View independent object classification by exploring scene consistency information for traffic scene surveillance

Zhaoxiang Zhang, Kaiqi Huang, Yunhong Wang, Min Li

Laboratory of Intelligent Recognition and Image Processing, Beijing Key Laboratory of Digital Media, School of Computer Science and Engineering, Beihang University, Beijing 100191, China
National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

Abstract

We address the problem of view independent object classification. Our aim is to classify moving objects in traffic scene surveillance videos into pedestrians, bicycles and vehicles. However, this problem is very challenging due to the following aspects. Firstly, regions of interest in videos are of low resolution and limited size due to the capacity of conventional surveillance cameras. Secondly, the intra-class variations are very large due to changes in view angles, lighting conditions and environments. Thirdly, real-time performance of algorithms is always required for real applications. Especially, perspective distortions of surveillance cameras make most 2D object features like size and speed related to view angles and not suitable for object classification. In this paper, we try to explore the hidden information of traffic scenes to deal with perspective distortions of surveillance cameras. Two solutions are given to achieve automatic object classification based on simple motion and shape features on the 2D image plane, both of which are free of large database collection and manually labeling. Abundant experiments of the two methods are conducted in videos of different scenes and experimental results demonstrate the performance of our approaches.

1. Introduction

Automatic object classification in videos is an important issue in image processing and video analysis with great potential for all kinds of real applications, which supplies a bridge between the low level feature extraction and the high level video interpretation. With object classification achieved, system operators can program the monitoring system by specifying events of interest for different object types. Object classification can also be treated as a preprocessing step for object recognition to supply coarse type information and improve efficiency by decreasing searching scopes. In addition, object classification is also indispensable for the semantic interpretation, which is the ultimate aim of computer vision.

In spite of its importance, object classification in wide view scene surveillance is also very challenging for the following aspects. Firstly, regions of interest in videos are of limited size and low resolution due to the capacity of conventional surveillance cameras. Secondly, the intra-class variance for every category is very large. Object appearance is not only determined by categories, but also related to view angles, illuminations and even seasons. Thirdly, since object classification has great potential for real applications, real-time performance is always required to be achieved which limits the complexity of object classification algorithms.

Since its importance and difficulties, various works have been done in the field of object classification. With regions of interest detected by motion information, many features and classifiers are adopted to construct classification systems. In the early works like [1–4], shape features like size, compactness, bounding box aspect ratio and motion features like speed, motion direction are extracted for training and classification. However, these shape and motion features are based on the 2D image plane so that they cannot avoid perspective distortions, which are much more significant in wide-view traffic scene videos. For example, nearby objects in images appear to be larger and move faster than those far away.

Evidently, simply extracting features from 2D regions of interest in images is not enough to conquer the effect of perspective distortions. Extra information should be added and interact with image features to enhance the performance of object classification. In the past, various strategies have been attempted with different kinds of information added. Learning based methods make use of information of a large labeled database, which supplies redundant sample information to improve the accuracy and robustness of object classification. Instead of simple shape and motion features on the
In this work, we propose to use an online learning method for view-independent object classification. The main idea is to use scene consistency information, which is combined with the common scene. In this paper, we propose an algorithm to achieve high performance object classification by exploring scene consistency information. Different models are attempted and an unsupervised strategy is applied to avoid large database collection and manual labeling. An online learning based framework is adopted to model the algorithm adaptive to scene and condition changes. Extensive experiments are conducted and experimental results demonstrate the performance of our proposed approaches.

The remainder of the paper is organized as follows. In Section 2, we will introduce a scene division based strategy as a rough description of scene consistency information, which is combined with an online learning based framework for view independent object classification. Then, a much refined model based on the ground plane rectification is proposed to enhance the description of scene consistency information in Section 3. Experimental results and analysis are described in Section 4. Finally, we draw our conclusions in Section 5.

2. Scene division based object classification

As we have described before, object appearance in videos suffers from perspective distortions of surveillance cameras, which is much more significant in wide-view scene video surveillance. Furthermore, object metrics on the 2D image plane have nonlinear relations to view angles, which makes it even more difficult to achieve view independent object classification. We try to make use of hidden scene information to solve this problem. It is based on an evident constraint that objects exist in the same surveillance scene, which should suffer from the consistent perspective distortion. In other words, closer objects in video scenes should suffer more similar perspective distortion than those who are far from each other. Even though the relation between the distortions and object position is nonlinear, we can adopt a divide and conquer strategy to model the scene consistency model roughly by scene division. Then an online learning based strategy can be applied to achieve view independent object classification, which is free of large database collection and manual labeling. The flowchart of the scene division based object classification is shown in Fig. 1. A previous and much shorter version of this work has appeared in [30].

![Flowchart of scene division based object classification](image-url)

**Fig. 1.** Flowchart of scene division based object classification [30].
2.1. Scene division

2D features like speed, size and velocity are quite effective in narrow view scene surveillance videos, which have achieved successful applications for object classification as illustrated in [1–4]. The reason is that objects of interest are close to each other in narrow view scene surveillance videos. Based on the scene consistency constraint, they suffer from the similar perspective distortions, which make the 2D features distorted accordingly to be still effective for object classification. In the case of wide view surveillance, objects are not always close to each other and the nonlinear effect of perspective distortions becomes evident to make 2D image features no longer effective for object classification. Based on this consideration, we can adopt a divide and conquer strategy to divide the wide view surveillance scene into many equal subregions. If the division level is appropriate, each subregion can be independently processed as a narrow view scene, in which 2D image features can be effective to achieve object classification. We can construct classifiers for each of these subregions and enhance the object classification performance by the decision level fusion of classifiers in different subregions. Evidently the more subregions the scene is divided into the more effective the 2D features will be. However, since small regions need longer time to collect enough samples for online learning and cause maintenance of larger number of classifiers, there is a trade-off between the accuracy and the computational cost. In our approach, the average size of moving objects is estimated and compared with the size of the scene, which is adopted as a criteria for the division level. Based on our experience, we should make sure that the ratio between the average size of objects in each subregion and the size of the whole subregion larger than 0.03. Taking the scene of Video 1 shown in Fig. 2(a) as an example, we divide the scene into 16 parts as 4 × 4 grids, which correspond to 16 classifiers initialized and maintained online later. For the case of Video 2 illustrated in Fig. 2(b) as a much narrower view scene, we only need to divide the scene into 2 × 2 grids to achieve object classification.

2.2. K-Mean clustering for automatic labeling

In order to classify moving objects automatically without supervised learning, we adopt K-Mean clustering and decision level fusion for automatic labeling. There are totally five shape and motion features used in our algorithm:

- **size**: the size of objects in pixels,
- **speed**: time derivative of centroid of the object,
- **compactness**: equals to size/perimeter²,
- **size’**: time derivative of size,
- **angle**: angle between motion direction and direction of major axis of the silhouette.

As we know, **size** and **speed** have the most significant perspective distortions among all the 5 features. As a result, we use the additional 3 features for K-Mean clustering and automatic labeling. After videos processed frame by frame for a period of time, K-Mean clustering is adopted to establish 3 clusters. One cluster corresponds to one category, respectively. The decision level fusion based on the following three intuitive rules is adopted to establish the correspondence [30]:

1. **compactness** has the advantages of distinguishing vehicles from pedestrians and bicycles;
2. **size’** has the advantages of distinguishing pedestrians from vehicles and bicycles; and
3. **angle** has the advantages of classifying pedestrians and vehicles. Using voting strategy, we can conveniently achieve automatic labeling.

2.3. Bayesian classification

After dividing wide view traffic scenes into equal grids to obtain many narrow view scenes, 2D motion and shape features can be effectively applied to construct classifiers for each of the subregion. We use \( \nu = (\text{size, speed, compactness}) \) as distinctive feature vector to achieve object classification in these subregions. The approach is based on the Gaussian assumption that the feature vector \( \nu \) of every category in each grid satisfies a multivariate Gaussian distribution, which is tested helpful with appropriate subregion size and denoted as

\[
P_i(\nu) = \mathcal{G}(\nu, \mu_i, \Sigma_i), \quad i = 1, 2, 3
\]  

(1)

Using Bayesian rules, we obtain the derivation as follows:

\[
P(\text{category} = i|\nu) \propto P_i(\nu) \cdot p_i, \quad i = 1, 2, 3
\]  

(2)

where \( P(\text{category} = i|\nu) \) and \( p_i \) are the posterior and prior probability of each category, respectively.

Classification is realized in each subregion after automatic labeling. If the number of moving objects passing the subregion has not reached a threshold \( N \), classification is simply realized by comparing the distance between the feature vector and every cluster. If it reaches \( N \), the classifier of this subregion is initialized. The prior probability is determined by the number of individuals belonging to each cluster and the Gaussian distribution is estimated in the following way:

\[
\mu_i = \frac{1}{N} \sum_{r=1}^{N} \nu_{ir}, \quad i = 1, 2, 3
\]

\[
\sigma_i^2 = \frac{1}{N} \sum_{r=1}^{N} (\nu_{ir} - \mu_i)(\nu_{ir} - \mu_i), \quad i, j = 1, 2, 3
\]

(3)

\[\mu_i = \frac{1}{N} \sum_{r=1}^{N} \nu_{ir}, \quad i = 1, 2, 3\]

\[\sigma_i^2 = \frac{1}{N} \sum_{r=1}^{N} (\nu_{ir} - \mu_i)(\nu_{ir} - \mu_i), \quad i, j = 1, 2, 3\]

\[\frac{1}{N} \sum_{r=1}^{N} \nu_{ir}, \quad i = 1, 2, 3\]

\[\frac{1}{N} \sum_{r=1}^{N} (\nu_{ir} - \mu_i)(\nu_{ir} - \mu_i), \quad i, j = 1, 2, 3\]

Fig. 2. Illustration of scene division of different levels: (a) scene of Video 1 and (b) scene of Video 2.
The category is determined by the posterior probability and the classifier is refined at the same time to be robust to condition changes

\[
\begin{align*}
\hat{p}_{k,\text{new}} &= (1-\eta)\hat{p}_{k,\text{old}} + \eta(M(k,t)), \\
\hat{\mu}_{k,\text{new}}^2 &= (1-\gamma)\hat{\mu}_{k,\text{old}}^2 + \gamma\mu^2_t, \\
\hat{\sigma}_{l,j,\text{new}}^2 &= (1-\gamma)\hat{\sigma}_{l,j,\text{old}}^2 + \gamma(\mu^2_t - \hat{\mu}_l^2)(\mu^2_t - \hat{\mu}_j^2)
\end{align*}
\]

(4)

where \(\eta\) and \(\gamma\) are the refinement rate. \(M(k,t)\) is 1 if \(t\) is classified as the category \(k\) and 0 otherwise. The prior probability \(\hat{p}_{k,\text{old}}\) is renormalized accordingly.

In every frame, there is posterior probability output for every moving objects. We can further make use of object tracking to enhance the performance object classification by the decision level fusion. In our work, decision level fusion is simply achieved by determining object category using the sum of posterior probability of tracked frames.

### 2.4. Scene change detection

Most of scene changes in video surveillance are abrupt transitions caused by zooming or moving of cameras rather than gradual transitions. We can simply detect scene changes when

\[\Sigma_{xy}(B(x,y) - B_{l-1}(x,y)) > T\]

(5)

where \(T\) is a threshold and \(B_{l}\) and \(B_{l-1}\) are recovered background of the current and previous frames, respectively. As we use the reflectance component for background modeling described in [31], the detection is robust to fast illumination changes. When scene changes are detected, subregions will be subdivided according to the new situation and the classifiers will be refined online to be adaptive to the new environment.

### 2.5. Discussion

Scene division based strategy can be seen as a rough modeling of scene consistency information to conquer perspective distortions of surveillance cameras. Its effectiveness to achieve view independent object classification will be demonstrated in Section 4. However, just as we have described, the scene division based strategy models the scene consistency information very roughly. It just achieves a transformation from the wide view scene to narrow view scenes. The performance of object classification is closely related to the division level. If the subregion is not narrow enough, the 2D features are not effective to achieve object classification and the Gaussian assumption cannot describe the distribution of features accurately. On the other hand, if the scene is over divided, it will cost a much longer time for each subregion to initialize classifiers and a larger computational cost to maintain more independent classifiers.

### 3. Ground plane rectification based scene consistency modeling

As we have described above, scene division based strategy can only model the scene consistency information roughly. More information of surveillance scenes, especially the hidden scene constraints, should be further explored to enhance the accuracy of scene consistency modeling.

It is a common fact that objects of interest in traffic scene surveillance videos are always moving on the ground plane, which can be seen as the Ground Plane Constraints (GPC) [16]. In this case, we need not achieve complete camera calibration to conquer perspective distortions, but only focus our interest on the ground plane rectification. Ground plane rectification can be seen as a more accurate model to describe scene consistency information and straightly adopted in the online learning based framework, which will be described in detail as follows [32].

Under perspective projection, the ground plane is mapped to the image plane by homography. Points on the image plane, \(x\), are related to points on the world plane, \(x'\), as \(x' = \mathbf{H}x\). As has been described in [33], the homography matrix \(\mathbf{H}\) can be decomposed uniquely into three matrices, \(S, A, \mathbf{P}\), representing the similarity, affine and pure-projective components of homography, respectively

\[\mathbf{H} = \mathbf{SAP}\]

(6)

The similarity component \(\mathbf{S}\) is a similarity transformation which can be further decomposed into rotation, translation and zooming components. Similarity transform does not generate perspective distortions of objects, which can be ignored to be processed in our applications.

The pure-projective component is characterized by a vanishing line \(l_\infty = (l_1, l_2, l_3)^T\) of the ground plane, which has the form as

\[\mathbf{P} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ l_1 & l_2 & l_3 \end{pmatrix}\]

(7)

Recovery of the pure-projective component \(\mathbf{P}\) achieves affine rectification of the ground plane.

Extending affine rectification to metric rectification involves estimation of the affine component \(\mathbf{A}\) with the form:

\[\mathbf{A} = \begin{pmatrix} 1 & \frac{1}{\beta} & \frac{1}{\alpha} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\]

(8)

This matrix has two degrees of freedom represented by \(\alpha\) and \(\beta\), which specify the image of the circular points [33]. The circular points are invariant under similarity transformations, but are transformed from metric coordinates \((1, \pm i, 0)^T\) to affine coordinate \((\alpha \pm i\beta, 1, 0)^T\) by the affine transformation \(\mathbf{A}\).

In the following, we will describe how to make use of traffic scene constraints to achieve ground plane rectification, which can be seen as a more accurate modeling of scene consistent information.

#### 3.1. Affine rectification

Affine rectification of the ground plane requires identification of the vanishing line \(l_\infty\) of the ground plane, which can be determined by two horizontal vanishing points. It is related to our previous work of making use of motion and appearance information of moving objects for convenient camera calibration [34], which is introduced in detail as follows.

We make use of coarse vehicle detection in surveillance videos to estimate two horizontal vanishing points. As we know, moving objects in traffic scenes can be detected accurately with shadows removed by improved GMM [31], but we need to further distinguish vehicles from pedestrians, bicycles and outliers. The difference in the following two directions are taken as a distinctive feature for coarse vehicle detection. The first direction is the velocity direction of objects in videos, which can be calculated due to position change of unit time. The second is the main axis direction \(\theta\), which can be estimated from moment analysis of silhouette:

\[\theta = \arctan\left(\frac{2\mu_{11}}{\mu_{20} - \mu_{02}}\right)\]

(9)
here $\mu_{pq}$ is the central moment of order $(p,q)$. It is evident that the angle difference is very small for moving vehicles, while it is significant for pedestrians and bicycles as illustrated in Fig. 3. Instead of $K$-Mean clustering, we adopt a more reliable strategy for coarse vehicle detection. Only those objects with angle difference less than $\theta_I = 5^\circ$ are labeled as vehicles with all of the rest discarded. The latter estimation of vanishing points benefits from this strict detection strategy.

The second constraint to use is the geometry constraint of vehicles in traffic scene surveillance videos. As we know, vehicles in videos are rich in line segments along two orientations corresponding to the symmetrical axis direction and its perpendicular direction. We make use of image gradient to extract these two accurate line equations for every vehicle detected from videos. As shown in Fig. 4, these two orientations are extracted by two stages of Histogram of Orientation Gradient (HOG). The first round HOG is of larger bins, which is adopted to obtain two peaks as rough orientations of vehicles. Then, the second round HOG is of much smaller bins, which are conducted respectively in the neighborhood of each peak to obtain precise estimations of the two orientations. The respective line equations are further determined by correlation to image data. Motion direction can help us to distinguish these two lines. The one with orientation closer to the motion direction corresponds to the symmetrical axis direction, while the other one corresponds to its perpendicular direction.

Imagine that there is only one straight roadway in the surveillance scene, which can treated as the simplest case. It is very common that most of vehicles should be parallel to each other since they should follow the road way in two opposite directions. It can be further derived that the symmetrical axis of most vehicles in 3D world should be parallel to each other. Due to the image projection, they are no longer parallel but intersect at the same point called horizontal vanishing point on the image plane. The perpendicular direction is of the same case. With these two lines estimated from each vehicle as we described before, we can make use of the voting strategy to estimate these two horizontal vanishing points. For every line $l$ extracted from vehicles, each point $s(x,y)$ lying on $l$ generates a Gaussian impulse in voting space with $(x,y)$ as its center. With time accumulated, a voting surface is generated and the position of its global extreme corresponds to the estimated intersection point. One example of voting space corresponding to the roadway direction is shown in Fig. 5.

Evidently, the simplest case of only one straight roadway is usually not satisfied in reality. There may be more than one roadway in the view filed, like a crossroad. The roadway may be not straight at all, like a roundabout. In these cases, the method described above cannot applied to estimate the two horizontal vanishing points. Fortunately, the variance of roadway layouts in reality can be seen as combinations of several primitive layouts which can be solved for horizontal vanishing points estimation.

Fig. 3. Direction difference (red arrowhead stands for velocity direction and blue arrowhead stands for main axis direction) [32]: (a) illustration of pedestrians; (b) illustration of bicycles and (c) illustration of vehicles. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fig. 4. Estimation of line equations from vehicles [32].
Details of the algorithm to achieve horizontal vanishing points estimation in all kinds of traffic scene layouts can be found in our previous work focusing on practical camera calibration from traffic scenes [34], which is directly adopted in our system.

3.2. Metric rectification

As described in [35], each known angle $\theta$ on the world plane between line $l_a$ and line $l_b$ on the image plane gives a constraint of $(\alpha, \beta)$ to lie on a circle with center $(c_a, c_b)$ and radius $r$:

$$\begin{align*}
(c_a, c_b) &= \left(\frac{(a+b)}{2}, -\frac{a-b}{2}\cot \theta\right), \\
r &= \frac{|a-b|}{2\sin(\theta)}
\end{align*}$$

(10)

where $a = -l_{a2}/l_{a1}$ and $b = -l_{b2}/l_{b1}$ are the line directions.

As described above, each detected vehicle gives two perpendicular directions on the world plane to determine a circle about $(\alpha, \beta)$. Since there are redundant detected vehicles from videos, we can determine $(\alpha, \beta)$ as the intersection of a large set of estimated circles as shown in Fig. 6(a).

Due to the symmetric property, we only focus on the intersection above the $z$ axis. Every two circles can determine the intersection conveniently by differential of these two circle equations. For $N$ circles, we can obtain $N(N-1)/2$ candidate points and $(\alpha, \beta)$ is determined simply based on a Gaussian voting strategy described above as shown in Fig. 6(b). With $(\alpha, \beta)$ estimated, we can calculate the affine matrix $A$ so that the metric rectification is realized.

3.3. Object classification

With the ground plane rectification achieved, 2D object features can be normalized based on the homography to be effective for object classification. We still adopt the online learning framework described in Section 2 for classification but change some details, which is shown in Fig. 7.

The 5 shape and motion features described in Section 2 are still adopted but normalized by ground plane rectification, which are denoted as $(size_e, speed_e, compactness_e, size_e', angle_e')$.
In the phase of automatic labeling, we still adopt the K-Mean clustering and decision level fusion. The rectified features ($\text{compactness}_R$, $\text{size}_R$, $\text{angle}$) are adopted in K-Mean clustering. Accordingly, the decision level fusion is based on the following three new rules:

1. $\text{compactness}_R$ has the advantages of distinguishing vehicles from pedestrians and bicycles;
2. $\text{size}_R$ has the advantages of distinguishing pedestrians from vehicles and bicycles; and
3. $\text{angle}$ has the advantages of classifying pedestrians and vehicles.

In the phase of Bayesian classification, we no longer need to maintain independent classifiers for each grids but only one classifier for the whole scene. We adopt a new Gaussian assumption that $\nu' = (\text{size}_R, \text{speed}_R, \text{compactness}_R)$ of every category satisfies a multivariate Gaussian distribution, which will be further tested in Section 4. Online learn based strategy is the same as we described in Section 2. The category is determined by the posterior probability and the classifier is refined at the same time to be robust to condition changes. Tracking can still be adopted to enhance the performance of object classification by determine the category using the sum of posterior probability of tracked frames.

4. Experimental results and analysis

Numerous experiments are conducted and experimental results are presented in this section to demonstrate the performance of the proposed approach. All experiments are conducted on a computer of P4 3.0 CPU and 512 M DDR.

4.1. Testing of scene division based strategy

Scene division based strategy achieves a transformation from a wide view surveillance scene to narrow view surveillance scenes, which is a rough modeling of scene consistency information and should improve the performance of object classification. The division level would definitely effect the performance of object classification. A video (Video 1) of a typical traffic scene illustrated in Fig. 2(a) is processed for testing. The scene is respectively divided into $1 \times 1, 2 \times 2, 3 \times 4, 4 \times 4, 5 \times 5$ and the classification performance is shown in Fig. 8. As we can see, the accuracy increases with higher level of division, which demonstrates the advantage of the scene division strategy to conquer perspective distortions. However, we should attention the trade-off between the performance and computational cost. In the case of Video 1, $4 \times 4$ should be a appropriate division level in practice.

With the scene of Video 1 divided into $4 \times 4$ grids, the classification performance is further shown in Table 1. It is based

<table>
<thead>
<tr>
<th>Grids</th>
<th>Pedestrian (%)</th>
<th>Bicycle (%)</th>
<th>Vehicle (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 x 1</td>
<td>98.2</td>
<td>1.8</td>
<td>0.0</td>
</tr>
<tr>
<td>2 x 2</td>
<td>3.4</td>
<td>90.4</td>
<td>6.2</td>
</tr>
<tr>
<td>3 x 4</td>
<td>100</td>
<td>2.7</td>
<td>97.3</td>
</tr>
<tr>
<td>4 x 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 x 5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 8. Different accuracy with different extent of division: (a) classification accuracy of the category of pedestrian; (b) classification accuracy of the category of bicycle and (c) classification accuracy of the category of vehicle.
on a strong assumption that the selected feature vector of every category satisfies a multivariate Gaussian assumption in subregions, which can then be combined with Bayesian theory for classification. If this assumption is ignored, we can still achieve online object classification by comparing distances between the feature vector and each of the $K$-Mean cluster. The results of this simpler strategy are shown in Table 2. As we can see, the classification based on the Gaussian assumption is much better than the simply distance based decision, which illustrates the practicability of Gaussian assumption for scene division based object classification. However, the inappropriate division level illustrated in Fig. 8 may cause even worse performance than the distance comparison based strategy. It is because low level scene division cannot conquer the effect of perspective distortions, which make the feature vector cannot satisfy Gaussian distributions at all.

The above experiments show that scene division based strategy and the Gaussian assumption can conquer perspective distortions to some extent and achieve better classification performance. However, the accuracy of Gaussian assumption is strongly related to the scene division level, which cannot be accurately determined in practice.

4.2. Testing of the ground plane rectification accuracy

Compared to scene division based strategy, ground plane rectification can be seen as a more refined model to describe scene consistency information.

Experiments are firstly conducted to test the accuracy of our ground plane rectification. As we know, with the affine and metric rectification of the ground plane realized, it is possible to recover the relative length or ratio of lines on the ground plane in an Euclidean ambiguity, which is also a good way to test the accuracy of the ground plane rectification. A typical surveillance scene is shown in Fig. 9. In this scene, 16 lines and 5 right-angles are measured from images. The measured value and the corresponding ground truth are shown in Table 3.

As we can see, the average error of measurement is less than 6%. The accuracy of the ground plane rectification makes it possible to achieve good performance of view independent object classification.

4.3. Illustration of appearance rectification

As we have described before, object appearance has significant distortions due to camera projections. The ground plane rectification enables normalization of appearance. Here, we take object silhouette as an example to illustrate the performance of appearance rectification. Experiments are conducted to a vehicle moving across a wide-view traffic scene. The original projected silhouettes and corresponding rectified silhouettes are shown in Fig. 10. As we can see, after rectification, the object silhouette is approximately invariant in wide-view scenes, which illustrates the effectiveness of appearance rectification.

4.4. Illustration of feature rectification

The ground plane rectification enables normalization of 2D object features to be robust to view angle changes. With abundant moving objects extracted by motion detection and labeled manually, we analyze the class-conditional densities of the three object features before and after normalization as shown in Figs. 11–13.

![Fig. 9. Metric recovery in surveillance scenes [32].](image1)

<table>
<thead>
<tr>
<th>Label</th>
<th>$l_1$</th>
<th>$l_2$</th>
<th>$l_3$</th>
<th>$l_4$</th>
<th>$l_5$</th>
<th>$l_6$</th>
<th>$l_7$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test</td>
<td>0.94</td>
<td>1.03</td>
<td>1.02</td>
<td>0.97</td>
<td>0.99</td>
<td>1.01</td>
<td>1.97</td>
</tr>
<tr>
<td>Real</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>2.11</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Label</th>
<th>$l_8$</th>
<th>$l_9$</th>
<th>$l_{10}$</th>
<th>$l_{11}$</th>
<th>$l_{12}$</th>
<th>$l_{13}$</th>
<th>$l_{14}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test</td>
<td>2.17</td>
<td>2.21</td>
<td>1.21</td>
<td>1.19</td>
<td>2.07</td>
<td>2.09</td>
<td>2.10</td>
</tr>
<tr>
<td>Real</td>
<td>2.11</td>
<td>2.11</td>
<td>1.13</td>
<td>1.13</td>
<td>2.11</td>
<td>2.11</td>
<td>2.11</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Label</th>
<th>$l_{15}$</th>
<th>$l_{16}$</th>
<th>$\theta_1$</th>
<th>$\theta_2$</th>
<th>$\theta_3$</th>
<th>$\theta_4$</th>
<th>$\theta_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test</td>
<td>1.09</td>
<td>1.11</td>
<td>87.3</td>
<td>91.9</td>
<td>88.6</td>
<td>89.2</td>
<td>90.9</td>
</tr>
<tr>
<td>Real</td>
<td>1.13</td>
<td>1.13</td>
<td>90</td>
<td>90</td>
<td>90</td>
<td>90</td>
<td>90</td>
</tr>
</tbody>
</table>

![Fig. 10. Rectification of silhouette for a moving vehicle ((a)–(f) are original silhouettes; (g)–(l) are rectified silhouette) [32].](image2)
As we can see, the rectified features are much easier to be classified than original features. Furthermore, the class-conditional densities of these three features approximately satisfy Gaussian distributions, which can further illustrate that the Gaussian assumption makes sense to help object classification.

4.5. Classification performance of ground plane rectification

We test the performance of our approach in four traffic scenes of different view angles as shown in Fig. 14 to test the performance of the proposed object classification approach. They are of 30 min videos (45,000 frames), with the first 15,000 frames adopted to achieve ground plane rectification and the rest to test the performance of object classification. The average classification accuracy of the proposed method is shown in Table 4.

As we can see, the performance of our method is still better than the scene division based strategy. It is because the ground plane rectification can model the scene consistency information much better than a simple scene division.

4.6. Discussion

In our approach, initialization and refinement of classifiers are carried out online. With the scene change detection modular, our approach can detect scene changes and adapt to new scenes automatically. Conventional object tracking can improve the performance of object classification with temporal information based on the decision level fusion. In our implementation, our algorithms can deal with videos with the speed of more than 15 frames per second, which basically achieves real-time performance. Even though the ground plane rectification based strategy requires an additional phase to achieve ground plane rectification, the computational cost is better than the scene division based strategy. It is because scene division based strategy requires to maintain many classifiers on the fly.

The degenerate case of ground plane rectification based approach corresponds to the top-down camera view. However, perspective distortions is not evident in top-down views. Since the ground plane rectification is taken to deal with perspective distortions, top-down view is not a big problem of our approach. Furthermore, traffic scene surveillance prefers to mount cameras with an oblique to the ground plane to obtain a wider view field.

From all of the above, we can see that our online learning based object classification algorithms have many desired convenient properties. It is efficient to be real-time, effective and view independent. Furthermore, the algorithms are free of manual labeling and supervised learning, and can deal with environment changes very well, which have great potential to be applied in real applications.

5. Conclusions

In this paper, we focus on view independent object classification by exploring hidden scene consistency information. Two kinds of models are proposed to describe scene consistency information and combined with the online learning based framework. The scene division based strategy can achieve scene consistency modeling
roughly and improve object classification in an appropriate scene division level. Ground plane rectification can model scene consistency information more accurately and achieves higher classification accuracy. Experimental results demonstrate the performance of our framework to improve object classification by scene consistency information modeling. It is shown that the classifiers are initialized and refined online, free of manual labeling, and adaptive to all kinds of condition changes. The scene consistency information based object classification has great potential to be adopted in real applications.

Acknowledgments

This work is funded by the National Basic Research Program of China (No. 2010CB327902), the National Natural Science Foundation of China (No. 61005016), the Fundamental Research Funds for the Central Universities, and the Open Projects Program of National Laboratory of Pattern Recognition.

References


J.Y. Bouguet, Camera calibration toolbox for MATLAB, 2008.


Zhaoxiang Zhang, Min Li, Kaiqi Huang, Tieniu Tan, Robust automated ground plane rectification based on moving vehicles for traffic scene surveillance, in: Proceedings of the International Conference on Image Processing, 2008.


Zhaoxiang Zhang, Tieniu Tan, Kaiqi Huang, Yunhong Wang, Practical camera calibration from moving objects for traffic scene surveillance, IEEE T. Circ. Syst. Vid. in press, http://dx.doi.org/10.1109/TCSVT.2012.2210670.


Kaiqi Huang received the B.Sc. and M.Sc. degrees from Nanjing University of Science Technology, Nanjing, China, and the Ph.D. degree from Southeast University, Nanjing, China, in 1995 and 1998, respectively. From 1998 to 2004, he was with the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Science, Beijing, China, and in 2005, became an Associate Professor. He is the author of over 70 papers in the international journals and conferences such as T-PAMI, TIP, T-CSVT, T-SMCB, Pattern Recognition, ECCV, CVPR, ICIP, and ICFR. He is in charge of several fundings, including National Science Funding and collaboration with some enterprises. He is also involved in several national research projects (e.g., 863, 973, and NSFC). His current research interests include visual-surveillance digital image processing, pattern recognition, biological-based vision, etc.

Yunhong Wang received the B.S. degree in electronic engineering from Northwestern Polytechnical University, Xi’an, China, in 1989 and the M.S. and Ph.D. degrees in electronic engineering from Nanjing University of Science and Technology, Nanjing, China, in 1995 and 1998, respectively. From 1998 to 2004, she was with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. Since 2004, she has been a Professor with the School of Computer Science and Engineering, Beihang University, Beijing, where she is also the Director of the Laboratory of Intelligent Recognition and Image Processing, Beijing Key Laboratory of Digital Media. Her research interests include biometrics, pattern recognition, computer vision, data fusion, and image processing.

Min Li received his B.Sc. degree in electronic information science and technology from the University of Science and Technology of China (USTC) in 2004, and his M.Sc. in electrical engineering from Institute of Electrical Engineering, Chinese Academy of Sciences in 2007, and his Ph.D. degree in pattern recognition and intelligent system from Institute of Automation, Chinese Academy of Sciences in 2011. His current research interests include computer vision, pattern recognition, and machine learning.