Online Tracking and Offline Recognition Using Scale Invariant Feature Transform

A. Bahmidha Banu¹; Dr. V. Venkatesa kumar²

¹PG Scholar, Department of CSE, Anna University Regional Centre, Tamilnadu, bahmidha@gmail.com; ²Assistant Professor, Department of CSE, Anna University Regional Centre, , Tamilnadu, mail2venkatesa@gmail.com

Abstract—Object tracking, is a challenging problem. In order to meet real-time requirements, Low computational complexity is achieved using a unique feature statistical morphological skeleton, accuracy of localization, and noise robustness has been considered for both object tracking and recognition has been proposed. Previous work used only low level features for tracking framework. Tracking is performed by applying a proposed Scale Invariant Feature Transform to a set of observable quantities derived from the detected skeleton and other geometric characteristics of the moving object. Then unified approach of tracking and recognition can be established. High-level offline models corresponding to the recognized category are then adaptively selected and combined with the proposed online tracking models so as to achieve better tracking performance. Experimental result provides better result when compare with the existing work.

Keywords—Object Recognition; Video Analysis; Visual Tracking;

1. INTRODUCTION

A. Object Tracking

Video tracking is the process of locating a moving object over time using camera. Tracker assigns the video into frames and provides object-centric information, such as area, shape and orientation of an object [2]. Object tracker tracking the object over time by locating its position every frame of the video. There are some tracking methods:

- Point tracking: To detect the object in consecutive frames are represented by points, and points is based on the previous state which can include object position and motion.

- Kernel Tracking: Tracking is based on the shape and its appearance of the object.

- Silhouette Tracking: Silhouette tracking methods used the information encoded inside the object region. Tracked the object by shape matching or contour evolution.

B. Feature Selection for Tracking

Feature is the uniqueness property of the object tracking. So it can be simply distinguished in the feature space. The main role of the tracking process selecting right features from the object. The common features for tracking are as follows:

- Color: The RGB (red, green, blue) color space is usually used to represent color in image processing. In tracking variety of color spaces are used.

- Edges: Boundaries of object generate strong changes in image intensities. This detection is used to identify the changes.

- Texture: Surface intensity variation is measure quantifies properties such as smoothness and regularity.

C. Object Representation

An object can be anything like vehicles in the road, people walking on the road, boats in the sea, bubbles in the water, in the tracking scenario. Objects are represented using object appearances and shapes. The shape of the appearance object for tracking can be representing as below:

- Points: The objects are represented by a point.

- Primitives Geometric Shapes: Object shape is represented by a square, rectangle, ellipse, etc.

- Object silhouette and contour: Boundary of an object is called contour. The region inside the contour is called silhouette of the object.

- Skeletal Models: Applying medical axis transform to the object extracted the object skeleton.

D. Object Detection

Moving objects in video images detection is very important. Moving object in monitoring system needs efficient algorithms for automatic detection of an object [3]. Object detection method mostly applied background subtraction method i.e. to subtract current image from background. The color based subtraction technique method is applied for object detection. There is some object detection techniques, these techniques are given below:
Point detectors: Region of interest in image is considered.

Segmentation: Images are partitioned into similar regions.

Background subtraction: To subtract the current image from the background.

Every moving object should be tracked correctly for subsequent frames. In other cases, the target might change because of object morphing and camouflage [2]. Morphing object pre-determined detectors are likely to fail. So this paper proposes an online tracking and offline recognition. In this approach, object is discovered and tracked and it result are forward to upper-level recognition scheme, in which Scale Invariant Feature Transform (SIFT) algorithm to recognize the category of the object in the current frame.

In this, Section 2 provides the related work of tracking and recognition; Section 3 gives brief description about the proposed model; Section 4 provides the architecture flow of the system; Section 5 describes the implementation of the system; Section 6 concludes with a discussion of this model and future work.

2. RESEARCH WORK

A. Tracking Online Model

Traditional online models for tracking include shape and appearance based templates [12]; Color regions based tracking [23], tracking based stable structures [14], and online classifiers trained by boosting [3]. An online model is based on past tracking performance and tends to drift when the object keeps changing and tracking errors are accumulated. So the errors are prevented from drifting [1], [3]; and [13].

Huang and Nevatia [4] proposed “Tracking by Detection” has been shown to give promising results at real time speeds. Discriminative classifier in an online manner to separate the object from the background. To extract positive and negative examples from the current frame using classifier. It Achieved Better Results with Real-Time Performance. An Object in a video and its location given, but no other information. Kernel-based objective functions optimized using the mean shift algorithm has been demonstrated as an effective means of tracking in video sequences [5]. Tracking methods were able to find a particular individual in image sequences, but were severely challenged by real-world scenarios [8].

Dalal and Triggs [4] the Subspace Adaptation Problem was considered. The best weak classifier is selected by the selector where the error of the weak classifier is estimated from samples seen so far. The complexity was determined by the number of selectors. Real-world tracking a tracker must be able to handle various appearance changes (i.e., illumination changes, occlusions, out-of-plane rotations, movement) which could occur in natural scenes. On-line Ada Boost algorithm which allows updating features of the classifier during tracking.

B. Tracking Offline Model

Offline model used for recognize the tracking object in online model. Pre-trained offline models and databases have been incorporated to the online tracking models for some specific tracking tasks [5], [7].

Avidan [2] proposed a strong classifier using Ada Boost and Mean Shift Algorithm. This algorithm is used added or removed at any time to reflect changes in object appearance or incorporate new information about the background. But the tracker is not designed to handle full and long term occlusions. Tension between adaptation and drifting appears. This algorithm is used to obtain Robust and Stable Tracking [7].

Yang and Kriegman [4] proposed Unified Framework and Appearance model combined the k-means Clustering, Linear Approximation and the Transition Probability methods. In that new framework, both the tracking and recognition components share the same appearance model to minimize the misalignment between the trackers output and the recognizer input [5]. The appearance of each face is modeled by a collection of linear subspaces in the image space [11] [12]. The experimental results had demonstrated that framework is capable of providing robust and stable results for video face recognition. However, illumination variation is an important class of image variation that is not modeled by this algorithm.

Hager and Dewan [7] proposed Kernels based algorithms and more traditional template tracking methods. There is a well known equivalence between the kernel-based objective function and an SSD-like measure on kernel-modulated histograms [10]. It is not yet clear how to properly adapt the histogram structure overtime to adapt to changing illumination, changing target appearance, or occlusion.

Yang and Hua [8] proposed algorithm is CAT steerable filters, often with descriptor permutation. This method suffers from high computational overhead of computing many filter orientations. Rotation as a circular shift and use the magnitude of the Fourier transform. This method is often not sufficiently robust to viewpoint variation. A context-aware tracker tracks the target and the set of auxiliary objects as a random field in a collaborative manner [8] [10].

Andriluka and Roth [9] proposed approximate articulation of each person was detected in every frame based on local features that model the appearance of individual body parts. Prior knowledge in possible articulations and temporal coherency within a walking cycle were modeled using a hierarchical Gaussian process latent variable model. The combination of these results improved hypotheses for position and articulation of each person in several subsequent frames.Babenko and Belongie[10] proposed algorithm is Histograms of Oriented Gradient (HOG) of Histograms of Oriented Gradient (HOG) descriptors significantly outperform existing feature sets for human detection.

The method is based on evaluating well-normalized local histograms of image gradient orientations in a dense grid. This can be done by accumulating a measure of local histogram energy over somewhat larger spatial regions (.blocks.) and using the results to normalize all of the cells in the block. It will refer to the normalized descriptor blocks as Histogram of Oriented Gradient (HOG) descriptors.Bibby and I. Reid [11] proposed to combine the features in the level sets process, enhance set all appearance models by using a probabilistic model determined via Expectation-Maximization (EM) clustering.
The Result set is used as the weighting kernel which improves the accuracy of the similarity measurement in the mean shift method. This begins with finding a person in the video stream. Then, the fore-ground and background models are computed through Expectation-Maximization (EM) clustering.

Takacs and Chandrasekhar [12] proposed unifies tracking and video content recognition with applications to Mobile Augmented Reality (MAR). The Radial Gradient Transform (RGT) yielding the Rotation Invariant Fast Feature (RIFF) descriptor. Any image recognition algorithm for handheld devices must be rotationally invariant. Typical feature descriptor based systems, such as SIFT and SURF, assign an orientation to interest points before extracting descriptors.

Grabner and Bischof [13] proposed Real-time automatic visual tracking, labelling and classification of a variable number of objects such as pedestrians or vehicles, under time varying illumination conditions. Shen [12] proposed inaccurate extraction of the foreground and background information in model adaptation would cause the model to drift and degrade the tracking performance. Drift problem was to obtain accurate boundaries of the target.

3. PROBLEMS IN EXISTING SYSTEM

Existing Online models for low-level correspondences are generally employed to adapt to the changing appearances of the target. The short coming of these online models is that they are constructed and updated based on the previous appearance of the target without much semantic understanding. Therefore, they are limited in predicting unprecedented states of the target due to significant view changes and occlusion, and easily drift in the case when the appearance of the target changes too fast. Some offline trained high level detectors with semantic meanings have already been introduced into the tracking by detection scheme for some specific tracking tasks, especially for human tracking, and vehicle tracking, which largely improves the tracking performance. However, these models assume the semantic meanings of targets are already known before tracking, and accordingly cannot be applied to many general applications.

In existing many object tracking and recognition algorithm has been used. Tracking is essentially a matching problem. All tracking is not the final goal of video analysis but an intermediate task for some succeeding high-level processing like event detection and scene understanding. Essentially, an ideal tracking system should actively understand the target, and adaptively incorporate high-level semantic correspondences and low-level image correspondences. Towards this end, the existing work used a unified approach tracking and Recognition the object. In existing system discovered and tracked the object, the tracking results are continuously fed forward to the upper-level video-based recognition scheme, in which dynamic programming is adopted to recognize the category of the object in the current frame. Based on the feedback from the recognition results, similar off-line models dedicated to specific categories are adaptively chosen, and the location of the tracked object in the next frame is determined by integrated optimization of these selected detectors and the tracking evidence.

Disadvantages of the existing system are the low-level feature information is not enough to estimate the correct scale in tracking scenario. The ambiguity of the tracking problem increases, as the number of object categories increases. The wrong recognition result probably leads to error propagation. It is not appropriate for some tracking dataset, due to data type inconsistency.

4. PROPOSED SYSTEM

A. The Proposed Framework

In proposed system, a unified approach of tracking and recognition can extend by using a unique feature for tracking purpose. Since The low-level feature information is not enough to estimate the correct scale in tracking scenario.

This unified feature of statistical morphological skeleton has been proposed. Low computational complexity is achieves using a unique feature statistical morphological skeleton, accuracy of localization, and noise robustness has been considered for both object tracking and Recognition. Tracking is performed by applying a proposed Scale Invariant Feature Transform to a set of observable quantities derived from the detected skeleton and other geometric characteristics of the moving object. High-level offline models corresponding to the recognized category are then adaptively selected and combined with the proposed tracking models so as to achieve better tracking performance. An overview of the proposed framework is presented in fig.1.

In this section, we propose input video split into frames. Background subtracts the frame and extracts the features. Online tracking and offline recognized the frames, objects are discovered and tracked. The tracking result was fed forward to the object recognition module. When the target was recognized properly, the recognition target model would be automatically incorporated to provide more information about the target to achieve better tracking performance. In the following paragraphs, these components are separately introduced.
B. Video Split into frames

Split the real time video into image blocks. Each splitted image blocks are interpolated to the dimension of the original image without blurring and display image in different screens. The process of dividing the video into non overlapping parts is called video splitting. Frames are without Blurring.

C. Background Subtraction

Background Subtraction comes from the simple technique of subtracting the observed image from the estimated image and thresholding the result to generate the objects of interest. Object detection can be achieved by building a representation of the scene called the background model and then finding deviations from the model for each incoming frame.

A moving objects any significant change in an image region from the background model. The pixels constituting the regions undergoing change are marked for further processing. Background subtraction is the method of removing pixels that do not move, only focus on objects. The method works like this:

- capture two frames
- Each frame compare the pixel colors
- if the colors are the same, replace with the color white
- Else, keep the new pixel

D. Feature Extraction

Feature is extracted using Scale Invariant Feature Transform (SIFT) algorithm. It transforms the image data into scale-invariant relative to local features. It generates large numbers of features that densely cover the image over the full range of scales and oations. For image recognition and matching, SIFT features are first extracted from a set of reference images and stored in a database. A novel image is matched by individually comparing each feature from the new image to this previous database and finding candidate matching features based on Euclidean distance of their feature vectors.

Key point descriptors typically uses a set of Sixteen histograms, it aligned in a 4x4 grid, each with Eight orientation bins, one for each histograms one for each of the main compass directions. It makes the tracking as easy. In this point correspondences the set of salient point on the target at time ‘t’ is denoted as $S_t^t$. For point correspondences we will find two terms. These are, Record the relative position and Correspondences.

- For each $S_t^{i-1}$ on the target, we record its relative position with respect to the target under $t^{i-1}$ normalized by the target size.
- For each $S_t^{i-1}$, we find its $b_t^i$ at time $t$, by SIFT matching, with the matching error $W_t^i$.

Assumptions: The relative position of the same salient point with respect to the target cannot change rapidly.

$$E_t(X_t) = \Sigma ||\text{P}^{i-1} - g^i||^2 e^{-\text{wi}x} \text{………………… (1)}$$

Where,

$X_t$ - Candidate Region.
$g^i$ - relative position of the target.

Condition: If the target movement is only translation (or) scaling, (i.e) $E_s=0$. So $E_s$ is related to the deformation of the target.

E. Tracking and Recognition Method

The Tracking target is $X_t^t$. This is considering as a maximizing likelihood term.

$$X_t^t = \arg \max P (X_t | M_{t-1}^L, M_{t-1}^H, I_t) \text{…………… (2)}$$

Where,

$M_{t-1}^L$ - Tracking model
$M_{t-1}^H$ - Recognition model
$I_t$ - Input image at time ‘t’
$I_t$ - Image sequence: $\{I_1, I_2, \ldots, I_t\}$
$X_t$ - Position of the target

Recognition Term: The current target state is $X_t$. The target category $c_t$ based on the target measurement $Z_t = I_t(X_t)$.

5. SYSTEM IMPLEMENTATION

Offline tracking method some samples of the tracking object given. It is used to achieve more information about the tracking target. The input video is split into frames. The threshold value is (±20%) using trial and error method. The backgrounds are subtracted by comparing the threshold value.

Tracking object features are extracted using SIFT in online tracking method and also offline method. The features are compare between online and offline model. The tracking result was fed forward to the object recognition module.

When the target was recognized properly, the recognition target model would be automatically incorporated to provide more information about the target to achieve better tracking performance.

6. EMPIRICAL RESULTS

The output obtained for this tracking object compared the performance using online and offline model alone, and both combined. The combination of offline and online model performs better than either model. In Fig 2 Evaluate Online, Offline and combined online and offline model.
Fig 2: Quantitative Evolution on Individuals Components of the Unified approach.

7. CONCLUSION AND FUTURE WORK

Visual tracking plays an important role for high-level semantic understanding or video analysis. Meanwhile the high-level understanding that is object recognition should feed back some guidance for low-level tracking. A unified approach is to tracking the object and recognition. Objects are discovered and tracked the tracking result was fed forward to the object recognition module. When the target was recognized properly, the recognition target model would be automatically incorporated to provide more information about the target to achieve better tracking performance. The recognition result was fed back to activate the framework of a system is explained in the fig1 where as the implementation is explained as follows: Input video split into Frames. Background subtract the frame and extract the Features using SIFT algorithm. Online Tracking and offline recognized the frames, objects are discovered and tracked. The algorithm is implemented by MATLAB and runs 2 ~0.5 frames per second on average depending on the object. The off-line model to help and improve tracking. Future work is to implement offline recognition model and online tracking model to improve both tracking and recognition accuracy.

REFERENCES


