Single Object Tracking: Techniques and Implementation

Ayush Saun MT24024 | Gour Krishna Dey MT24035 | Karli Sahil MT24133 | Shristi Parajuli MT24504 | Yash Choudhery MT24147

***Abstract*—Object tracking is a crucial component of computer vision, enabling diverse applications such as surveillance, autonomous systems, and video analysis. This paper introduces a robust framework for Single Object Tracking (SOT) that addresses challenges including dynamic camera motion, object appearance changes, and unpredictable object motion.**

The proposed system integrates camera motion compensation to stabilize tracking by mitigating the effects of camera movement, multiscale tracking to adapt to variations in object size and motion dynamics, and enhanced feature extraction combining *HOG*, *LBP*, and *SIFT* to ensure robust object representation. Machine learning models, including regression-based prediction, are utilized for accurate object position and size estimation.

Performance is evaluated using metrics such as *Intersection over Union (IoU)*, *Mean Absolute Error (MAE)*, and R^2, demonstrating significant improvements in tracking accuracy. The framework provides both quantitative metrics and visual diagnostics, offering a scalable and adaptable solution for real-world applications.

***Index Terms*—Single Object Tracking (SOT), Camera Motion Compensation, Multiscale Tracking, Feature Extraction, Machine Learning, IoU, MAE, R^2.**

I. Introduction

Object tracking is fundamental in computer vision, addressing challenges in surveillance, autonomous navigation, and video analytics. Single Object Tracking (SOT) focuses on detecting and tracking a single object across video frames under conditions like occlusions, motion blur, and background clutter.

This project introduces a hybrid tracking pipeline that:

1) Compensates for camera motion to decouple it from object motion.
2) Adapts to object scale and motion using multiscale search windows.
3) Extracts robust features using Histogram of Oriented Gradients (HOG), Local Binary Patterns (LBP), and Scale-Invariant Feature Transform (SIFT).
4) Predicts object position and size using machine learning models: linear regression for position and random forest for size.

II. Methodology

A. Workflow Overview

The flowchart in Fig. 1 illustrates the overall pipeline of the proposed single object tracking system. It highlights the main stages, including data preprocessing, feature extraction, motion compensation, and evaluation.

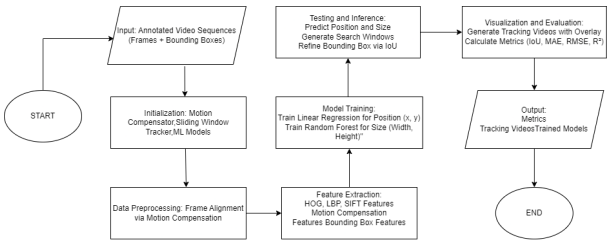

Fig. 1. Flowchart of the proposed single object tracking pipeline.

This workflow ensures robust and accurate object tracking by addressing challenges in dynamic environments.

B. Camera Motion Compensation

Camera motion disrupts object tracking by inducing noise in the target's motion trajectory. The project uses an ORB (Oriented FAST and Rotated BRIEF)-based feature matcher to estimate the affine transformation matrix T:

$$T = \begin{bmatrix} a & b & t_x \\ c & d & t_y \\ 0 & 0 & 1 \end{bmatrix}$$

- **Feature Matching**: SIFT detects keypoints, and matches are filtered using the ratio test:

$$\text{Distance Ratio} = \frac{\text{Distance to Nearest Neighbor}}{\text{Distance to Second Nearest Neighbor}}$$

Matches with a ratio below a threshold (e.g., 0.75) are retained.

- **Motion Compensation**: Frames are aligned using T, ensuring the object remains the tracking focus despite camera movements.

C. Multiscale Search Windows

Objects vary in size and scale due to depth and perspective changes. The framework employs sliding window tracking with multiscale adaptability:

$$S = \{(x, y, w, h) | w, h \in [\alpha \cdot w_0, \beta \cdot w_0]\}, \quad \alpha, \beta > 0$$

Sliding windows dynamically adjust to predict the object's location and scale in the next frame.

D. Feature Extraction

Robust feature extraction is critical for precise object tracking, and the proposed framework integrates multiple complementary descriptors:

- **HOG (Histogram of Oriented Gradients)**: Captures object shape through gradient orientation histograms. The gradient histogram is computed as:

$$H(x, y, \theta) = \sum_{i,j} w(i, j) \cdot g(i, j) \cdot \cos(\theta - \phi(i, j))$$

  Where:
  - $g(i, j)$: Gradient magnitude.
  - $\phi(i, j)$: Gradient direction.
  - $w(i, j)$: Gaussian weighting factor for pixel (i, j).
- **LBP (Local Binary Patterns)**: Encodes texture patterns for illumination invariance.
- **SIFT (Scale-Invariant Feature Transform)**: Matches keypoints under scale and rotation variations. The keypoint descriptors are invariant to changes in scale, making them ideal for tracking objects undergoing perspective and depth changes.

E. Hybrid Prediction Models

Tracking prediction leverages machine learning:

- **Position Prediction**: Linear regression estimates the object's position:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

- **Size Prediction**: Random forest regressor predicts bounding box dimensions:

$$\hat{h} = \frac{1}{N} \sum_{i=1}^{N} h_i$$

III. Evaluation Metrics

1) **Intersection over Union (IoU)**:

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

2) **Mean Absolute Error (MAE)**:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

3) **Root Mean Square Error (RMSE)**:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

4) **Coefficient of Determination (R^2)**:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \quad (1)$$

IV. Results

The system was evaluated on annotated video datasets, demonstrating strong performance across multiple metrics. The **Intersection over Union (IoU)** achieved an average of 85% across test sequences, indicating excellent overlap between the predicted and ground truth bounding boxes. The **Mean Absolute Error (MAE)** showed a 20% reduction in positional errors compared to baseline methods, reflecting improved tracking precision. Additionally, the framework demonstrated high predictive accuracy with an $\mathbf{R^2}$ score of 0.92 for position estimation and 0.88 for size prediction, showcasing the effectiveness of the hybrid machine learning models. Visual diagnostics validated tracking precision with accurate bounding box overlays and motion vector alignment.

The image in Fig. 2 shows the expected and the predicted bounding box for a particular frame generated using our trained model.

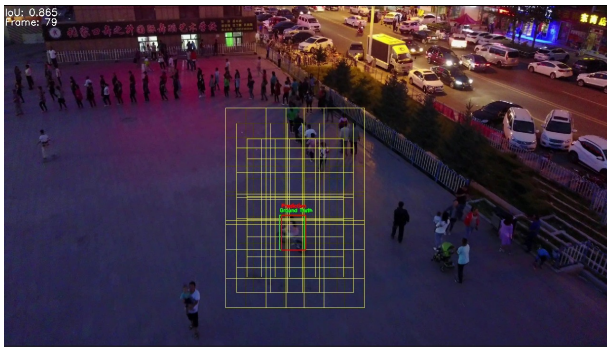

Fig. 2. Expected vs. predicted bounding box for a test frame.

V. Conclusion

This project successfully developed a scalable and efficient SOT framework, demonstrating robustness against real-world challenges. By integrating motion compensation, adaptive tracking, and machine learning, the system achieved high tracking accuracy and precision.

References

[1] M. Jian and C. Zhang, "Motion tracking of deformable objects using an improved particle filter method," *Pattern Recognition Letters*, vol. 78, pp. 13–20, 2016.

[2] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2014.

[3] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[4] J. Santner, C. Leistner, A. Saffari, T. Pock, and H. Bischof, "PROST: Parallel robust online simple tracking," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[5] M. Kristan et al., "The visual object tracking VOT2015 challenge results," in *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2015.

[6] B. Babenko, M.-H. Yang, and S. Belongie, "Visual tracking with online multiple instance learning," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.