Impact Factor (SJIF): 4.542

International Journal of Advance Research in Engineering, Science & Technology

e-ISSN: 2393-9877, p-ISSN: 2394-2444

Volume 4, Issue 3, March-2017

SCALE INVARIANT FEATURE EXTRACTION METHODS: A BROAD STUDY

Jaykumar Limbasiya[1], Hitul Marvaniya[2], Radhika Kotecha[3]

[1]*PG Student, Information Technology Department, V.V.P. Engineering College, Rajkot, India*
[2]*Assistant Professor, Information Technology Department, V.V.P. Engineering College, Rajkot, India*
[3]*Assistant Professor, Information Technology Department, V.V.P. Engineering College, Rajkot, India*

Abstract — *Object detection and tracking are important and challenging tasks in many computer vision applications such as surveillance, vehicle navigation, and autonomous robot navigation. Object tracking becomes even more challenging in the presence of variable illumination conditions, background motion, complex shaped object, partial and full object occlusions, etc. Furthermore, detecting and tracking of the human body are very important to understand the human activity. Several human detection and tracking systems have been developed which use a new class of local image features that are invariant to image scaling, translation and rotation. In addition, these features are partially invariant to illumination changes, affine or 3D projection. In any surveillance system, the camera location is stationary, but there is the change in ambience, orientation, and scaling of the human body. This paper describes scale invariant feature extraction methods for human detection and tracking from surveillance video, which originally consists of two components: feature extraction for human detection and tracking. A detailed study of these methods is presented in the paper along with the advantages and disadvantages of each of these methods.*

Keywords- *Scale Invariant Feature; Object Detection; Feature Extraction; Object Tracking; Human Detection & Tracking*

I. INTRODUCTION

The detection and tracking of the human being in Video surveillance is a vibrant research topic in computer vision. It replaces old traditional method of monitoring camera feed by the human being by automatically detect and tracking along with the understanding of human behaviour. Human detection and tracking are vibrant and challenging tasks in several computer vision applications such as surveillance system, vehicle navigation and self-directed robot navigation. Human detection includes detecting humans in each frame of video. Either each tracking method requires an object detection mechanism in every frame or when the object first appears in the video [1]. Human tracking is the method of detecting human over time on the camera feed. The fast and efficient camera, high-quality videos, low – priced camera and increasing need of the automatic system for detection and tracking will create interest in human tracking algorithms.

There is four key step in video analysis, i.e. Pre- processing, Feature Extraction, Ob ject Detection, Object Tracking. The pre-processing step is required on the sequence of frames if the sequence of frames having the area of low contrast and any other such bad condition. Pre- processing step prepares the sequence of frames from the video for the further video analysis steps. It is easy to implement further step if the pre-processing step is implemented correctly.

Autonomous human detection and tracking is the critical task for the wide range of application like in-house, business building, industries, urban development and traffic management. These applications were not including consumer electronics because it strongly requires perfect working condition, special and costly hardware, complex instalment and difficult setup steps. A large amount of work in computer vision has been done to build autonomous system, which automatically detects track human being, which leads to reduce the effort required to monitor manually.

Surveillance systems must be self-directed to improve the performance and eliminate such operator errors. Ideally, an automated surveillance system should only require the objectives of an application, in which there are real-time understanding of activities. Then, the challenge is to provide robust and real-time performing surveillance systems at a reasonable price. As the cost of hardware decreases for sensing and computing, and the processor speed increase, surveillance systems have become commercially available, and they are now applied to a number of different applications, such as traffic monitoring, airport and bank security, etc. However, occlusions, shadows, weather conditions, etc. likes shortcoming still affect machine vision algorithms (especially for a single camera).

After studying the literature, it is found that detecting the object from the video sequence and track them throughout video is a challenging task. Object tracking can be a time- consuming process due to the amount of data that is contained in the video. From the literature survey, it is found that there are many background subtraction algorithm exist, which work efficiently in both indoor and outdoor surveillance system [2].

It will be better if the shadow will be removed at the time of the foreground object detection by designing an efficient algorithm, which can properly classify the foreground object and background by removing false foreground pixel from detection [3]. Then there will no extra computation needed for shadow detection and removal.

Automatic tracking of objects can be the foundation for many interesting applications. A precise and proficient tracking capability is required at the base of such a for building higher-level vision-based intelligence. Tracking is also very importatnt process resulted into their motion, the non-deterministic nature of the subjects, and the image capture process itself.

It also very widely used in traffic monitoring, automated surveillance, Motion-based recognition, video indexing, human-computer interaction, vehicle navigation.

Next, Section-2 describes the basic understanding of Feature Extraction and object detection and tracking, Section-3 consists of detail survey of point-based methods for object detection and tracking, Section-4 is the observations derived based on the survey done in this paper.

II. FEATURE EXTRACTION & OBJECT DETECTION AND TRACKING

Object tracking is a very challenging task in the presence of variability Illumination condition, background motion, complex object shape, partial and full object occlusions. In [3], the modification is done to overcome the problem of illumination variation and background clutter such as fake motion due to the leaves of the trees, water flowing, or flag waving in the wind. Sometimes object tracking involves tracking of a single interested object and that is done using normalized correlation coefficient and updating the template.

In [4] very basic framework to detect and track moving object is very straight forward. When the video is ready for the video analysis step, the feature extraction method extracts the candidate feature from the frame extracted from the sequence of the frame from the video. The candidate feature may describe the feature of the interested object and other objects too.

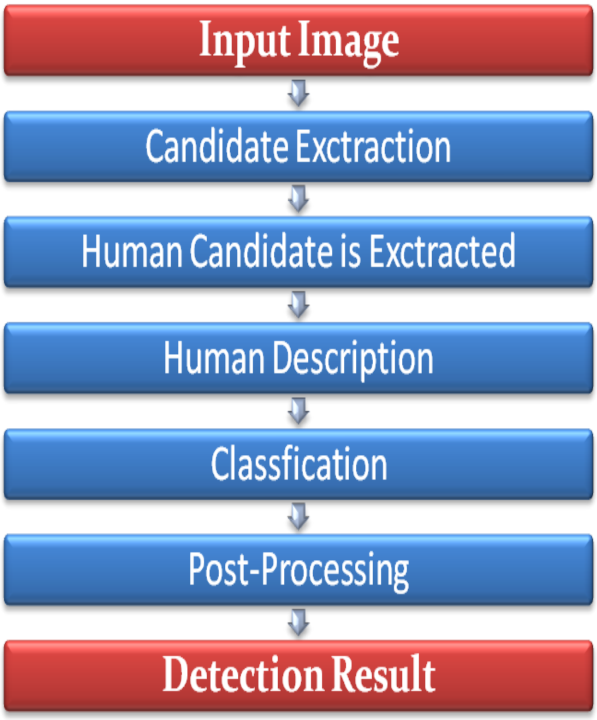

Figure 1. A General Human Detection Framework

As shown in Fig. 1, human detection process, it may include the candidate feature of both human as well as non-human like the statue. A human descriptor that is already stored will help us to detect the interesting feature, which describes the human being.

In [5], the process of detecting human objects from images/videos is demonstrated, in which the exact process to be carried out in the following manner steps: the candidate regions are to be extracted that are hypothetically covered by human objects, then that extracted regions is described, after that the exctracted regioons is being classified as human and non-human, and finally the post-processing is carried out (e.g. integrating the positive regions or adjusting the size of those regions).

This process is illustrated in Fig. 2. Note that this framework different from that presented. In particular, foreground segmentation proposed in is not assumed in our framework. This is because the segmentation step is not mandatory in current human detection systems, e.g. Moreover, the tracking step in is also not appropriate for detecting humans from fixed images.

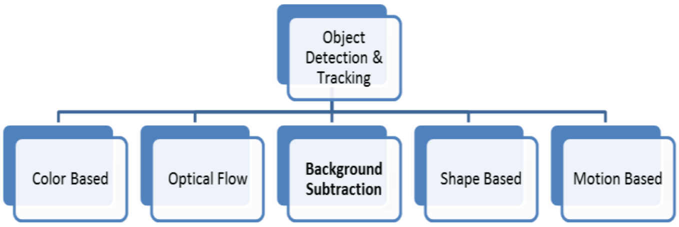

Figure 2. Techniques for Object Detection and Tracking

Background Model is built based on video sequence and each video sequence is compared with that Background model to Detect Object. Background subtraction, also known as Foreground Detection, is a technique in the fields of image processing and computer vision wherein an image's foreground is extracted for further processing (object recognition etc.) [6]. generally, a region of image of interest are objects in its foreground. After the stage of image pre-processing, (which may include removing of noise, post processing like morphology etc.) object localization is required, which may make use of this technique. Background subtraction is a widely used approach for detecting moving objects in videos from static cameras. It is highly dependent on background availability; it provides a complete object information if the background is known. However, it has poor anti-interference ability.

Firstly, it calculates the image optical flow field using displacement vector of each pixel in the region, and do clustering processing on image optical flow field. Optical flow or optic flow is the pattern of deceptive motion of objects, surfaces, and edges in a visual scene caused by the relative motion between an observer and a scene [7]. The concept of Optical flow is introduced by an American psychologist James J. Gibson. Optical flow used to describe the visual stimulus provided to animals moving through the world. Gibson stressed the importance of optic flow, which deals with the ability to detect possibilities for action within the environment. Optical flow method relies on optical flow field; therefore, we can get the complete movement information. It required large calculation and it is also sensitive to noise, therefore, it has poor anti-noise performance.

Color based creates a Gaussian Mixture Model to describe the color distribution within the sequence of images and to segment the image into background and objects [8]. It has the low computational cost of the algorithms because it depends on Gaussian mixture model and many pixels has the same or nearer value. However, Colors might not always appropriate for detecting and tracking objects because any object in the real world has lakhs of color shades.

Shape based uses different descriptions of shape information of motion regions such as representations of points, boxes and blobs are available for classifying moving objects [8]. A Simple pattern-matching approach can be applied with appropriate templates. Many objects may not have fixed shape and therefore it does not work well in dynamic situations.

Non-rigid articulated object motion shows a periodic property, so this has been used as a strong cue for moving object detection [9]. It does not require predefined pattern templates as in shape based methods. There might be non-moving human being, so Motion based method struggles for detection of non-moving human being.

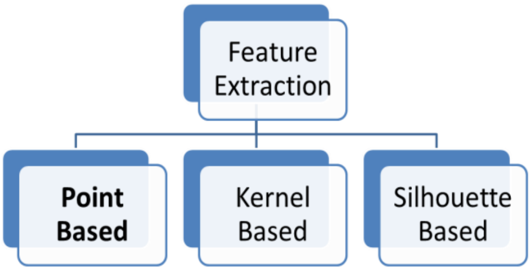

Figure 3. Techniques for Feature Extraction

Kernel tracking is usually performed by computing the moving object, which is represented by an embryonic object region, from one frame to the next. It is very suitable for Real Time Tracking. The problem arises when parts of the objects may be left outside of the defined shape while portions of the background may exist inside.

The aim of a silhouette-based object tracking is to find the object region in every frame by means of an object model generated by the previous frames [10]. It is Capable of dealing with the variety of object shapes, Occlusion and objects split and merge. Whenever the scenario is of real- time streaming video, it is Difficult to generate Model from each frame. In an image structure, their feature points represent moving objects during tracking. Recognition can be done relatively simple, by thresholding; identification of these points might become easy [6]. Sometimes, it becomes complex in the incidence of occlusions.

III. LITERATURE SURVEY

There are numerous methods for extracting distinctive invariant features from images that can be used to perform consistent matching between different views of an object or scene. The features are invariant to image scale and rotation and are shown to provide robust matching across a substantial range of affine distortion, change in 3D viewpoint, addition of noise, and change in illumination.

The features are highly distinguishing, in the way such that a single feature can be precisely matched with high probability against a large database of features from many images [11]. The recognition of the object is carried out by comparing individual features to the database of feature extracted from the known object using fast nearest-neighbor algorithm which is followed by transformation to identify clusters belonging to a single object using Hough transform and finally verification is performed through the least-squares solution for consistent pose parameters. This approach to recognition can vigorously identify objects among clutter and occlusion while achieving near real-time performance.

3.1. Scale Invariant Feature Transform (SIFT)

Image matching is a fundamental phase of many problems in computer vision, including entity or scene recognition, solving for 3D structure from multiple images, stereo correspondence, and motion tracking. In [12], image features consist of many properties which build them ready for suitably matching different images of an object or scene. The features then extracted are invariant to image scaling and rotation, and partially invariant to change in illumination and 3D camera viewpoint. They have widely spread in both the domains i.e. spatial and frequency domains, which reduce the probability of disturbance by occlusion, clutter, or noise [13]. Large numbers of features can be extracted from typical images with many of efficient algorithms available. In addition, the basis for object and scene recognition is that a single feature can be correctly matched with a higher probability with the large sized database of features although the features are highly distinctive.

The cascade filtering approach is to be used to minimise the cost of extracting features, in this filtering approach the most costly operations is being applied to the location that passes through the initial test. Following are the major stages of computation used to generate the set of image features [12]:

3.1.1. Scale-space extrema detection

This phase includes computation to searches over all possible and defined scales and image locations too. It is implemented proficiently to identify potential interest points that are invariant to scale and orientation by using a difference-of-Gaussian (DoG) function [14].

3.1.2. Keypoint localization

In this phase the Keypoints are being selected based on measured stability, which is present at every candidate point locations, a detailed model is used to determine location and scale [15].

3.1.3. Orientation assignment

In Orientation assignment phase, there can be the possibility of assigning more than one orientations each keypoint locations calculated in above phase based on their local image gradient directions. All further operations are being performed on image data that has been transformed relative to the assigned orientation, scale, and location for each feature, thereby providing invariance to these transformations [16].

3.1.4. Keypoint descriptor

The local image gradients are measured at the selected scale in the region around each keypoint. These are transformed into a representation that allows for significant levels of local shape distortion and change in illumination.

This approach has been named the Scale Invariant Feature Transform (SIFT), as it transforms image data into scale-invariant coordinates relative to local features [12].

3.2. Speeded Up Robust Features (SURF)

It is a unique scale and rotation-invariant interest point detector and descriptor, called SURF (Speeded Up Robust Features) [17]. It has higher repeatability, distinctiveness, and robustness compares to previously proposed schemes to approximate or has higher performance than that of previous methods, yet can be computed and compared much faster.

The task of finding the relation between two images of the same scene or object is part of many computer vision applications. Camera calibration, 3D reconstruction, image registration, and object recognition are just a few methods of computer vision that deals with such applications. The goal of this work is to search for correspondences through discrete image, there for the goal of this work can be divided into three main steps [17]. First, the different locations in images such as such as corners, blobs, and T-junctions is being searched for "Interest points‟. The most important property of an interest point detector is its repeatability, i.e. whether it reliably finds the same interest points under different viewing conditions. Next, a feature vector is calculated for each and every neighborhood of every interest point. This descriptor has to be robust to noise, detection errors, and geometric and photometric deformations as well as distinctive at the same time. Finally, the descriptor vectors are matched between different images. The matching is often based on a distance between the vectors, e.g. the Mahanalobis or Euclidean distance. The dimension of the descriptor has a direct impact on the time this takes, and a lower number of dimensions is, therefore, desirable [18].

A novel detector-descriptor scheme, called SURF (Speeded-Up Robust Features). The detector uses the Hessian matrix, but it is based on a very basic approximation, it is a very basic Laplacian-based detector just as DoG [19]. It depends on integral images to reduce the computation time and therefore, sometimes it also called as "Fast-Hessian" detector. The descriptor describes a distribution of Haar-wavelet responses throughout at each neighborhood of the interest point. Again, they exploit integral images for speed.

Furthermore, this method uses only 64 dimensions, which increases robustness and reduces the time required for feature vector calculation and matching that feature vector with others [18]. It also introduces a new indexing step based on the sign of the Laplacian, which increases not only the matching speed but also the robustness of the descriptor.

3.3. Binary Robust Independent Elementary Features (BRIEF)

This method uses binary strings as an efficient feature point descriptor, which is called BRIEF. BRIEF can compute using simple intensity differences tests and it also comes with the capability of being highly discriminative even when using relatively few bits [20].

BRIEF is very fast for both to compute and to match. It resulted in a similar or better recognition performance while running in a little bit of the time required by others. It was shown in [20] that floating-point value could be quantized for the descriptor vector using very few bits per value without the cost of recognition performance. This whole computation can be shortened by directly computing binary strings from image patches.

The individual bits are computed by just associating the intensities of pairs of points along the same lines as in but this step doesn't require any training phase, therefore BRIEF is very efficient in both computations as well as to store in memory [20]. Furthermore, the Hamming distance is used to compare strings which generally uses bit count or XOR Operation, which can be done tremendously fast on modern CPUs that often provide a specific instruction to perform a XOR or bit count operation more rapidly, as is the case such that latest SSE instruction set. Therefore, it furthermore overtakes them in terms of recognition rate in many cases, as demonstrated in [20] using benchmark datasets.

3.4. Oriented Fast and Rotated Brief (ORB)

The necessity of rotation invariant is full filled by the new descriptor called ORB (Oriented Fast and Rotated Brief). The SIFT keypoint detector and descriptor established a standard benchmark for many typical types of such applications visual features, including object recognition, image stitching, visual mapping, etc., although SIFT is over a decade old. Nevertheless, it imposes a large computational encumbrance, particularly for the application over low-powered devices such as cell phones and for real- time systems such as visual odometry. This has led to a rigorous search for substitutions with lower computation cost; questionably, the best of these is SURF. In [21], specify that there are many researcher whose research aimed to speed up the computation of SIFT, conspicuously with GPU devices.

In [21] a survey is carried out on the computationally efficient replacement to SIFT such that it has similar or better matching performance and it is less pretentious by image noise and it should have the ability to be used for real-time performance. In [21], main inspiration is to improve many common image-matching applications, e.g., to perform panorama stitching, to enable low-power devices without GPU acceleration, and patch tracking and to reduce the time for feature-based object detection on regular PCs.

In ORB, a new Fast and accurate orientation component are added such as in FAST. In oriented BRIEF, a new efficient technique which uses analysis of variance and correlation for computation of features is presented. The decorrelating BRIEF features using a learning method under rotational invariance, leading to better performance in the nearest-neighbor application is presented in [20].

IV. OBSERVATIONS

After Studying Many Scale Invariant Feature Extraction Methods for Human Detection and Tracking, it is observed that there are many limitations of different methods. SIFT having high computation; therefore, it is slow for real-time surveillance video. SURF code can also be optimizing for additional speed. BRIEF method also somehow lack for orientation and scale invariance. ORB focuses on orientation but it lacks scale invariance up to some instant. Therefore, there is no method, which could be taken as perfect with complete working accuracy.

V. CONCLUSION AND FUTURE WORK

Feature extraction is a key step for human detection and tracking. From the survey conducted and after analysis from several perspectives, it is concluded that SIFT is an efficient technique for feature extraction. As a part of future work, SIFT can be improved by using the integral image whenever the calculation is required to be carried out on the image. To improve the matching and to reduce computation burden, the indexing step of SURF method can be used. Also, the size of descriptor vector can be reduced by rounding off the floating point value of the descriptor vector. An additional orientation component can be added to increase the accuracy for matching rotated images.

REFERENCES

[1] A. Yilmaz, O. Javed, and M. Shah, “Object tracking: A survey”, Acm Computing Surveys (CSUR), 38(4):13, 2006.

[2] C. Stauffer and W. Grimson, “ Adaptive background mixture models for real-time tracking”, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, volume 2. IEEE, 1999.

[3] J. Jacques, C. Jung, and S. Musse, “Background subtraction and shadow detection in grayscale video sequences”, Computer Graphics and Image Processing IEEE, SIBGRAPI 2005. 18th Brazilian Symposium, pages 189–196, 2005.

[4] D. Nguyen, W. Li and P. Ogunbona, “Human Detection from Images and Videos: A Survey”, Pattern Recognition, http://dx.doi.org/10.1016/j.patcog, 2015

[5] H. Parekh, D. Thakore, U. Jaliya, “A Survey on Object Detection and Tracking Methods”, IJIRCCE, Vol. 2, Issue 2, February 2014

[6] J. Athanesious, P. Suresh, “Systematic Survey on Object Tracking Methods in Video”, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) 242-247, October 2012.

[7] A. Chauhan, P. Krishan, “Moving Object Tracking Using Gaussian Mixture Model and Optical Flow”, International Journal of Advanced Research in Computer Science and Software Engineering, April 2013

[8] M.Sankari, C. Meena, “Estimation of Dynamic Background and Object Detection in Noisy Visual Surveillance”, International Journal of Advanced Computer Science and Applications, 77-83, 2011.

[9] H. Patel, D. Thakore, “Moving Object Tracking Using Kalman Filter”, International Journal of Computer Science and Mobile Computing, pg.326 – 332A, pril 2013.

[10] J. Athanesious, P. Suresh,” Implementation and Comparison of Kernel and Silhouette Based Object Tracking”, International Journal of Advanced Research in Computer Engineering & Technology, pp 1298-1303, March 2013.

[11] “Scale-invariant feature transform”, https://en.wikipedia.org/wiki/Scale-invariant_feature_transform

[12] D. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints”, International Proc. of the International Conference on Computer Vision, Corfu, September, 1999.

[13] A. Witkin, “Scale-space filtering. In International Joint Conference on Artificial Intelligence”, pp. 1019-1022, Karlsruhe, Germany, 1983.

[14] J. Koenderink., “The structure of images.”, Biological Cybernetics 50:363-396, 1984.

[15] T. Lindeberg, “Scale-space theory: A basic tool for analysing structures at different scales”, Journal of Applied Statistics, 21(2):224-270, 1994.

[16] M. Brown and D. Lowe, “Invariant features from interest point groups. In British Machine Vision Conference”, pp. 656-665, Cardiff, Wales 2002.

[17] H. Bay, T. Tuytelaars, L. Gool, “SURF: Speeded Up Robust Features”, Computer Vision and Image Understanding Volume 110, Issue 3, Pages 346–359, June 2008.

[18] D. Lowe, “Distinctive image features from scale-invariant keypoints, cascade filtering approach”, IJCV 60, 91 – 110, 2004.

[19] K. Mikolajczyk, C. Schmid, “A performance evaluation of local descriptors”, PAMI, 27, 1615–1630, 2005.

[20] M. Calonder, V. Lepetit, C. Strecha, P. Fua, “BRIEF: Binary Robust Independent Elementary Features”, Computer Vision – ECCV, Volume 6314 of the series, pp 778-792, 2010.

[21] E. Rublee, V. Rabaud, K. Konolige, G. Bradski, “ORB: an efficient alternative to SIFT or SURF”, Computer Vision (ICCV), IEEE International Conference on Computer Vision, Nov. 2011