A SIFT-based Mean Shift Algorithm for Moving Vehicle Tracking*

Liang Wei¹ Xie,Xudong² Wang,Jianhua³ Zhang Yi⁴ Hu,jianming⁵

Abstract—The classical mean shift algorithm is easy to pass into local maxima, which is caused by the lack of appropriate target model updating mechanism. In this paper, a SIFT-based mean shift algorithm is proposed, which can be used for continuous vehicle tracking in complex situations, such as the shape and the illumination of the vehicle object change. In our algorithm, the mean shift algorithm is utilized to determine the candidate target region, and then a judgment on the tracking effect is made according to the Bhattacharyya coefficient. If tracking fails, the candidate area is matched with the target model by SIFT feature, and a new track position is determined. Otherwise, the target model is periodically updated by SIFT feature matching, and the target model can be constantly updated according to the state change of the moving vehicle. In the scenes of moving vehicle target deformations, such as the variation of scale and illumination, the algorithm is tested and compared with other algorithms. The experimental results show that the proposed method can effectively track an object under the condition of varying illumination and shape deformation.

I. INTRODUCTION

Moving vehicle tracking is an important research direction in the field of intelligent transportation systems, which can be used for video surveillance, artificial intelligence, object recognition, and military defense, etc.

Mean shift algorithm is a non-parametric matching algorithm based on density estimation. Because of its efficiency in feature extraction and matching, it has been widely used in moving object tracking in recent years[1,2]. However, when the deformation of a target is caused by changes due to external illumination, shadowing and turning movements of vehicles, mean shift algorithm fails easily to track because the color histogram is employed to describe the target model. Chen, Ai-hua et al.[3] and Gao, Tao et al.[4] combined the vehicle position prediction and particle filter with mean shift separately to track moving vehicle. Han, Pengchong et al.[5] combined the wavelet analysis with mean shift to track moving satellites, vehicles, pedestrians and other targets. The mean shift algorithm is to maximize the correlation between two color histograms between the object model and the candidate in different images so that it is one of the most efficient tracking algorithms for real-time applications. However, because the color histogram is easily affected by varying illumination or shape, it is prone to cause wrong tracking results.

The Scale Invariant Feature Transform (SIFT) algorithm[6] extracts local features and is employed to track objects. Because SIFT local features have a good invariance of the image rotation, scaling transformation, affine transformation, and illumination changing, SIFT features matching has a great advantage relatively to traditional gray correlation method. Lowe[6] described the implementation process of SIFT algorithm in detail and applied it to image retrieval. Psyllos[7] proposed a vehicle logo recognition algorithm, which is based on an enhanced SIFT feature-matching scheme. Zhao, Xuqiang et al.[8] achieved better results by using the SIFT matching algorithm for object tracking. Qian, Zhiming[9] combined the SIFT matching algorithm and SVM, which can perform target recognition and tracking in multi-vehicle scenes. The proposed method is able to handle tracking problems well in complex situations, such as occlusions, rotations, and illumination changing. However, there is a time-consuming problem for features matching, which severely restricts its applications in practice. The matching of SIFT feature points fails when the object has the large variation of scale and rotation and the CAMSHIFT algorithm fails when the object moves too fast between two sequent frames. Wang Z et al.[10] combined the advantages of the two algorithms to generate a new one. They got two candidate models by SIFT and CAMSHIFT, computed histogram of color probability, compared their Bhattacharyya coefficients, and selected the algorithm whose coefficient value is smaller to track object adaptably. Although, the object model is not renewed essentially so that the SIFT feature matching points become too few if the shape of the vehicle object varies so large. The effect of the new algorithm is the same as that of the CAMSHIFT algorithm. Chen, Ai-hua et al.[3] also proposed a tracking algorithm combining mean shift and SIFT to solve the problem of the variation of the object’s scale. It uses mean shift algorithm to get approximate candidate regions of the object, uses SIFT to abstract the features, and computes transformation coefficients of two images by matching points, through which whether to renew the object model can be decided. But,
when the shape of the object varies large, the number of
feature matching points is not enough so that the object
model cannot be renewed by the transformation coefficients
and the track fails.

In this paper, a SIFT-based mean shift algorithm is
proposed for moving vehicle tracking. In our method, the
candidate area of moving object is selected by mean shift,
and then the effect of tracking is estimated according to
the Bhattacharyya coefficient. If the coefficient is lower
than a preset threshold, which indicates that the tracking
by the mean shift fails, the target model will be used to
match with the candidate areas by SIFT features. After the
target model is renewed based on the deformation coefficient
matrix, the new track position can be obtained. Otherwise,
if the Bhattacharyya coefficient is larger than the thresh-
old, which means the tracking successes, the target model
will be periodically updated using SIFT matching for the
vehicle movements. In summary, our method combines the
advantages of SIFT and mean shift algorithms in case of
vehicle target deformation and illumination changing.
The proposed algorithm adopts the update mechanism of moving
target model based on SIFT features, and provides a detailed
flowchart of moving vehicle tracking algorithm combining
SIFT and mean shift. The proposed algorithm is tested based
on a real traffic video and compared with the presented
tracking algorithms, i.e. the BWH(background-weighted his-
togram) mean shift algorithm[1] and the CAMSHIFTand-
SIFT algorithm[10]. The experimental results show that our
proposed method can track a moving vehicle changing in
size and pose more accurately.

The rest of paper is organized as follows. The proposed
approach is introduced in Section II with classical mean shift
tracking algorithm and SIFT theory. The experimental results
and analysis are presented in Section III, and Section IV
shows the conclusions and future work.

II. PROPOSED ALGORITHM

A. Mean shift Algorithm

Mean shift algorithm is proposed by Fukunaga et al.[11]
in the estimation of probability density function. Now mean
shift algorithm[1,2] is an iterative process, in which the RGB
color histogram of the original target in the first frame is
iteratively compared with that of the target candidate regions
in the following frames. The algorithm does not search in the
whole parameter space and has real-time performance. The
process of object tracking is shown as below.

1) In the initial frame, the tracking object is manually
selected and the object model is built with its probability
distribution of color histogram. Suppose $y_0$ is the center
of an object, the positions of pixels are \( \{ x_i \}_{i=1}^{n_h} \), where \( n \) is the number of pixels of the object. The grayscale range of the
target area is \([0, H - 1] \). The statistical histogram distribution
model of the target area is

\[
\hat{p}_h(y_0) = C_h \sum_{i=1}^{n_h} k(||y_0 - x_i||^2) \delta[b(x_i) - h]
\]

where \( \delta \) is an impulse function, \( C \) is the normalization
factor, \( b(x_i) \) is the grayscale value of the pixel at \( x_i \),
\( h \in [0, H - 1] \). \( ||x_i|| \) is the distance between \( x_i \) and the

(2) At \( y_0 \) in the current frame, its pixels are \( \{ x_i \}_{i=1}^{n_h} \)
and the statistical histogram distribution model is

\[
\hat{p}_h(y_0) = C_h \sum_{i=1}^{n_h} k(||y_0 - x_i||^2) \delta[b(x_i) - h]
\]

(3) Computing the similarity measurement between the
object model and the candidate template by the Bhattacharyya
coefficient.

\[
\rho(\hat{p}_h(y_0), \hat{q}) = \sum_{h=0}^{H-1} \sqrt{\hat{p}_h(y_0) \hat{q}_h}
\]

(4) Computing the weights of pixels in the tracking
window

\[
w_i = \sum_{h=0}^{H-1} \delta[b(x_i) - h] \sqrt{\frac{q_h}{p_h(y_0)}}
\]

(5) Searching a new object position by the mean shift
value.

\[
y_1 = \sum_{i=1}^{n_h} w_i g(||y_0 - x_i||^2) \sum_{i=1}^{n_h} w_i g(||y_0 - x_i||^2)
\]

And computing the Bhattacharyya coefficient.

\[
\rho(\hat{p}_h(y_1), \hat{q}) = \sum_{h=0}^{H-1} \sqrt{\hat{p}_h(y_1) \hat{q}_h}
\]

(6) Comparing the Bhattacharyya coefficients, and updat-
ing the candidate template.

If \( \rho(\hat{p}_h(y_1), \hat{q}) < \rho(\hat{p}_h(y_0), \hat{q}) \)
then \( y_1 = \frac{1}{2}(y_0 + y_1) \)

763

(7) If \( ||y_1 - y_0|| < \xi \), the iteration stops.
otherwise, \( y_0 = y_1 \) and going to step(2).

B. SIFT Theory

The object tracking based on SIFT matching has two steps:
SIFT local features extracting and local features matching
between two images [6], where the 5 steps are shown in
detail.

(1) Scale-space peak selection: the keypoints, which is
invariant of scale and rotation, are searched in all the scale-
spaces of the original image and found through the DoG
functions.

The scale-space \( L(x, y, \sigma) \) of the image is generated to
build up its DoG pyramid sequence, i.e.

\[
L(x, y, \sigma) = G(x, y, \sigma) * I(x, y)
\]

In (4), * is the convolution operation and \( G(x, y, \sigma) \) is a
Gaussian function.

The difference of the Gaussian function \( D(x, y, \sigma) \) is used
to detect the stable feature point locations in its scale-space.
\( D(x, y, \sigma) \) is computed by using the difference between two
adjacent images.

\[
D(x, y, \sigma) = [G(x, y, k\sigma) - G(x, y, \sigma)] * I(x, y)
\]

\[
= L(x, y, k\sigma) - L(x, y, \sigma)
\]

\[
= (k-1)\sigma^2\nabla G(x, y, \sigma)
\]
And then the maxima and minima of the scale-space function are determined by comparing each pixel in the pyramid to its neighbors.

(2) Keypoint localization: scale-space extrema detection produces too many candidate keypoints, some of which are unstable. This step is to perform a detailed fit to the nearby data for accurate location, scale, and ratio of principal curvatures.

Because the DoG function has strong responses along edges, even if the candidate keypoints are not robust to small amounts of noise. Therefore, in order to increase the stability, low contrast candidate points and edge response points along an edge are discarded through the Laplace operation of keypoints.

(3) Orientation assignment: each keypoint is assigned one or more orientations based on the local image gradient directions. This is the key step in achieving the invariance to rotation as the keypoint descriptor can be represented relative to the orientation and therefore the invariance to image rotation can be achieved.

Each key location is assigned a canonical orientation so that the image descriptors are invariant to rotation. In order to make it as stable as possible against lighting or contrast changes, the orientation is determined by the peak in a histogram of local image gradient orientations.

(4) Feature points descriptor: a descriptor vector is computed for each keypoint such that the descriptor is highly distinctive and partially invariant to the remaining variations such as illumination, 3D viewpoint, etc. This step is performed on the image closest in scale to the keypoint’s scale. Given a stable location, scale, and orientation for each key, it is now possible to describe the local image region in a manner invariant to these transformations. A set of orientation histograms is created on $4 \times 4$ neighborhoods with 8 bins each. These histograms are computed from the magnitude and orientation values of samples in a $16 \times 16$ region around the keypoint, so that each histogram contains samples from a $4 \times 4$ subregion of the original neighborhood region. The vector of feature points descriptor has 128 elements.

(5) Similarity measure determination: the Euclid distance between the keypoints of two images is used as similarity measurement. A keypoint of image A is taken, two nearest keypoints from which are found in image B. If the time by nearest keypoint and the second one is less than the certain threshold, a pair of matching points can be obtained.

C. Proposed Algorithm

In order to alleviate the effects of scale change and illumination of vehicle object on tracking performance, the main ideas of the proposed algorithm are mean shift tracking and SIFT matching adjusting. When mean shift tracking fails, SIFT algorithm is used to match the object, which can adjust the tracking window and update the object model in time. The detailed steps of our algorithm is shown as follows.

Firstly, the moving vehicle object concerned is manually selected in the first frame to create a rectangle template. Its probability density function of color histogram is calculated and the corresponding SIFT features is extracted. Then,

1) Track the vehicle object by the mean shift algorithm in the next frame.

2) Judge the tracking performance according to the Bhattacharyya coefficient. If it is less than the preset threshold, the tracking deviates. The tracking window of mean shift is enlarged to include the vehicle object as a candidate template and the SIFT features are extracted.

3) Match the SIFT features between the object model and the candidate, and then compute the deformation coefficient matrix $H$ by the matched feature pairs, which can be used to determine the new tracking window. Then, update the tracking position and the object model.

4) If tracking successes, update the object model periodically by SIFT matching and the deformation coefficient matrix $H$.

5) Repeat the above steps and track continually.

The flowchart of proposed algorithm is described as Fig.1. In the algorithm, the determination of the deformation matrix $H$ of the vehicle object in step (3) and step (4) is the key point.

The deformation matrix $H$ of a vehicle object consists of rotation, scale, and translation. Supposed $I$ and $I'$ are two images including the vehicle object concerned, the deformation relation of them is described by (6):

$$I' = H \times I$$

where $H$ is a $3 \times 3$ nonsingular matrix which is described as below,

$$H = \begin{bmatrix}
\rho \cos \theta & -\rho \sin \theta & t_x \\
\rho \sin \theta & \rho \cos \theta & t_y \\
0 & 0 & 1
\end{bmatrix}$$
where \( \rho \) is a scale factor, \( \theta \) is a rotation angle, \((t_x, t_y)\) is the shift vector. According to the matrix \( H \), the relation of a vehicle object model and the candidate template is expressed distinctly. In order to compute 4 parameters in \( H \), 4 equations linearly independent are required. These required equations can be achieved from the deformation relation of the correct matching features by SIFT. An example of the result of SIFT features matching is showed in Fig.2, which shows the state change of vehicle in addition.

Supposed \( P \) is the positions of the correctly matched features in the vehicle object model image, and \( P' \) is the corresponding positions of the correctly matched features in the vehicle candidate image, i.e.

\[
P = \{(x_i, y_i, 1), i = 1...N\}
\]
\[
P' = \{(x_i', y_i', 1), i = 1...N\}
\]

where \( N \) is the number of the correctly matched features. For each correctly pixel pair, we have:

\[
\begin{align*}
x_i'' &= x_i \rho \cos \hat{\theta} - y_i \rho \sin \hat{\theta} + t_x \\tag{7} \\
y_i'' &= x_i \rho \sin \hat{\theta} + y_i \rho \cos \hat{\theta} + t_y
\end{align*}
\]

Then, we can see that there are \( 2 \times N \) equations which can be obtained from the \( N \) matched couples, and these equations can be used to compute the transform parameters. This is usually an overdetermined problem. In other word, the number of equations is more than the number of unknown parameters. To compute the unknown parameters, an optimization algorithm in MATLAB is adopted to minimize the least mean squares as below:

\[
\begin{align*}
\min f &= \min \sum_{i=1}^{N} [(x_i' - x_i'')^2 + (y_i' - y_i'')^2] \\
x_i'' &= x_i \hat{\rho} \cos \hat{\theta} - y_i \hat{\rho} \sin \hat{\theta} + \hat{t}_x \\
y_i'' &= x_i \hat{\rho} \sin \hat{\theta} + y_i \hat{\rho} \cos \hat{\theta} + \hat{t}_y \\tag{8}
\end{align*}
\]

where \( \hat{\rho}, \hat{\theta}, \hat{t}_x, \hat{t}_y \) are the estimated parameters in \( H \).

After the deformation coefficient matrix \( H \) is solved, the new tracking window and object model can be updated by (7).

### III. EXPERIMENTAL ANALYSIS

In order to estimate the performance of our proposed algorithm, it is tested using MATLAB programs on an INTEL® Core™ i3-2120, 3.30GHz PC machine. The video named as "AVSS_PV_Easy_Divx.avi" [12] is selected, which has twenty-five frames per second. The video includes the scenes of vehicle object turning, illumination changing, and vehicle object moving from far to near. The image size is \( 576 \times 720 \) pixels and the initial object template size is \( 68 \times 90 \) pixels. The color histogram is computed in RGB space, and each channel is divided into 16 bins. The threshold of the Bhattacharyya coefficient is 0.90, and the period of updating the object model is 10 frames. In order to show the performances of our method, we compare it with the BWH mean shift algorithm[1] and the CAMSHIFTandSIFT algorithm[10].

Fig.3 shows the tracking results of the BWH mean shift algorithm and the CAMSHIFTandSIFT algorithm, from which we can see that the two methods have similar performance. Fig.4 shows the tracking results of our algorithm. Fig. 5 and fig.6 compare the Bhattacharyya coefficients and tracking errors of the three algorithms, respectively.

From Fig.3, we can see that the tracking window deviates in the frame 55(fig.3b), frame 75(fig.3c) and frame 85(fig.3d). This is because the color histogram of the candidate template is great changed when the moving vehicle is turning left and the illumination is affected by the shadow of the building. In fact, although the BWH mean shift algorithm adjusts the tracking window by 10\%, it cannot tackle the problem when the tracking vehicle changes its shape greatly. The CAMSHIFTandSIFT algorithm utilizes the SIFT features to match vehicle object. When the shape of the object changes, the number of matched points greatly decreases and the performance of the method will degrade as the SIFT matching does not contribute to the method actually. From Fig.4, we can see that the moving vehicle can be tracked more accurately by our algorithm, as shown in the frame 55(fig.4b), frame 75(fig.4c) and frame 85(fig.4d).

From Fig.5, we can see that the proposed algorithm updates the object model with SIFT matching per 10 frames or when the Bhattacharyya coefficient is less than 0.90. Therefore, in the whole process of vehicle tracking, the Bhattacharyya coefficient is always larger than 0.90, which indicates a good performance for tracking.

Further, from Fig.6 the tracking error of the proposed algorithm is obviously less than that of the other two algorithms. The ground truth is created as the rectangle which is manually set by a trained professional. The tracking error is defined as the Euclid distance of the center of rectangle window. The mean error of the proposed algorithm is 5.15 pixels, while the mean error that of BWH mean shift algorithm is 37.78 pixels and that of the CAMSHIFTandSIFT algorithm is 38.89 pixels. The accuracy of tracking algorithm will be beneficial to the subsequent processing of the tracking target, such as target segmentation and extraction. The results show clearly the better performance of the proposed method.

### IV. CONCLUSIONS

The tracking problem of moving vehicle is important in ITS and computer vision. In fact, the moving vehicle tracking...
at outdoor has many challenges, such as the changing of illumination, scale, occlusion, and rotation. All the factors result in complex surroundings for object tracking. To some extent, SIFT features are invariant of the above factors. In this paper, a new mean shift tracking algorithm based on SIFT matching is proposed, which combines SIFT and mean shift algorithm. Our method is tested based on actual traffic video and the experimental results show that our method can achieve a good tracking accuracy. Similarly, the proposed method can also be utilized in other traffic scenes, such as pedestrian tracking and multi-view vehicle tracking.

In our future work, we will research the tracking method which is used for moving vehicles with occlusions. In addition, besides the SIFT features, some other features can also be considered to express the vehicle object more accurately and to improve the tracking performance.

REFERENCES


