Object Motion Perception and Tracking Using Sift With K-Means Clustering

Deekshitha P [1], B Tarun Reddy [2], Shravani Badadha [3], K K Dhruthi [4], S Pandiaraj [5]
[1, 2, 3, 4] U.G. Student, Department of Computer Engineering, SRM institute of science and technology, Chennai, Tamil Nadu, India.
[5] Associate Professor (S.G), Department of Computer Engineering, SRM institute of science and technology, Chennai, Tamil Nadu, India.

Abstract – Object tracking is a very complex technique, but vital task to be addressed in any video surveillance system. Detecting a dynamic object accurately in a video is a difficult job. Moving object tracking can be done by comparison of frames. In this paper, we propose a technique which compares the frames and helps to track the object. Based on the motion information, key points corresponding to moving objects are extracted from every frame by using The Laplacian of Gaussian. The corner detector is used to eliminate the background points. SIFT (Scale Invariant feature transform) approach is specifically used for object recognition and to detect feature points when the scale and orientation are in place. Tracked SIFT capabilities provide the displacement of every feature point in the image, along with image coordinates and frame number constitute a feature vector. All the key points along with the feature vector are integrated and clustered in order to track the objects. Improved k-means clustering is used to group the feature points and detect the moving object.

Index Terms – Feature Extraction, object detection and clustering analysis.

1. INTRODUCTION

Efficient and accurate matching of image capabilities or regions is an essential prerequisite to many complex issues in image processing domain. The achievement of higher level strategies such as object detection, re-identifying, classification or tracking depends closely on the quality of image . An easy and rapid method to matching is through area association but there are some problems like occlusion, illumination and noise. This method has a wide range of applications globally. There are various issues which are faced by the objects due to occlusion either partial or complete. There is a failure due to various other algorithms during the image matching. With the usage of key points for matching among the frames has reduced the problems faced so far which has given positive results in object tracking and detection. Studies prove that SIFT Algorithm helps the most in this aspect.

Object tracking in any surveillance system is very important as it has various applications. There are many techniques which face lot of challenges in object tracking. The most common one is the blob tracking, which normally emphasizes the distinction between the present day image observation and a model of the heritage. Blob tracking generally face problems related to shadows and occlusions. There is another technique known as active contours which is related to blob tracking. Active contours (snakes) are used to separate the image as foreground and background. There are certain limitations as this technique is not very accurate in the segmentation process.

Feature based model is used for object tracking as the feature points are extracted and then the points are clustered into various groups and the comparison between the frames are checked. Feature-based model algorithms can be classified into three subcategories: local feature-based, global feature-based and dependence-graph-based techniques. In the local-based method, features used are corner vertices, line segments, curve segments. Color, centroid, and perimeters features come under global feature-based. In the dependence-graph-based method the features are viewed with the help of graphs.

Image pre-processing is the time period for operations on pictures at the bottom level of abstraction. These operations do not increase picture facts content material but they lower it if entropy is an information measure. The purpose of pre-processing is an improvement of the picture statistics that suppresses undesired distortions or enhances some picture capabilities applicable for further processing and evaluation project. Image pre- processing uses redundancy in pictures. Neighboring pixels similar to one real object which have the identical or similar brightness price. If a distorted pixel can be picked out from the photograph, it is able to be restored as an average cost of neighboring pixels.

A video is converted into frames. Each frame is represented with GMM model. In this method, the foreground and the background is extracted. SIFT features points are extracted in all the regions then after further processing it is narrowed down to only the foreground feature points. These key points are used compare with the various frames. Every matched pair of SIFT functions gives a displacement vector. We gather function vectors that consist of this displacement, the image coordinates and the frame range. The range of the frames is

normalized. All the key points are processed by using improved k-means clustering. The best clusters are formed by an iterative process. This work contributes to accurate object tracking by overcoming the challenges faced.

The remainder of the paper is organized as follows: Section 2 describes the Architecture and implementation of object tracking and re-identification, Section 3 presents and discusses our experimental results and Section 4 provides the conclusions and future enhancement of this work.

2. RELATED WORKS

2.1 Generalized Stauffer–Grimson historical past subtraction:

Generalized Stauffer-Grimson background separation is an algorithm which is used to separate the background from that precise picture or frame. Background separation is one of the most necessary steps in object detection. In this, we use a k-means algorithm for updating the parameters the use of the vital records of the model. Then primarily based on the analysis of the movement video the historical past separation is done. Usually, the background separation algorithms are robust. There are various strategies like Dynamic textures, Background models, Background subtraction, mixture models and adaptive fashions used. It is suitable only two for static scenes. This is a sturdy trouble for scenes with spatiotemporal dynamics.

2.2 A People-Counting System primarily based on BP Neural Network:

A people-counting system is primarily based on a back propagation (BP) neural network. The system makes use of affordable photoelectric sensor to collect information and introduces BP neural network for counting and recognition, and it is advantageous and bendy for the purpose of performing people counting. There are new techniques for segmentation and characteristic extraction which are developed to enhance the classification performance. Promising consequences have been acquired and the evaluation shows that the device based on BP neural network presents proper consequences with low false rate and it is effective for people-counting. There are a range of methods like segmentation, feature extraction and back propagation neural network used. To completely put into effect a standard neural network architecture would require lots of computational sources which is a major drawback.

2.3 Tracking and Counting People in Traffic Surveillance Systems:

The most difficult challenge in a visible Surveillance system is to count and track people. If there is only one person in the video then there would not be any hassle but if there are multiple people appearing in the video then there exist various methods to track them. Before a person is tracked the background separation is achieved in order to perceive the object correctly. Once the background is separated to identify the object or person precisely there are selection points which are taken. The decision factors are similarly subtle to identify the object clearly. By this method, the averaged detection ratio has been extended by means of about 10% when in contrast to the traditional method. The primary drawback of it is very sensitive to light variations.

3. PROPOSED MODELLING

3.1 System Architecture and Implementation

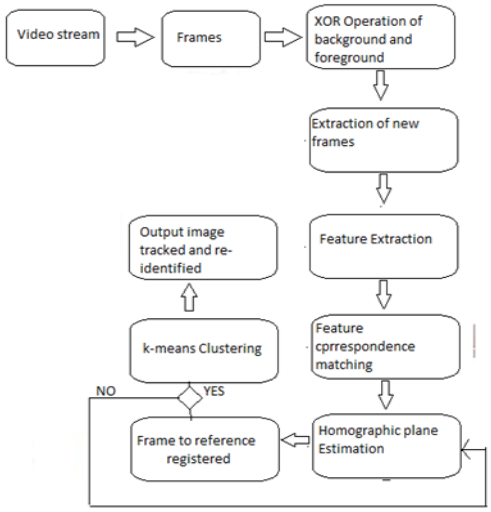

Fig 1: Architecture of detection and tracking of object

3.2 Background and Foreground Detection

Background modeling and foreground detection plays a vital role in video processing which is used in detection of motion based object in various challenging region robustly. It is preceded by detecting various foreground region in each and every frame of an image.

Fig 2: Example of Background subtraction

The process of separating moving objects is defined as foreground from the non-dynamic information is called as background.

Here, we use a static multiple cameras which are used to capture the video in our application domain. Due to continuous changes in the background regions like moving objects or humans then the video will be in quasi-stationary. The background of the video is said to be quasi-stationary if the multiple cameras are in static positions but if the background

keeps changing inherently. Detection of interested region from video sequences is an important task which requires a high level processing.

Hence, we propose a background subtraction methodology for detecting moving objects since in our application static cameras are mounted. Background subtraction is also called as foreground detection. Foreground detection is the first step in video processing which helps to detect a moving objects as in Fig. 2 . It deals with dynamic changes and illumination which requires a low memory in a real-time.

It helps in detecting an image such as human object, text or other moving object by performing XOR operation between current frame and the registered reference frames. This approach is often called as "Background image" or "Background model" In general background subtraction method is used for the identification process i.e. determination of people and objects identity. Background subtraction is applied in various applications such as surveillance tracking or human poses estimation. Background subtraction is based on a static background hypothesis which mostly not applicable in real-time environments like wavering flags and trees.

One of the conventional approach used in background subtraction technique is Median filtering. The median filter is a nonlinear digital filtering technique which is used to remove noise from an image. Thus noise reduction is a pre-processing step helps to improve the results of later processing like edge detection on an image .For calculating the image containing only the background, a series of preceding images are averaged. For calculating the background image at the instant *t*,

$$B(x, y, t) = \frac{1}{N} \sum_{i=1}^{N} V(x, y, t - i)$$

Foreground detection is then achieved by threshol Background subtraction is also known as foreground detection. Thus It helps in detecting an image such as human object, text or other moving object by performing XOR operation between current frame and the registered reference frames. This approach is often called as "Background image" or "Background model" In general background subtraction method is used for the identification process i.e. determination of people and objects re-identity. Background subtraction is applied in various applications such as surveillance tracking or human poses estimation. Background subtraction is based on a static background hypothesis which mostly not applicable in real-time environments like trees. One of the conventional approach used in background subtraction technique is Median filtering. The median filter is a nonlinear digital filtering technique which is used to remove noise from an image. Thus noise reduction is a pre-processing step helps to improve the results of later processing like edge detection on an image .For calculating the image containing only the background, a series of preceding images are averaged. For calculating the at the instant *t*

Where *N* denotes the number of preceding images which are taken for averaging. This averaging refers to average of total number of corresponding pixels of a given image. Here N depends on the total amount of movements in the video along with video speed i.e. total number of images per sec in the video. Firstly background is calculated $B(x, y, t)$ and later then subtract it from the image $V(x, y, t)$ at time t = t and thresholding is done. Thus the foreground is given by

$$|V(x, y, t) - B(x, y, t)| > \mathrm{Th}$$

Where Th denotes threshold. Similarly we are using concept of median instead of mean in the background calculations B(x, y, t). Due to Usage of global and time-independent thresholds (same Th value for all pixels in the given image) which may limit the accuracy of the above two conventional approaches.

3.3 SIFT based Algorithm

Fig 3: Example of matching of feature key points

The interest factors are decided by way of locating local extrema in the difference-of-Gaussian (Dog) scale-space. The SIFT functions are extracted for all such interest points detected. each function is categorized as corresponding to one of the foreground blobs or to the background. For each feature corresponding to a particular blob, we try to find a match with the capabilities determined in the next body inside a neighborhood window. We recall that a match is observed between SIFT capabilities when the Euclidean distance among the best match feature and the next best match function is more than 60%. All the matching features for a corresponding blob are averaged to find the displacement vector. The new position of the blob is decided by the previous position of the blob and the displacement vector.

Our technique proceed with a degree of background modeling and foreground extraction. SIFT features are extracted at points of interest within the regions detected as foreground, then used for matching from one frame to the subsequent. Every matched pair of SIFT functions gives a displacement vector, equal to the picture speed of the focus. Within a temporal buffer we gather function vectors that consist of this displacement, the image coordinates and the frame range. The frame range is normalized with the aid of the scale of the temporal window

taken into consideration. All feature vectors are processed the usage of an advanced k-means algorithm to discover the maximum dominant clusters, in which the number of clusters is determined the usage of a novel cluster data metric. The contribution of this work is the improvement of a sturdy, real-time SIFT- primarily based method for item monitoring that enforces the inherent temporal coherence throughout photo frames, consequently dealing with difficult situations brought about by using considerable object acceleration and partial occlusion. Fig. 3 shows an example of detected SIFT features in a given image. The arrows depicts the direction and position of the detection of the SIFT feature key points.

3.4 K means Clustering

K-means Clustering is one of the most effective unsupervised learning algorithms that deal with the well-known clustering problem. It is a partitioning approach. The feature k means partitions data into k mutually unique clusters, and returns the index of the cluster to which it has assigned each observation. K means treats each observation in your data as an object having a location in area. It finds a partition in which objects within every cluster are as near to each other as possible, and as far from objects in other clusters as viable.

Every cluster within the partition is defined by its member objects and by its centroid, or center in order to classify a given facts set into a certain wide variety of k clusters fixed a priori, the algorithm defines k centroid, one for each cluster. The centroid for each and every cluster is said as the point to which the sum of distances from all objects in that cluster is to be minimized. K means computes cluster centroid differently for every distance metric, to minimize the sum with respect to the measure that you specify.

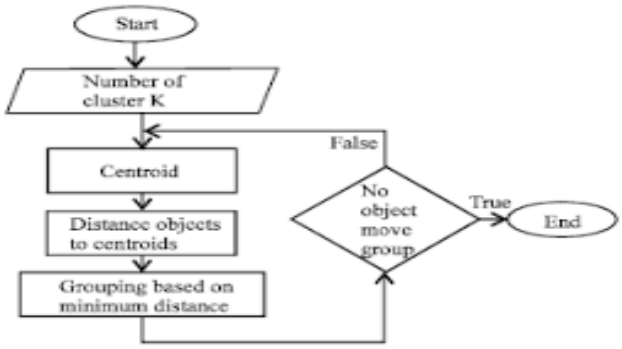

Fig 4: Workflow of k-means clustering

We can manage the details of the minimization the use of numerous input parameters to k means, including preliminary values of the cluster centroid and for the most range of iterations. k-means might also produce a poor clustering as the very last result rather relies upon on preliminary cluster picks .A less difficult manner to prevent poor clustering due to insufficient seeding is to alter the fundamental k means algorithm to achieve the improved k means algorithm. In improved k-means algorithm, we begin with a huge range of uniformly distributed seeds in the bounded five- dimensional feature area, but we reduce them extensively by removing the ones that are too close to each other.

At some stage in the iteration of k-means clustering, the detected target center might be updated in both the 2d position space and the 3-D shade area to follow the illumination modifications. However, rapid color shifting (e.g. high light reflecting) might also occur because of the glossy surface of the target object (e.g. eyeballs). If we simply follow the ordinary update principle, the goal belongings (here, the shade) can be completely destroyed. To solve this hassle, we use a simple averaging filter to regularly lower the effect of rapid color adjustments as:

$f(t)\ T = \gamma f(new)\ T + (1 - \gamma)\ f(t-1)$

$0 < \gamma < 1$ is the pre-defined coefficient(e.g. 0.6) , the superscript (t),(new),(t−1) denote the time, f (new) T represents the new target feature after clustering in frame t, f (t−1) T means target feature in frame t − 1.This ratio is calculated for all mixtures of clusters and the two clusters for which the ratio is maximal are selected to be merged. using this merging technique together with the minimum distance merging approach has been shown to produce very right outcomes in most of our experiments.

THE K-MEANS CLUSTERING ALGORITHM

Description of Algorithm Programming

1 Choose k center of cluster randomly

2 Assign initial values for cluster means c1to ck

3 Repeat (Repeat the following steps until the cluster no longer changes)

4 for i=1 to n do

5 Assign each data point to cluster where || Ci-Cj || is the minimum

6 end for

7 for j=1 to K do

8 Recalculate cluster mean of cluster

9 end for

10 until convergence

11 return C

4. EXPERIMENTAL ANALYSIS

We used the proposed approach for tracking and re-identification of an object using SIFT based algorithm and k-

means clustering in a video surveillance application. Fig 5 shows an example of human detected using the conventional approach of background subtraction i.e. median filtering where the noise and the background is removed using XOR operation and only the human is detected.

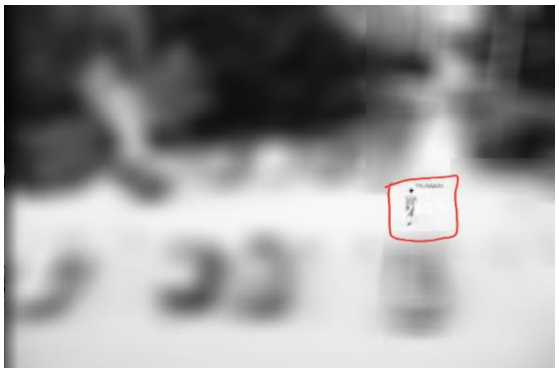

Fig 5: Background subtraction using media Filtering Technique

Fig 6: Human is detected in Yellow color box

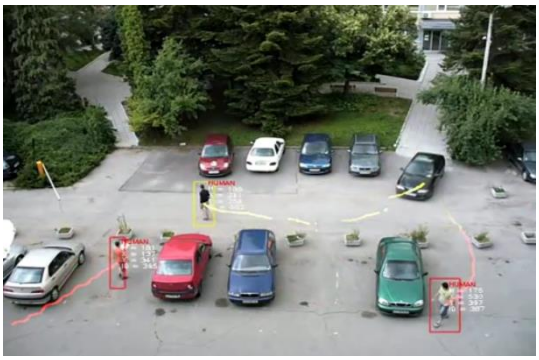

Fig 7: The same human is re-identified and Tracked in red color box and red mark Traces

Here the yellow square represents the bounding box for the detection of human and the red box represents the re-identification of the same human along with the red traces mark which denotes the tracking of that respective person. Fig 6 shows that a human is identified in yellow color box in the foreground region. Fig 7 shows an example of re-identification and tracking of same human using piecewise feature clustering of same interested points. The successful tracking and re-identification of an object shows that our proposed is able to work even with such changes in the image velocity (which may be quite significant when the human is close to camera).

5. CONCLUSIONS AND FUTURE WORK

In this paper we have described a novel approach for object tracking using SIFT algorithm and improved k-means algorithm . Our experimental results show the contribution of this work, real-time SIFT- based approach that tracks the object accurately and also handles and overcomes the challenges like partial occlusion, image frames.

Although the proposed algorithm works well in general, the performance might decrease when there are large number of frames. As the number of frames are more the processing time is increased. When there are many feature points nearby then the cluster heads will be very close and two clusters might merge together. So the new cluster might group with various distinct clusters which eventually forms a similar cluster.

Here the clustering is done piecewise where it is done for each piece separately. This problem can be overcome if the data from the previous frame can be used for the next frame. Further the work can be expanded to finding the facial expressions of human in the frame and accurately identify the person accurately.

REFERENCES

[1] V. Ferrari, T. Tuytelaars, and L. Van Gool, “Simultaneous Object Recognition and Segmentation by Image Exploration,” Proc. Eighth European Conf. Computer Vision, pp. 40-54, 2004.

[2] S. Lazebnik, C. Schmid, and J. Ponce, “Sparse Texture Representation Using Affine-Invariant Neighborhoods,” Proc. Conf. Computer Vision and Pattern Recognition, pp. 319-324, 2003.

[3] K. Mikolajczyk and C. Schmid, “Indexing Based on Scale Invariant Interest Points,” Proc. Eighth Int’l Conf. Computer Vision, pp. 525-531, 2001.

[4] R. Polana and R. Nelson, “Low level recognition of human motion,” Proceedings of IEEE Workshop Motion of NonRigid and Articulated Objects, pp. 77–82, 1994.

[5] T. J. Fan, G. Medioni, and G. Nevatia, “Recognizing 3-D objects using surface descriptions,” IEEE Trans. Pattern Recognit. Machine Intell. v. 11, pp. 1140–1157, 1989.

[6] C. Bregler, “Learning and recognizing human dynamics in video sequences,” IEEE Computer Vision and Pattern Recognition (CVPR), June 1997.

[7] H. Greenspan, J. Goldberger, A. Mayer, “Probabilistic space-time video modeling via piecewise GMM,” IEEE Transactions on Pattern Analysis and Machine Intelligence, v. 26(3), pp. 384 – 396, March 2004.

[8] Tavakkoli, M. Nicolescu, G. Bebis, “Robust Recursive Learning for Foreground Region Detection in Videos with Quasi-Stationary Backgrounds,” Proceedings of International Conference on Pattern Recognition, pp. 315318, 2006.