Vision Based Human Tracking:An Overview

Dr. Anshul Pareek
Deptt Of ECE,
Maharaja Surajmal Institute Of Technology,
New Delhi,India
er.anshulpareek@msit.in

Dr Poonam
Deptt Of ECE,
Maharaja Surajmal Institute Of Technology,
New Delhi,India
poonam.dahiya@msit.in

Dr. Shaifali Madan Arora
Deptt Of ECE,
Maharaja Surajmal Institute Of Technology,
New Delhi,India
shaifali04@msit.in

Abstract—
Human/object tracking when done on an array of frames is an operation of tracking any mobile target object over a span of time with the help of any mobile or immobile camera. It has been a critical issue in the arena of computer vision as it is used in a number of application fields like security, surveillance, human-computer interaction, augmented reality, video communication, and compression, medical imaging, traffic control, video editing, and assistive robotics. This is a highly studied problem and remains to be a complex problem to solve. In object tracking in any given video, the major task is to trace the target object in upcoming video frames. To track the object efficiently a complete knowledge of the target is needed, also the tracking parameters are required and the type of video is being analyzed is to be known. Object tracking is a principal segment of human-computer interaction in a real-time environment, where the computer obtains a finer model of a real-time world.

Keywords—Vision Based Tracking, Object Tracking, Feature Descriptors
DOI Number: 10.48047/NQ.2022.20.13.NQ88453 **Neuroquantology 2022; 20(13): 3691-3698**

I. INTRODUCTION

1.1 Vision-based tracking

How do we feel about the world we live in? It is by looking and seeing. There are countless objects present in our surrounding environment with their impressions. Vision can be better explained as a way to understand the environment that surrounds us. Even after decades, the exact working of the visual system is a mystery for the scientist involved in its investigation. When eye based vision in living creatures is replaced by computational instruments it is defined as computer vision. In other words, computer vision is artificial mimicry of vision in living creatures, where digital images and videos which are captured by cameras are further analyzed by computers obtaining an optimum level of understanding from it. Human/object tracking when done on an array of frames is an operation of tracking any mobile target object over a span of time with the help of any mobile or immobile camera. It has been a critical issue in the arena of computer vision as it is used in a number of application fields like security, surveillance, human-computer interaction, augmented reality, video communication, and compression, medical imaging, traffic control, video editing, and assistive robotics [1-3]. This is ahighly studied problem and remains to be a complex problem to solve. In object tracking in any given video, the major task is to trace the target object in upcoming video framesTo track the object efficiently a complete knowledge of the target is needed,also the tracking parameters are required and the type of video is being analyzed is to beknown[13].

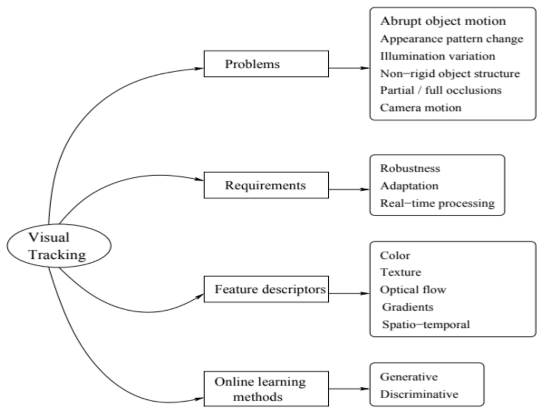


Fig 1: Visual tracking at a glance.

also the tracking parameters are required and the type of video is being analyzed is to be known[13].

Object tracking is a principal segment of human-computer interaction in a real-time environment, where the computer obtains a finer model of a real-time world. For example, when autonomous vehicles are talked about, a human being cannot transmit the exact state of surroundings precisely and speedily enough.

The wide-ranging scope of the application review the significance of dependable, exact, and efficacious object tracking. To obtain an effective tracking the two most important parameters to be included are first, selection of the model and secondly, the tracking method worthy for the task.

The fundamental necessities of any tracking structure are first, a robust system, secondly, an adaptive system, and lastly, real-time processing requirement [3-4]. The famous state-of-art tracking strategies are Interest point-based tracking [10], multiple hypothesis tracking [5-6], kernel-based tracking [7-8], and optical flow-based tracking [9]. This area has observed a remarkable elevation due to available low cost, advanced technology cameras, and low computing complexity, corresponding to the inclination of ingenious approaches for image and video processing. Excellent reviews on the state-of-art techniques in this area have been provided in [11-12]. An overview of vision-based tracking is provided in Fig 1.1.

1.1.1 Difficulties in visual tracking

Regardless of the achievements acquired in the tracking field, it still faces challenges that may crop up because of any reasons stated below.

• Unanticipated object motion.
• Partial/ full occlusion
• Illumination variation.
• Non-rigid object structure.
• Pose change.
• Camera motion.

1.1.2 Required features of visual tracking

To develop a robust visual tracking system, these difficulties should be addressed properly along-with some basic requirements of visual tracking.

• Robustness: The tracking algorithm should be able to follow the target object even in adverse conditions.
• Adaptation: The tracking algorithm should be adaptive to the changes in the environment as well as to the changes in the target itself.
• Real-time processing: The algorithm should be computationally least intensive to enable it to be implementable in real-time. This is one of the primary concerns as the work is related to the navigation of the robot while tracking a target human.

The first step for visual object tracking is to have a depiction of the target being tracked. Description can be anything like color, texture, shape, or a template image of the object. The details of the feature descriptors commonly used for visual tracking are provided in the next subsection.

1.1.3 Feature descriptors for visual tracking

In visual tracking, image feature selection is a critical task. The basic entity required for feature points selected is its onliest property which differentiates it in the feature space. The visual features often used are described as mentioned below :

Color:- Two physical factors influence the appearance of color, one is the spectral power distribution of the illuminant and the other is the surface reflectance properties of the object. RGB which is a non-uniform color space is commonly used to define color in image processing. Whereas HSV is defined to be one of the uniform spaces.

Recent advances divide the color descriptors into two categories that are histogram and SIFT-based In [3], it is declared that when the hue samples are weighted by their saturation it results in a robust hue histogram. The authors also proposed a link of hue with SIFT. RG histogram based on RGB color model is used by Geverset al.[4]

• Texture:- Texture computes the variation of illumination on any surface by finding its levelness [5-6]. The most common feature is the Gabor wavelet [7]. In recent years, the focus is laid on attaining superior detection by finding local patterns of the given image.

Ojala*et al.* [8] proposed a gray-scale invariant texture measure that exhibits high tolerance against intensity variation and low computation time.

• Optical Flow:- Optical flow can be defined as an intense area containing the displacement vectors of pixel translation. The luster constancy is assumed of the proximate pixels in successive frames [9]. Lucas and Kanade tracker a famous one [10].

• Gradient:- The gradient features are mainly divided into two categories. The first is to define the contours[11] and in the second statistical summary of gradients is used. Lowe[12] proposed SIFT for detection. Bay *et al.* introduced SURF [13], which is three times speedier than SIFT. Dalal and Triggs [14] gave the most robust algorithm ever based on the Histogram of Oriented Gradient (HOG) descriptor to train the SVM classifier.

• Spatio-temporal features:- Local space-time features are very useful in extracting motion patterns by providing spatio-temporal shifts [15].

Selecting a feature for an initial object model description is a crucial task, there is a linear relationship between the quality of the description to the success of the tracking. Even after having a high-quality object description given a priori, continual adaptation to appearance change is necessary to achieve robust tracking. The important task is to manage the appearance changes of the target object. The appearance variations can be divided into two categories (i) intrinsic, and (ii) extrinsic. When intrinsic variation is stated it includes targets Pose change and shape malformation. Whereas any changes occur due to full/partial occlusion, illumination variation, and unanticipated camera motion count under the extrinsic category. These variations can only be handled with adaptive methods that can incrementally update their representations. Thus, there is a need for online algorithms that can learn continuously. A brief description of the online learning algorithms is presented in the following subsection.

1.1.4 Online learning algorithms

There are two categories for Online algorithms generative and discriminative. In former, model is generally updated in online mode whereas in the latter offline training using datasets is used.

1.1.5 Applications

There are a lot of applications that require object tracking by a robust tracker. Object tracking is applied to various fields from electron microscopy images [16-19] to understanding trajectories[20-22] and many others. A few of these applications are recorded below.

- Visual surveillance:
- Human-Computer Interface (HCI)
- Virtual and Augmented Reality
- Robotics
- Sports Analysis
- Customer Analysis.
- Biomedical Analysis

2. Object detection and tracking:

In computer vision detection and tracking are crucial challenges to meet to date. When we say object detection it means locating and detecting the target object in all the frames of the given video sequence. Whereas, tracking means keeping a record of all the detected images concerning time. In other words, the trajectory of the target is computed in the image plane. Researchers are showing immense interest in this field because of the wide area of applications, increase in computational power, and last but not least low-cost cameras with high picture quality. The video analysis involves three major steps

- Detection of target among other moving objects.
- Tracking of the target in real-time scenario facing all challenges.
- To understand the behavior, an analysis of the trajectory is done.

2.1 Object representation

Object tracking is implemented in a number of applications. The target in each case is different depending on the application like in the medical field it can be human or the cluster of cells in the body. The same way the traffic is dealt with, the target is either car or human. When satellites are talked about, planets are the targets. Animals in zoo applications, missiles in defense, humans in the supermarket, etc. are a few of many examples where tracking is implemented. The target object can be defined by its shape and appearance. Object representation comes into the picture when it is decided how it is to be presented by the tracker. Some of the state-of-art methods are discussed below.

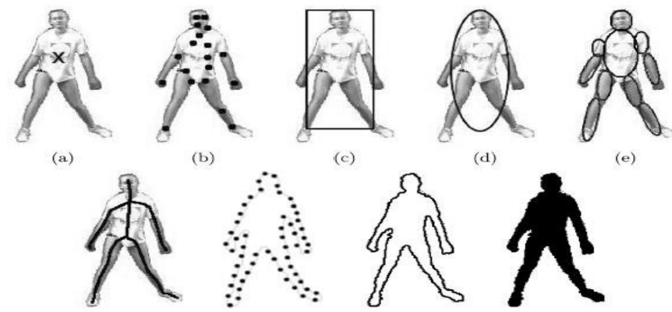
- Point: The target is represented by a single point or multiple points. When single point representation is considered it is generally the centroid of the target. This representation is used when the surroundings are small or the target is small in itself.

- Geometric shapes: In this case, the target is represented by shapes such as square, circle, rectangle, or ellipse. This is commonly used for rigid objects, also successful when simple non-rigid objects are being tracked.
- Silhouette and contour: This technique is used for complex non-rigid objects. Boundary drawn over the target is termed as its contour whereas silhouette is the region that is around this contour.
- Articulated shape models: Articulated shapes are the segments in the body that are held by joints like hands, legs, and torso. These segments are represented by geometric shapes like rectangle ellipse etc. The relationship among the body parts is established by using some kinetic motion model.
- Skeletal models: To represent the target objects its skeleton is created inside its silhouette or its boundary. It represents the shape of the target. Successfully used for rigid and non-rigid structures.
- Templates: The template is generated from the basic structure of the target i.e. its geometry or the silhouette. It contains the appearance as well as any special information regarding the target, which is the major advantage. The demerit associated with it is that it is not effective in case of pose change during tracking as the template has only one view of the object.

So, we can conclude that there's a strong bonding between target representation and target tracking. Point representation is perfect for the cases where the target is present in a small frame. The geometric shape is suitable for cases where the target is easily fitted in rectangles, circles, or ellipse. When the target is complex in structure for example human body contour or silhouette representation is the best.

2.2 Object detection

It is a process of detecting or locating the target in each frame of any video sequence. It is important to avoid false detection while tracking. There are some techniques for object detection stated below:



(f) (g) (h) (i)

Figure2: Representations of targets (a) single point, (b) multiple points, (c) rectangular shape, (d) elliptical shape, (e) part-based multiple geometric shape, (f) skeleton of object, (g) contour of object, (h) targets contour, (i) silhouette of target.

- Point Detectors: Here, the feature extraction technique provides the interest points over the targets. These points are stored from the frame where the target is located. Later, in all the frames the same points are looked for and the target is detected where the points are found. Literature study shows many such detectors like SURF, SIFT, HARRIS, KLT detectors, etc.
- Background Subtraction: It is a technique that allows foreground extraction which can be used for tracking the object. The word foreground is used for the target in the scene and background for the surroundings of the target. Feature extractor finds the keypoints for foreground and all other keypoints are treated as background. This technique is commonly used for the static background.
- Segmentation: It is a technique that partitions any digital image into a number of segments. The main aim of segmentation is to transform the image into something more purposeful. The common methods of segmentation are thresholding, clustering, motion and interactive segmentation, compression-based, graph-based, histogram-based, etc.

2.3 Object tracking

It is a technique that traces the movement of a single object or multiple objects in a sequence of frames captured by a single or multiple cameras. The main aim of tracking is to associate target in all the video frames. It estimates the trajectory of the moving target in any scene over an image plane. Fig 1.3 shows various tracking methods. Point tracking: This technique is used when targets are represented as points. This technique associates points over different frames. This association is termed correspondence. This in itself is a complicated problem. The complexity increases when real-time challenges like occlusion or pose change are encountered. This approach is further divided into two categories statistical and deterministic. The difference between the two is in their approach how to reduce correspondence cost. Correspondence cost is defined as a process of matching all objects present in frame t-1 to that of the object in frame t.

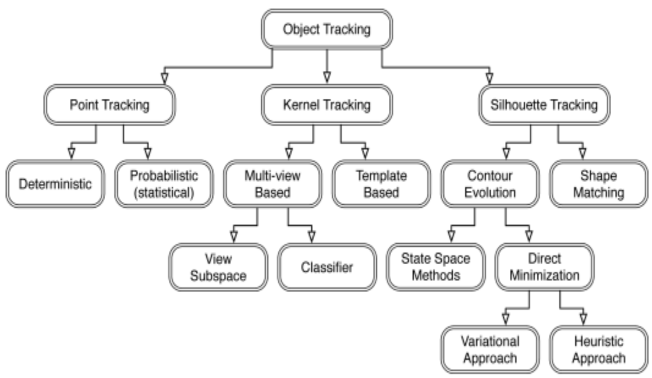


Fig 3: Object tracking methods

- Kernel tracking: This technique involves the usage of the motion model in the tracking process. The motion model helps in computing the motion of the target from one frame to the next and also predicting the position of the target. They are sub-categorized into two parts one is template based and the other is multi-view based. The objects are represented by geometrical shapes
- Silhouette tracking: This is a region-based tracking overcoming the limitations of geometrical shape-based tracking. It provides the exact shape of the target structure. During tracking object models in form of object contour or shape or color histogram is generated, once they are obtained they are matched with regions on current frames that match the model. They are of two types' contour evolution and shape matching.

2.4 Tracking vs. Detection

Object detection is implemented in real-time and easily it can detect the object in every frame. So, the question arises why there's a need for tracking. Here are some reasons why tracking objects is preferred in any video instead of repeated object detections on each frame.

- Tracking is speedy than detection: Normally it is observed that tracking is faster in the process when compared to detection. The reason behind this is that while detection a lot of information is already obtained from the previous frame including the expected location and speed of its motion. These information's are useful in predicting the location of the target incoming frame with the help of very timid information about the surrounding of the previous location. This gives us the exact location of the target. A detection based algorithm has to be started from scratch whereas a tracking algorithm will use previous information it has stored about the object. Detection algorithms have to be trained offline on a number of datasets, which leads to more information about that general class of object. Whereas when we talk about tracking algorithms they know about the class of target they are tracking.
- Tracking comes to the rescue when detection failure is encountered: Whenever the face is occluded during object detection, in that case, the detection fails. Whereas a good tracking algorithm can handle occlusion to some extent.
- Tracking conserves identity: Object detection results in an array of rectangles including the object and it's not clear that what's the match of the rectangle and object whereas in the case of a tracking algorithm can relate to the object it is dealing with.

2.5. Prediction Methods

An important part of a tracking system is the ability to predict where an object will be in the next frame. This is needed to aid the matching of the tracks to detect objects and to predict the position during occlusion. There are four common approaches to predict the objects' positions:

- Block matching
- Kalman filters
- Motion models
- Particle filter

2.6 Terminology

The major terminologies referred to in this thesis are discussed in this section.

Detection: The process of identifying the target during tracking is the first and foremost task. This is usually done by the user through the manual selection and later detected in all other frames using feature extraction techniques. For example, in the first frame, the target is located manually. This becomes the reference and feature keypoints for it are looked for in the rest of the sequence.

- Tracking: It is a process of detecting a moving target in due course of time by a stationary or dynamic camera. This process is done for all the frames of any video sequence. In other words, it is a process of linking trajectories. For example, in this thesis a human is manually selected in a frame later it is tracked in real-time conditions. If tracking is lost due to challenges as discussed before a motion model using estimators and predictors is used to predict the location of the target in consecutive frames.

2.7 Categorization

To provide a clear picture of what actual tracking is about the entire process is divided into four major categories and then reveal the state of this thesis based on this.

Based on the number of targets to be tracked the tracking process is categorized as:

- Single target tracking: As the name suggests the target is one in the real-time scenario. If the target is the only moving object in the background the tracking is comparatively easy. At the same time if the background is dynamic in nature and other than the target there are other moving objects the complexity increases.
- Multiple target tracking: Here during tracking more than one target to chased and tracked. This is a more complex process as compared to single tracking. Also, in such cases initially, the number of targets is undisclosed. This makes it a little more complex problem to tackle because the tracker needs to figure out the exact number of targets and their associated trajectories.

After the target the next important subject is the number of cameras used while tracking:

- Single Camera: such tracking is also defined as monocular tracking [23]. In such cases, there's no surety about calibration variables details given a-priori. If in such case the variables are pre-specified, it is image space-based tracking. So, the trajectories obtained are in 2D and not in 3D.
- Stereo vision or multiple cameras: In this set up multiple cameras are used to give 3D trajectories[24]. It gives a more accurate location of the target when compared to a single camera set up. Also, it is very helpful in occlusion cases because the target is captured by different cameras at different angles. But it makes it a more complex and computationally expensive setup.

The next category that comes into the picture is the movement of the camera whether moving or steady.

- Steady Camera: A steady camera has all the calibration variables constant in the entire course of tracking.
- Dynamic Camera: The cameras are mobile during the video capturing process. So the calibration variables change and have to be re-estimated time and again.

The last and the most important category of tracking is based on detection or motion model evolution.

- By detection: The targets are detected by implementing detectors on each frame. Later the entire tracking process is treated as a data association process. The aim is to generate trajectories with these detections over the period taken. This is a successful technique when multiple targets are taken into account. It is more robust as compared to the other technique stated below .
- By motion model evolution: this technique is more successful for single target based tracking. It is based on Bayes formulation where probability density function is estimated depending on the monitoring. The most common types of approach are Kalmanfilter[5] and particle filter[25].

The focus of the thesis involves single target tracking using a single steady /dynamic camera. The motion model evolution is used based on the interest point technique.

2.8 Performance parameters

- Overlap Percentage: Overlapping area of tracker window on ground truth window [24].

- Overlap % = [area ($W_g \cap W_t$) / area ($W_g \cup W_t$)] X 100
- W_t = Tracker window
- W_g= Ground truth window
- It should be high, above 50% for successful tracking.
- Overlap percentage is a measure of the accuracy of tracking which in turn affects the success rate.
- Success Rate: The ratio of successful tracking frames(n) to the total number of frames(N)[25].
- Success Rate = (n/N) X 100

- It should be high during tracking.
- The success rate is an integral measure of the robustness of the entire tracking process.
- Average Time: Overall average time required for computation while tracking in msec. For efficient tracking, it should be low.

  Less average time signifies the compatibility of real-time tracking.

2.9 Visual tracking methodology

The following flowchart in Fig 2 shows the working of the visual tracking model. Initially, the video is captured using a stationary or moving camera. This can be real-time or pre-recorded video. This video is segmented into a number of frames and from it, its first or any random frame is captured. In this frame, the target is located and bounded in any shape. Once the target is captured the background is subtracted from the foreground image. In the foreground target feature extraction technique is applied like in my research SURF is used. Now tracking process initiates the motion model predicts the location on all the frames. If tracking is lost the motion model updates the new location of the target and tracking is continued.

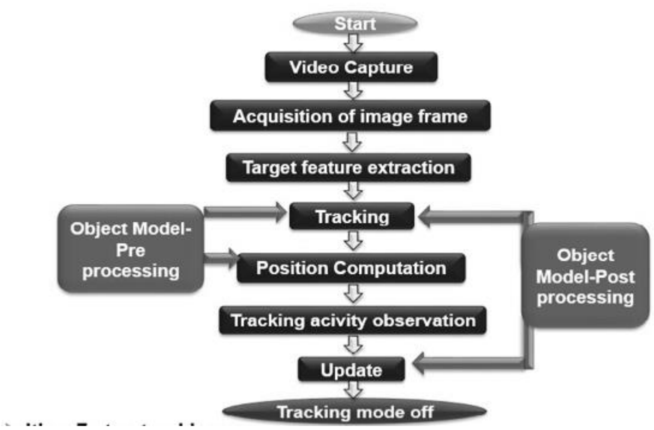


Fig 4 : Flow chart of visual tracking

REFERENCES

1. Yilmaz, A., Javed, O., & Shah, M. (2006). Object tracking: A survey. *Acm computing surveys (CSUR)*, *38*(4), 13-es.
2. Yang, H., Shao, L., Zheng, F., Wang, L., & Song, Z. (2011). Recent advances and trends in visual tracking: A review. *Neurocomputing*, *74*(18), 3823-3831.
3. Van de Weijer, J., Gevers, T., &Bagdanov, A. D. (2005). Boosting color saliency in image feature detection. *IEEE transactions on pattern analysis and machine intelligence*, *28*(1), 150-156.
4. Lukac, R., &Plataniotis, K. N. (Eds.). (2018). *Color image processing: methods and applications*. CRC press.
5. Shotton, J., Winn, J., Rother, C., &Criminisi, A. (2009). Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *International journal of computer vision*, *81*(1), 2-23.
6. Winn, J., Criminisi, A., &Minka, T. (2005, October). Object categorization by learned universal visual dictionary. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1* (Vol. 2, pp. 1800-1807). IEEE.
7. Manjunath, B. S., & Ma, W. Y. (1996). Texture features for browsing and retrieval of image data. *IEEE Transactions on pattern analysis and machine intelligence*, *18*(8), 837-842.
8. Ojala, T., Pietikainen, M., &Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, *24*(7), 971-987.
9. Horn, B. K., &Schunck, B. G. (1981). Determining optical flow. *Artificial intelligence*, *17*(1-3), 185-203.
10. Lucas, B. D., &Kanade, T. (1981, April). An iterative image registration technique with an application to stereo vision.
11. Sabzmeydani, P., & Mori, G. (2007, June). Detecting pedestrians by learning shapelet features. In *2007 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1-8). IEEE.
12. Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, *60*(2), 91-110.
13. Bay, H., Ess, A., Tuytelaars, T., & Van Gool, L. (2008). Speeded-up robust features (SURF). *Computer vision and image understanding*, *110*(3), 346-359.
14. Dalal, N., &Triggs, B. (2005, June). Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)* (Vol. 1, pp. 886-893). Ieee.

15. Feichtenhofer, C., &Pinz, A. (2013). Spatio-temporal good features to track. In *Proceedings of the IEEE International Conference on Computer Vision Workshops* (pp. 246-253).
16. Gerstner, W., &Kistler, W. M. (2002). *Spiking neuron models: Single neurons, populations, plasticity*. Cambridge university press.
17. Rantz, M. J., Banerjee, T. S., Cattoor, E., Scott, S. D., Skubic, M., &Popescu, M. (2013). Automated fall detection with quality improvement "rewind" to reduce falls in hospital rooms. *Journal of gerontological nursing*, *40*(1), 13-17.
18. Nissan Multi-sensing system with front camera, 2015, http://www.nissan-global.com/
19. Liu, J., Carr, P., Collins, R. T., & Liu, Y. (2013). Tracking sports players with context-conditioned motion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1830-1837).
20. Kanda, T., Glas, D. F., Shiomi, M., Ishiguro, H., &Hagita, N. (2008, September). Who will be the customer? A social robot that anticipates people's behavior from their trajectories. In Proceedings of the 10th international conference on Ubiquitous computing (pp. 380-389).
21. Kausler, B. X., Schiegg, M., Andres, B., Lindner, M., Koethe, U., Leitte, H., ...&Hamprecht, F. A. (2012, October). A discrete chain graph model for 3d+ t cell tracking with high misdetection robustness. In *European Conference on Computer Vision* (pp. 144-157). Springer, Berlin, Heidelberg.
22. Pareek, A., &Arora, N. (2019, February). Evaluation of Feature Detector-Descriptor Using RANSAC for Visual Tracking. In *Proceedings of International Conference on Sustainable Computing in Science, Technology and Management (SUSCOM), Amity University Rajasthan, Jaipur-India*.
23. Pareek, A. and Arora, N., 2020. Re-projected SURF features based mean-shift algorithm for visual tracking. *Procedia Computer Science*, *167*, pp.1553-1560.
24. Pareek, A., Arora, V. and Arora, N., 2021. A Robust Surf-Based Online Human Tracking Algorithm Using Adaptive Object Model. In *Proceedings of International Conference on Artificial Intelligence and Applications* (pp. 543-551). Springer, Singapore.
25. Pareek, A., Dahiya, P. and Arora, S.M., SURF-Based Algorithm to Deal with Pose Change Challenge in HumanTracking.