using supervised fuzzy c-means clustering, which is an iterative process is described in the following steps.

- Step 1: The color values belonging to each object i.e target area are extracted and stored, such that the object to which they belong is always known.
- Step 2: For each object a centroid (mean) of the color values belonging to it is calculated. This centroid color value is now the representative of the object.
- Step 3: For each color value in the feature space its responsibility to every object (i.e. centroid of the object) is calculated. This is done as follows.

\[
U_{ij} = \frac{1}{\sum_{k=1}^{C} \frac{\text{dist}(x_i, c_j)^2}{m-1}}
\]  

where

- \(U_{ij}\) is the responsibility of a sample \(x_i\) to object \(j\) based on its centroid \(c_j\).
- \(C\) is the total number of centroids.
- \(\text{dist}(x_i, c_j)\) is the Euclidean distance between \(x_i\) and \(c_j\) given by \(\sqrt{\sum_{k=1}^{n} (x_i(k) - c_j(k))^2}\) where \(n\) is the number of dimensions. \(K = n\) represents the \(n^{th}\) dimension.
- \(m\) is the fuzziness factor, a greater value of \(m\) indicates that a larger proportion of the color values can belong to more than one object. Generally \(m\) is taken as 2, a value closer to 1 indicates a stricter partitioning of color samples with only few belonging to more than one object cluster.
- Step 4: The centroid of each object is now updated using the newly calculated responsibilities in step 3. In the original unsupervised C-means fuzzy clustering all samples in the clustering space are used in the update of each object cluster’s centroid. However in the supervised fuzzy clustering used here, the update of each object cluster’s centroid is done using only those samples in the clustering space that belong to the object as follows.

\[
c_j^* = \frac{\sum_{x_i \in \text{object} j} U_{ij}^m x_i}{\sum_{x_i \in \text{object} j} U_{ij}^m}
\]  

where

- \(c_j^*\) is the new updated centroid of object \(j\).
- \(U_{ij}\) is the responsibility of a point \(x_i\) to object \(j\) to the object’s centroid before the update in this step.
- Step 5: Steps 3 & 4 are repeated for a fixed number of iterations or until the maximum of change in responsibility values among all samples exceeds a certain threshold i.e. \(\max(U^{k+1} - U^k) > \varepsilon\). \(U^{k+1}\) is the set of all responsibilities in iteration \(K+1\) and \(U^k\) in iteration \(K\). This is the stopping criterion.
Determination of discriminatory colors

By the end of clustering each object is represented by a centroid closer to its unique color values. To grasp this closer proximity of object’s centroid to its unique colors, the process by which the responsibilities are assigned to a color sample has to be understood. The responsibility of color value towards a centroid closer to it is higher than to one farther away. In case the color is common to both the objects then its responsibilities to them will be nearly the same. This leads to lower values of responsibilities than those for the unique features. The updated centroid of an object is a weighted mean of all colors belonging to it (supervised). The weights are the responsibilities of the color values towards that object’s centroid before the update. Thus in the calculation of every new centroid the distinctive colors have a greater influence. This results in the centroid straying towards these discriminatory colors during the clustering. The benefit of having a representative centroid closer to discriminatory color features will be evident during the back projection phase. For the sake of better understanding a 2 dimensional (hue & saturation) illustration of this process is depicted in Figure 3.1.

![Figure 3.1: Discriminatory color appearance model](image)

(a) Color space filled with object samples and their centroids at the start of clustering.

(b) Assignment of responsibilities.

(c) Centroid drifts towards unique colors by the end of clustering.

Figure 3.1: Discriminatory color appearance model
Figure 3.2 shows the discriminatory color appearance model for two objects being tracked in one of the videos used for testing in chapter 4. Figure 3.2(b) depicts the HSV color space filled with the color values representing objects 1 and 2 shown in figure 3.2(a). The points depicted in red represent the colors belonging to object 1 and those shown in green belong to object 2. From the figure it can be observed that there are predominantly 3 groups of colors. One group completely belongs to object 1, another belongs entirely to object 2. A third group of colors is shared by both the objects. The resulting color space on performing the supervised fuzzy c-means clustering is depicted in figure 3.2(c). It can be observed that the group of colors belonging to either of the objects but not both (discriminatory colors) have a higher responsibility assigned to them. Where as the colors shared by both the objects (common colors) depicted in color black have lower responsibility. It is clear that the centroids of the object clusters representing each object end up near their discriminatory colors after the supervised fuzzy c-means clustering. Thus the discriminatory color appearance model is constructed.

(a) Tracked objects.

(b) Joint object color space before clustering.

(c) Color space after supervised fuzzy c-means clustering.

Figure 3.2: Result of discriminatory color appearance model
3.1.2 Backprojection

On obtaining the discriminatory color appearance model, centroids (weighted mean colors) representing each object are known. As explained in previous section they are closer to the discriminatory colors. The task of backprojection is to use these centroids to emphasize the discriminatory colors and suppress the common color features shared among the objects. The back projection is done in an area centered at the target area and slightly larger than it, this is the search area. For each pixel inside the search area the responsibility of its color value to the corresponding object’s centroid is calculated and then assigned back to it. As obvious the discriminatory colors of the object within the search area will get a higher responsibility, because of the centroid’s close proximity to them. Similarly colors shared among the objects will be assigned lower responsibilities. This can be thought of as a responsibility backprojection. Figure 3.3 depicts the effect of the method discussed so far. Figure 3.3(a) shows the two objects being tracked. Figures 3.3(b) and 3.3(c) depict the backprojected search areas of objects 1 and 2 respectively. It can be observed that only the discriminatory colors corresponding to each object have been highlighted in their respective backprojections.

Figure 3.3: Responsibility back projection
3.2 Discriminatory color appearance model based tracking algorithm

So far in this chapter the fundamental functional components like the discriminatory color appearance model and backprojection have been described. The position localization is performed like in CAMSHIFT using the moments describing the spatial distribution of responsibilities within the search area. In this section the algorithm for tracking using these functional components is presented. It can be noticed that the process is very similar to the CAMSHIFT except for the way in which the appearance model is constructed and how backprojection is done. The discriminatory color appearance model is constructed only once at the start of tracking. Hence it is static and is not updated in each frame. This is because color features not belonging to the object may be incorporated in to the appearance model in case of improper position localization. During extraction of an object’s color data for appearance model update, all the pixels within a box centered at it spatial centroid in the corresponding frame and with dimensions equal to it’s initialization in reference frame are chosen. This may be lead to inclusion of spatial locations not belonging to the object causing inaccuracies in the appearance model. The focus of this algorithm is on tracking multiple objects.

- Step 1: The target areas at the positions of all objects being tracked are specified at the reference frame.
- Step 2: The discriminatory color appearance model is constructed.
- Step 3: For each object, the appearance model calculated in step 2 is used to backproject the responsibilities in a search area centered at the target area and slightly larger than it.
- Step 4: Position localization for each object using mean-shift algorithm in search area is performed resulting in a spatial centroids for the target areas in the reference frame.
- Step 5: In all subsequent frames do steps 6 to 9.
- Step 6: In the current video frame, the target areas are positioned at the centroids obtained in the previous frame.
- Step 7: In the current frame the appearance model calculated at the reference frame is backprojected in search areas of all corresponding objects like in Step 3.
- Step 8: Meanshift is used to perform position localization to find the new spatial centroids of all target areas. Go to next frame.
- Step 9: Go to step 6.

Figure 3.4 depicts the flowchart describing the discriminatory color appearance model based tracking.
3.3 Extensions

The discriminatory color appearance model in its basic form has several challenges, like the effectiveness of back projection when using the original responsibilities, interference by background colors similar to the object’s color features and non-sufficiency of the number of centers per object to claim all the discriminatory colors. All these issues and the contributions made to solve them are discussed in this section.

3.3.1 Backprojection with modified responsibilities

During backprojection if the responsibilities of the colors are directly assigned to the spatial locations, it leads to significantly emphasized pixels in the background within the search area. This affects the position localization. The lack of background information in the discriminative color appearance model is the cause for this. These undesirably emphasized regions are the colors shared by the background and objects. These will be henceforth referred to as background interference. The results shown in figure 3.3 were obtained after modifications to suppress the background interference. The backprojection with the original responsibilities is depicted in the figures 3.6(a) and 3.6(b). As can be observed several pixels belonging to the background (background pixels) within the search area have been assigned moderately high values.

Backprojection approach using an exponent of responsibilities

In order to suppress the background interference two approaches were proposed. One approach is to assign an exponent of the responsibility to the spatial locations in search area. For example a responsibility of 0.5 in the background raised to power of 2 will be reduced to 0.25. However the object pixel colors have responsibility values closer to 1 and hence are less affected by this approach. This way the background interference is reduced. In figures 3.7(a) and 3.7(b) the results of backprojection with responsibilities raised to the power of five have been shown. The original responsibilities are modified...
as shown in equation 3.3

\[
\hat{U}_{ij} = \frac{1}{\left( \sum_{k=1}^{C} \frac{\text{dist}(x_i, c_j)}{\text{dist}(x_i, c_k)} \right)^{\frac{2}{m-1}}}^{p}
\] (3.3)

where the power \( p = 5 \) and \( \hat{U}_{ij} \) is the modified responsibility.

**Backprojection approach with responsibilities calculated using lower fuzziness**

Another approach is by using a lower fuzziness \( m \) (stricter partitioning) for calculating the responsibilities during backprojection. By doing this as the fuzzyness \( m \) goes to 1 the responsibility goes to 1 for all colors close to the centroid and the responsibility goes to 0 for those colors farther away in the color space. This is expected to suppress the moderately highlighted pixels in the background. Generally a fuzziness of \( m = 2 \) (preferred by literature) [2] is used while constructing the discriminatory color appearance model. A lower fuzziness of \( m = 1.4 \) has been used for the demonstration of backprojection in figures 3.7(c) and 3.7(d). A fuzziness of \( m = 1.7 \) was empirically chosen as the best fuzzyness for backprojection.

A third approach to backprojection may be the combination of the two approaches mentioned above, that is an exponent of five for the responsibilities calculated using a sharper fuzziness value. A comparison of the three approaches is evaluated in chapter 4. It was observed that backprojecting with a lower fuzziness leads to improved emphasis on object pixels (pixel locations belong to the object), however background interference still persists substantially. The use of an exponent of responsibilities for backprojection eliminates the background interference very effectively, but highlights the object pixels to a lesser extent. Hence it can be considered to be tradeoff between filtering background interference and better emphasis of object colors.
(a) Backprojection for object 1 using original responsibilities.

(b) Backprojection for object 2 using original responsibilities.

Figure 3.6: Backprojection using original responsibilities.

(a) Backprojection for object 1 using exponent of 5 for responsibilities.

(b) Backprojection for object 2 using exponent of 5 for responsibilities.

(c) Backprojection for object 1 using sharper fuzziness $m = 1.4$.

(d) Backprojection for object 2 using sharper fuzziness $m = 1.4$.

Figure 3.7: Responsibility modification for improved backprojection
3.3.2 Including background information in discriminatory color appearance model

In subsection 3.3.1 approaches to improve the responsibility backprojection were discussed. In spite of these improvements, background interference continues to persist. An extension to further filter this interference is the inclusion of background information during construction of the discriminatory color appearance model. This addition enables the suppression of color features shared between the objects and the background. The background information is obtained by spatially sub-sampling pixels over the entire image to create a model representing its dominant colors. In the horizontal direction of the image one pixel per 10 spatial positions and in vertical direction one row per 4 rows of pixels is included in the background model. In the supervised fuzzy clustering process the background data model is also treated like an object and is represented by a centroid.

Figures 3.8(b) and 3.8(d) show the result of backprojection with the inclusion of the background information in the appearance model. It is clear that addition of background information in the discriminative color appearance model brings a significant improvement in the back projection approaches. This is an effective extension to eliminate the persistent background interference observed in backprojections without a background model (figures 3.8(a) and 3.8(c)). The backprojection approaches are compared to select the best method in chapter 4.

(a) Backprojection using exponent of 5 for responsibilities without background model.
(b) Backprojection using exponent of 5 for responsibilities with background model.
(c) Backprojection using sharper fuzziness \( m = 1.7 \) without background model.
(d) Backprojection using sharper fuzziness \( m = 1.7 \) with background model.

Figure 3.8: Responsibility backprojections without and with a background model
3.3.3 Extension to multiple centroids per object

The supervised fuzzy c-means clustering is not always effective in finding the discriminatory colors of an object with just one centroid. The joint color space distribution of objects being tracked is not always predictable and simple, like the illustration in figure 3.1. In some distributions, there is a strong possibility that the centroid of the object ends up near color features that do not belong to it. Another possibility is that centroid converges such that not all discriminatory colors are emphasized. An extension to tackle this problem is to use two or more centroids per object in the discriminatory color appearance model. For every object the multiple centroids are initialized by adding a random offset to the mean of the object’s color values calculated in the step 2 of a discriminatory color appearance model construction (section 3.1.1). The responsibilities are calculated in the same way as in equation 3.1. The only difference is that each sample will now have certain responsibility to all the centroids representing an object. The centroids belonging to an object cluster are updated as follows.

\[
c_jz^* = \frac{\sum_{\forall x_i \in \text{object}_j} U_{ijz} m \cdot x_i}{\sum_{\forall x_i \in \text{object}_j} U_{ijz} m}
\]

where \( c_jz^* \) is the \( z^{th} \) centroid of an object \( j \). \( U_{ijz} \) is the responsibility of a color sample \( x_i \) to the \( z^{th} \) centroid of an object \( j \).

With multiple centroids the clustering process will ensure that discriminatory colors missed before are now claimed by the object. Backprojection in an object’s search area is done by assigning each pixel the sum of the responsibilities to all centroids representing an object.

3.3.4 Extension to multiple centroids per background

The background in a scene can be very complex. Using just one centroid may not always yield the best of background suppression. By using more centroids for the background model in the supervised fuzzy c-means clustering, the influence of the object on the color features shared by the object and background is reduced. This is because the responsibility of a common color is now distributed among more centroids, which inevitable leads to a reduction in the color’s responsibility to the object. It is possible to keep on adding more centroids to the background model, however too many of them will lead to some undesirable suppression of color features belonging to the objects being tracked. It is recommendable to maintain more or less same number of centroids for the background as there are for each of the objects.

3.3.5 Implementation details

In this subsection details of the discriminatory color appearance model based tracker implemented during the thesis are presented. The tracker was implemented entirely in Matlab using the image processing and statistics toolbox. The supervised fuzzy c-means clustering used during the construction of the appearance model is performed in a 3-D color feature space. The three dimensions are the three channels of the HSV color space, i.e hue, saturation and value.
In order to determine the responsibilities of a color sample, the Euclidean distance between the sample and the object centroids have to be calculated in the HSV color space. However the hue space is circular in nature with a normalized range from 0 to 1 (originally 0 to 360 degrees) as shown in figure 3.9. Therefore a centroid and a color point separated by a normal Euclidean distance of 1 unit are in fact one and the same. But using the normal Euclidean distance in the hue space, they might appear to be farther away thus resulting in a very low responsibility of the color to the centroid. To tackle this issue a modification was made to the way the distance is calculated in the hue dimension. Instead of using \( \sqrt{(x_{i(hue)} - c_{j(hue)})^2} \) where \( x_{i(hue)} \) and \( c_{j(hue)} \) are the dimensions of the color sample \( x_i \) and the centroid \( c_j \) in hue channel, \( \sqrt{\min((x_{i(hue)} - c_{j(hue)}), (1 - (x_{i(hue)} - c_{j(hue)})))^2} \) is done in equation 3.1. This way even when distance between the centroid and the color point is 1 unit according the normal Euclidean distance the modified distance will be 0 units, which is consistent with the circular nature of the hue channel.

A fuzziness factor of \( m = 2 \) is chosen for the supervised fuzzy c-means clustering used in the construction of the discriminative color appearance model. Three approaches to backprojection have been proposed in this chapter and compared in chapter 4. One approach is the backprojection with a lower fuzziness value (\( m = 1.7 \)) than that chosen for the clustering process in the discriminative model. In the second approach an exponent of the responsibilities (exponent = 5) was chosen upon observing the results of several backprojected images. A third approach combining the 2 methods mentioned so far has also been included in the comparison.

The tracker implementation has been extended to track two objects at a time. The discriminatory color appearance model was enabled to accommodate 2 centroids per object and up to 3 centroids for the background model. This background model was created by sampling pixels color values in the reference frame. Several variants of the discriminative appearance model based tracking with various extensions have been evaluated and compared in the next chapter.
Chapter 4

Performance evaluation

In this chapter the performance of the discriminatory color appearance model based tracking is evaluated. Its performance is compared to the CAMSHIFT algorithm extended to multi-object tracking. The metric and dataset used are also described.

4.1 Evaluation metric

The evaluation in this work is performed offline on recorded videos. The well known metrics in object tracking literature use the hand labeled ground truth of the recorded test videos to evaluate the tracking performance. Ground truth is the actual trajectory of the object marked according to human observation [14]. However extracting it is a tedious and time consuming process prone to error due to the human involvement [12]. In this thesis a simple automated, trajectory based metric has been devised.

The metric uses the deviation in the generated trajectories of the object in forward and time-reverse direction to estimate the tracking performance. For a given object at each frame 2 estimated positions at a certain time instant will be obtained, one from forward tracking and the other from time-reversed tracking. These are henceforth referred to as forward position (while tracking through sequences in normal order) and reverse position (tracking through the sequences in reverse order). The Euclidean distance between these two positions is calculated and is referred to as the deviation in object trajectory in the corresponding video frame. The logic in using such a metric is to determine frames where there is a large deviation, indicating a possible inaccuracy in tracking. The underlying assumption is that if the forward and backward trajectories coincide then they represent the true trajectory. This has been empirically observed while using the tracker over various recorded test videos.

The deviation in the forward and the reverse estimated positions is used to determine a percentage accuracy in tracking performance. In each frame of the test video sequence if the deviation exceeds a certain threshold, it is classified as a mis-hit. If the deviation is within the threshold, then it is a hit. The threshold for the deviation is equal to half the largest dimension of a box used to initialize the object. This threshold means that in case of accurate tracking the two positions will be located anywhere on the object or its immediate vicinity. The percentage tracking accuracy is defined as follows:

\[
\text{Tracking accuracy(\%)} = \frac{\text{Number of frames with hits}}{\text{Total number of frames}} \times 100
\] (4.1)
4.2 Dataset

For the evaluation a total of six test videos were selected. Video 1 was made by the author at Philips Research during the thesis. It depicts the specific case where confusion in tracking generally arises because of common color features of the objects. Videos 2, 3 and 4 were taken from the Imagery Library for intelligent detection systems (i-Lids) dataset for AVSS 2007 \cite{AVSS2007}. Videos 5 and 6 were taken from the Context Aware Vision using Image-based Active Recognition (CAVIAR) dataset for object tracking \cite{CAVIAR}. The details of the videos are presented in table 4.1. The lengths of the video segments were chosen such that the objects being tracked are always visible without exiting the scene. Video 4 represents a case where an object and person are tracked. It should be kept mind that the discriminatory color appearance model based tracking can be extended to human tracking which is a special case of object tracking discussed so far. The starting positions of the objects are fixed prior to the evaluation. For each object two sets of starting conditions are provided, one for forward tracking and the other for time reversed tracking. The initializations are done so that entire object is encompassed and little or none of the background is selected.

<table>
<thead>
<tr>
<th>Video</th>
<th>Number of Frames</th>
<th>Source</th>
<th>Original frame rate(frames/second)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Video 1</td>
<td>80</td>
<td>Self-made</td>
<td>30</td>
</tr>
<tr>
<td>Video 2</td>
<td>50</td>
<td>i-Lids</td>
<td>25</td>
</tr>
<tr>
<td>Video 3</td>
<td>60</td>
<td>i-Lids</td>
<td>25</td>
</tr>
<tr>
<td>Video 4</td>
<td>150</td>
<td>i-Lids</td>
<td>29</td>
</tr>
<tr>
<td>Video 5</td>
<td>65</td>
<td>CAVIAR</td>
<td>29</td>
</tr>
<tr>
<td>Video 6</td>
<td>60</td>
<td>CAVIAR</td>
<td>29</td>
</tr>
</tbody>
</table>

Table 4.1: Test Videos

Figures 4.1 and 4.2 shows a sample frame from each of the videos and the corresponding objects being tracked.

(a) Video 1.

Figure 4.1: Sample frame and objects being tracked from the self made video
Figure 4.2: Sample frames and objects being tracked from the CAVIAR and i-Lids dataset

(a) Video 2.
(b) Video 3.
(c) Video 4.
(d) Video 5.
(e) Video 6.
4.3 Evaluation

In this section the performance evaluation of the discriminatory color appearance model based tracking in relation to its parameters and its comparison with the CAMSHIFT tracking algorithm are presented.

4.3.1 Backprojection evaluation

In this subsection the two backprojection approaches described in chapter 3 i.e using responsibilities calculated with a lower fuzziness or using an exponent of the responsibilities are compared. In addition a third approach combining the two methods is also compared. It is expected that the third approach will yield a better result. A lower fuzziness factor will provide improved emphasis on the object pixels and the exponent of these responsibilities will eliminate the interference in the background within the search area as already mentioned in subsection 3.3.1. In order to evaluate the best backprojection the discriminatory color appearance model in its most basic form, i.e. one center per target with no incorporated background information, is chosen. Table 4.2 shows the tracking accuracies of the discriminatory color appearance model based tracker using different approaches of backprojection. Figure 4.3 presents the average of tracking accuracies for objects 1 and 2 calculated over the entire dataset. The tracking accuracies for each video presented in the table are the average values taken over three trials as there is some variation in the clustering outcome due to a random initialization of cluster centers. The following backprojection approaches have been compared:

- **Backprojection 1**: Backprojection using an exponent of 5 for responsibilities calculated with the same fuzziness factor as in the appearance model (m=2).
- **Backprojection 2**: Backprojection using responsibilities calculated with a lower fuzziness (m=1.7) than in the appearance model (m=2).
- **Backprojection 3**: Backprojection combining both the approaches, using using an exponent of 5 for responsibilities calculated with a lower fuzziness factor (m=1.7).

<table>
<thead>
<tr>
<th>Test video</th>
<th>Tracking Accuracy %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Backprojection 1</td>
</tr>
<tr>
<td></td>
<td>Object 1</td>
</tr>
<tr>
<td>Video 1</td>
<td>100</td>
</tr>
<tr>
<td>Video 2</td>
<td>100</td>
</tr>
<tr>
<td>Video 3</td>
<td>100</td>
</tr>
<tr>
<td>Video 4</td>
<td>14</td>
</tr>
<tr>
<td>Video 5</td>
<td>0</td>
</tr>
<tr>
<td>Video 6</td>
<td>100</td>
</tr>
<tr>
<td>Average over dataset:</td>
<td>69</td>
</tr>
</tbody>
</table>

Table 4.2: Evaluation of backprojection approaches
Figure 4.3: Average of tracking accuracies for objects 1 and 2 over the entire dataset using different backprojection approaches.

To determine the better performing backprojection method a combined tracking accuracy is calculated. It is average of tracking accuracies corresponding to all the objects. Given below are the combined tracking accuracies for different backprojection approaches using their average metric values over the entire test dataset.

- Backprojection 1: \( \frac{69+71}{2} = 70\% \)
- Backprojection 2: \( \frac{73+66}{2} = 69.5\% \)
- Backprojection 3: \( \frac{69+73}{2} = 71\% \)

Backprojection 3 which combines the first two approaches performs slightly better overall and is selected for backprojection in further tests. It can be observed that for videos 4 & 5 tracking fails severely in some or all cases. This may be attributed to the lack of background information in the appearance model configuration used in this test. In spite of using the backprojection approaches with the intent to reduce the background interference (emphasis of background pixels), it tends to persist in some magnitude depending on the scene as discussed in subsection 3.3.2. On incorporating the background model the performance of tracking in these cases is expected to improve substantially. The performance of the tracker with incorporated background information and multiple centers per target/background will be studied in the subsequent sections.
4.3.2 Effect of Background model

In this section the impact on tracking performance due to the inclusion of background information in the fuzzy clustering based discriminative color appearance model is studied. Prior to the addition of the background information, the colors shared by the objects and the background are not suppressed leading to a background interference as mentioned in section 3.3. By incorporating the background information in to the appearance model, this interference is eliminated. This is expected to result in an improved tracking performance. Backprojection 3 (combined approach) is selected as it was determined to be the best approach in the previous subsection. The configuration of the tracker using the discriminatory color appearance model without the background information is denoted by 1O + 0B (1 center per object and 0 background centers). The configuration with the background is represented by 1O + 1B (1 center per object and 1 center per background). The tracking accuracy values for each video presented in table 4.3 are the average values taken over three trials as there is some randomness in the clustering outcome. The combined tracking accuracy in the last two columns is the average for tracking accuracies of objects 1 and 2 for each configuration of the tracker and the corresponding video in the test dataset. Figure 4.4 shows combined tracking accuracies using configurations of the discriminatory color appearance model based tracker with and without the background information.

<table>
<thead>
<tr>
<th>Test video</th>
<th>Tracking Accuracy %</th>
<th>Combined tracking accuracy %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object 1</td>
<td>Object 2</td>
<td>Object 1</td>
</tr>
<tr>
<td>------------</td>
<td>---------------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>Video 1</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Video 2</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Video 3</td>
<td>100</td>
<td>63</td>
</tr>
<tr>
<td>Video 4</td>
<td>13</td>
<td>74</td>
</tr>
<tr>
<td>Video 5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Video 6</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 4.3: Effect of backprojection on tracking performance

From Table 4.3 and figure 4.4 it is clear that there is an improvement in tracking performance on inclusion of background information in the discriminatory color appearance model. This is profoundly evident from combined tracking accuracies of videos 3, 4 and 5. For video 5 without the background information the combined tracking accuracy was zero. On incorporating the background information in to the appearance model it increases substantially to 45%. Similarly an increase in performance is observed for videos 4 and 5. By adding more centroids per background in the appearance model a further gain in performance can be expected. In the remaining videos the combined tracking accuracy is more or less the same, indicating a definite improvement in overall performance.
4.3.3 Effect of multiple centroids per object or background

In this subsection the effect on tracking performance as multiple centroids per object or background are enabled for the fuzzy c-means clustering based discriminatory appearance model, is studied. During this thesis the tracker was extended to accommodate up to two centroids per object and three centroids per background. Various configurations tested and their abbreviations are given below.

- 1 centroid per object/background (1O + 1B).
- 2 centroids per object and 1 centroid for the background (2O + 1B).
- 2 centroids per object and 2 centroids for background (2O + 2B).
- 2 centroids per object and 3 centroids for background (2O + 3B).

Table 4.4 shows the tracking accuracies for each video using the different tracker configurations. These tracking accuracies are the average values taken over three trials. Table 4.5 present the combined tracking accuracies for each video and the corresponding tracker configurations.

<table>
<thead>
<tr>
<th>Test video</th>
<th>Tracking Accuracy %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1O+1B</td>
</tr>
<tr>
<td>Object 1</td>
<td>Object 2</td>
</tr>
<tr>
<td>Video 1</td>
<td>94</td>
</tr>
<tr>
<td>Video 2</td>
<td>100</td>
</tr>
<tr>
<td>Video 3</td>
<td>100</td>
</tr>
<tr>
<td>Video 4</td>
<td>100</td>
</tr>
<tr>
<td>Video 5</td>
<td>29</td>
</tr>
<tr>
<td>Video 6</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 4.4: Tracking accuracies for different extensions of the discriminatory color appearance model based tracker.
Table 4.5: Combined tracking accuracies for different extensions of the discriminatory color appearance model based tracker

<table>
<thead>
<tr>
<th>Test video</th>
<th>1O+1B</th>
<th>2O+1B</th>
<th>2O+2B</th>
<th>2O+3B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Video 1</td>
<td>97</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Video 2</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Video 3</td>
<td>100</td>
<td>71</td>
<td>82</td>
<td>90</td>
</tr>
<tr>
<td>Video 4</td>
<td>94</td>
<td>76</td>
<td>86</td>
<td>82</td>
</tr>
<tr>
<td>Video 5</td>
<td>45</td>
<td>76</td>
<td>82</td>
<td>100</td>
</tr>
<tr>
<td>Video 6</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

Figure 4.5 depicts the bar graph for the combined accuracy values for the various extensions of the discriminatory color appearance model based tracker to multiple centroid per object/background depicted in table 4.5.

Figure 4.5: Combined tracking accuracies for extensions enabling multiple centers per object/background.

From the figure 4.5 it can be observed that on adding an additional centroid to the object clusters in the appearance model, it doesn’t always yield an improvement in performance. In videos 1 and 5 the configurations with two centroids per object show a higher combined tracking accuracy in comparison to the configuration with a single centroid per object. In videos 3 and 4 however this trend is reversed and the configuration with single centroid per object reigns superior. It can be interpreted from these contradicting trends that the optimal configuration may vary according to the scene and the color distribution of the objects. In complicated scenes with multiple discriminatory colors it may be preferred to employ two centroid per object in the discriminatory color appearance model. It is also observed that as the number of background centers are increased there is a an upward trend in the combined tracking accuracy. This is evident from the results of videos 3, 4 and 5. This is because by adding more centroids for the background the responsibility of a color value common to both object and background is shared among more centroids. This reduces the overall responsibility of the color to the object, thus suppressing the background noise in the search area more effectively.
4.3.4 Performance comparison of discriminatory color appearance model based tracking with CAMSHIFT

In this subsection the combined tracking accuracy of the CAMSHIFT algorithm is compared with the best performing configuration of the discriminative color appearance model for each corresponding video. Table 4.6 shows the tracking accuracies of the CAMSHIFT and best discriminative color tracker. Figure 4.6 shows a bar plot of the combined tracking accuracies corresponding to the two trackers.

<table>
<thead>
<tr>
<th>Test video</th>
<th>CAMSHIFT</th>
<th>Best discriminatory tracker</th>
<th>Combined tracking accuracy %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Object 1</td>
<td>Object 2</td>
<td>Object 1</td>
</tr>
<tr>
<td>Video 1</td>
<td>76</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Video 2</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Video 3</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Video 4</td>
<td>55</td>
<td>18</td>
<td>100</td>
</tr>
<tr>
<td>Video 5</td>
<td>100</td>
<td>18</td>
<td>100</td>
</tr>
<tr>
<td>Video 6</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 4.6: CAMSHIFT v/s Discriminatory color appearance model based tracking

From the table 4.6 and figure 4.6 it is clear that the CAMSHIFT tracking algorithm is outperformed by the best performing configuration of the discriminative color appearance model for each test video. The discriminatory color tracker performs on par with the CAMSHIFT in test videos 2,3 and 6. An improvement in the combined tracking accuracy is observed by using the discriminatory color tracker in test videos 1, 4 and 5. Test video 1 was specifically made representing the case where shared colors between the objects can cause confusion in tracking. As evident from the results the discriminatory tracker shows an 100 % combine tracking accuracy against the 88 % shown by the CAMSHIFT in this case.
4.4 Computational complexity of the discriminative color appearance model based tracker compared to CAMSHIFT

In this section the computational complexity of the discriminatory color appearance model based tracker is compared with the CAMSHIFT tracker. The construction of the discriminative color appearance model is computationally more intensive than the histogram appearance model in CAMSHIFT. The histogram calculation is quite simple involving a single access to each pixel value belonging to the objects and basic arithmetic operations. In contrast the discriminative color appearance model performs an iterative supervised fuzzy clustering of large amounts of color data belonging to all the objects and the background. On the other hand, in the proposed method only few centroids representing the objects are needed during tracking. CAMSHIFT in comparison uses a complete multi-dimensional histogram per target during tracking. The computational complexity of the discriminative color appearance model may be reduced by performing the supervised fuzzy c-means clustering on sub-sampled object color data. The higher resource requirement of the proposed method is justifiable considering the improvement in tracking accuracy exhibited by the discriminative color appearance model based tracker in comparison to the CAMSHIFT. In table 4.7 the computational times taken by each algorithm’s Matlab implementation per frame corresponding to a video in the test data set are shown. It was observed that backprojection consumed the most time over the entire video for both the algorithms. The times shown are seconds per frame.

<table>
<thead>
<tr>
<th>Test video</th>
<th>Total frames</th>
<th>CAMSHIFT</th>
<th>backprojection</th>
<th>Discriminatory color tracker</th>
<th>backprojection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Video 1</td>
<td>100</td>
<td>0.47</td>
<td>0.10</td>
<td>0.59</td>
<td>0.221</td>
</tr>
<tr>
<td>Video 2</td>
<td>50</td>
<td>0.54</td>
<td>0.173</td>
<td>0.74</td>
<td>0.351</td>
</tr>
<tr>
<td>Video 3</td>
<td>50</td>
<td>0.6</td>
<td>0.161</td>
<td>0.76</td>
<td>0.31</td>
</tr>
<tr>
<td>Video 4</td>
<td>120</td>
<td>0.625</td>
<td>0.15</td>
<td>0.74</td>
<td>0.2125</td>
</tr>
<tr>
<td>Video 5</td>
<td>65</td>
<td>0.49</td>
<td>0.11</td>
<td>0.64</td>
<td>0.22</td>
</tr>
<tr>
<td>Video 6</td>
<td>60</td>
<td>0.46</td>
<td>0.17</td>
<td>0.58</td>
<td>0.19</td>
</tr>
</tbody>
</table>

Table 4.7: Computational speeds of the two algorithms

It is clear that backprojection consumes a significant portion of the time spent on processing a single frame. The other time consuming functionality observed was the conversion of the images to HSV format, which is common to both the algorithms. It can also be seen that the algorithm proposed in this thesis consumes slightly more time than the CAMSHIFT, this is due to the more complex backprojection as evident from the table.
In this thesis a discriminatory color appearance model based tracking algorithm has been designed and implemented. Existing color based tracking algorithms like the CAMSHIFT are not effective in tracking multiple objects. In cases where the objects possess common colors, a confusion in tracking may arise. This is due to the independent, stand-alone nature of the tracker instance assigned to each object. This lack of a joint perspective to determine the discriminative colors unique to each object is the main cause for their failure. The tracker designed during this thesis is the discriminatory color appearance model based tracker. It employs a supervised version of the fuzzy C-means clustering algorithm in a joint object color space to determine the discriminatory object colors. Thus the confusion in tracking is avoided. A fuzzyness factor of 2 was used to construct the discriminatory color appearance model based tracking. By the end of clustering the cluster centroids (mean colors) representing each object are determined. Backprojection for each object in the proposed tracking is done by using the responsibilities of the pixels in the search area to the corresponding object centroids.

Three backprojection approaches were proposed in order to suppress the background interference observed in case the original responsibilities were directly assigned to search area pixels. Among them using an exponent of 5 for the responsibilities calculated with a lower fuzziness of 1.7 yielded slightly better performance. To further suppress the background an extension to the discriminatory color appearance model incorporating the background information in the fuzzy clustering was proposed. This shows a significant improvement in tracking. A further gain in performance was observed as more centroids were added to the background cluster in the appearance model. Another extension of the appearance model enabling the use of two centroids per object cluster was proposed. This allows the determination of multiple discriminatory colors of an object that would have been otherwise lost by using just once centroid per object. In some test videos using two centroids per object shows a some what diminished performance in comparison to the tracking with single centroid per object and an improved performance in some other videos. This indicated that the optimal number of centroids per object may vary from scene to scene. However it is recommended to utilize multiple centroids per object in complicated scenes with many discriminative colors.

A comparison of the tracking performance of CAMSHIFT and discriminatory color appearance based algorithm was done using a heuristic trajectory based tracking metric devised during the thesis. The best performing extension of the discriminatory color appearance model and the CAMSHIFT trackers implemented in Matlab were compared for each video in the test set containing videos made by the author and from video surveillance dataset repositories like CAVIAR and i-LIDS. From the results obtained it is clear that the the discriminatory color based tracking algorithm outperforms the CAMSHIFT algorithm.
5.1 Future Work

In this thesis an offline Matlab implementation of the discriminative color appearance model based tracker has been made. The tracking is performed on objects specified at the start of the procedure in the reference frame. A real-time implementation that automatically detects the objects in a scene and tracks them using the proposed algorithm could be made. The performance and speed of the algorithm in such a real-time implementation may be an interesting direction to explore.

The discriminatory color appearance model based tracker uses color information in the HSV (hue, saturation, value) color space. Since various color spaces differ in the way colors are characterized and distributed it may be of interest to examine the performance of the this tracker in different color spaces.

The results obtained after comparing the backprojection approaches suggest a slightly improved combined tracking accuracy on using the exponent of the responsibilities calculated with a lower fuzziness value. In order to obtain a more decisive result for determining the better performing backprojection, the dataset could be expanded to accommodate more number of videos.

Finally in configurations of the discriminatory color appearance model based tracker where more than one centroid per object/background is used, a random initialization of the centroids in the supervised fuzzy clustering is used. The final result of clustering may vary depending on the random initialization. This leads to variation in tracking accuracies. An initialization resulting in a more stable tracking performance should be researched.


