Common Fate Hough Transform for Concurrent Detection and Recognition

Anonymous CVPR submission

Paper ID 745

Abstract

This paper proposes a method for concurrent detection and recognition of multiple multi-class objects by extending the probabilistic Hough transform to incorporate grouping results of the voting elements. In the method, each object part casts votes about the center and the class label of the object that generates itself. These object parts that will vote are firstly grouped by their motion patterns. The votes of each object part are assigned different priors according to the votes of the other object parts in the same motion group. The traceability of the Hough image formed by these votes is greatly enhanced. In such a manner, the proposed method meets the challenges of separating near objects and separating similar objects, while using a codebook trained from images with a clustered background. Experimental results are provided to show the merit in terms of detection and recognition accuracy.

1. Introduction

The goal of the proposed method is to detect and recognize multiple multi-class objects on a given video frame. This is of importance for many applications, like surveillance and video analysis.

One reason, the generalized Hough transform [1] is favorable for detection and/or recognition, is its robustness to partial deformation, besides eigen window methods [15], pictorial structures [4], constellation models [5], sliding window classifiers [10, 22], and branch-and-bound methods [22]. Object parts act as voting elements in the Hough transform based methods, in the form of keypoint descriptors [11], image patches [6, 16], or image regions [7]. Each voting element votes for hypotheses that generate itself. The votes from different voting elements are added up to form a Hough image. The peaks of the Hough image are considered as detection hypotheses with the height of each peak as the confidence of the corresponding hypothesis. The simplicity of the learning procedure is another reason that the Hough transform based methods are attractive. An appearance codebook of object parts is built from a set of images on each of which the object center is annotated. Each code encodes the appearance and the offset to the object center of an object part. This offset acts as a Hough vote at run time, when object parts from the current image are matched against the codebook.

Besides the recent usage for various vision tasks [13, 21, 9, 17], the Hough transform based methods are extended for better detection performance. The implicit shape model [11, 12] is extended by notifying correspondences between the object parts and the hypotheses [2] for the detection of multiple near objects. The Hough transform is placed in a discriminative framework for object detection [13] in a way that the codes are assigned different weights by the co-occurrence frequency of their appearance and offset to the object center.

The proposed method not only detects multiple objects, but also recognizes each detected object. For recognition, the class label of each training image is annotated besides the object center. Then each code contains three aspects of information: the appearance, the offset to the object center, and the class label. The Hough image is formed by adding up votes with class labels, and the peaks of the Hough image are detection hypotheses with class labels. For each object, its location and class label are found from the hypotheses. However, for realistic scenarios with multiple multi-class objects, the Hough image formed in this manner is difficult to search over for concurrent detection and recognition.

One reason is that the object parts of the codebook are from multi-class objects, which even of the same class vary in poses. Thus the portion of the correct votes is low. For the traceability of the Hough image, the codebook needs to be very effective. So, the training images where objects appear as foreground need to have a very clean background. Otherwise, keypoints on the background act as noise and lead to false votes. The noisy votes make the portion of the correct votes even lower. Though computer graphics rendering provides a good source of clean-background images for the training of voting based methods [14], in order to use real-world images with a clustered background, manual efforts are needed to mark the foreground, or a very large number of training images are needed to build the codebook, which

harms efficiency.

Even if a very effective codebook is successfully built, the Hough image built upon a scene containing near objects and similar objects is still difficult to search over. This is due to two reasons: (1) votes casted by parts from two near objects make the peaks corresponding to different objects mixed up, and (2) different-class object parts are sometimes very similar, and this leads to tough decisions on the class label of the peaks.

To utilize the codebook built from training images with a clustered background and to meet the challenges of detecting near objects and recognizing similar objects, this paper proposes a concurrent detection and recognition method based on the common fate principle [20]. The principle is one of the four visual perception principles as theorized by gestalt psychologists. For humans, tokens moving coherently are perceptually grouped, and this provides an intuition to group the object parts (keypoint descriptors) that will vote by their motion patterns. The motion of each keypoint is represented by a trajectory generated by tracking the keypoint through frames. The keypoints are then grouped using the pairwise similarities of their corresponding trajectories. Among the votes of each object part, those which are more "agreeable" by the votes of the other object parts in the same motion group are assigned higher priors. This results in correct votes being more likely to be assigned higher priors. These votes with different priors enhance the traceability of the Hough image for concurrent detection and recognition.

To the best of our knowledge, this is the first to incorporate motion information in the Hough transform based methods for concurrent detection and recognition. Additionally, the proposed method has several appealing properties:

- The method realizes concurrent detection and recognition of multiple multi-class objects. The existence of three types of objects makes the task challenging: near objects, similar different-class objects, and multi-pose same-class objects.
- Its ability to use a codebook trained by images with a clustered background.
- The manner in which the grouping results by motion is integrated into the Hough transform is very general, and it can be used for incorporating other grouping results.

The rest of the paper is organized as follows: The probabilistic Hough transform formulism for a concurrent detection and recognition problem is given in section 2. In Section 3 the formulism of the common fate Hough transform is given. The inference for concurrent detection and recognition is described in Section 4. Then experimental results are given in section 5, and section 6 concludes.

2. Probabilistic Hough Transform

This section describes how a Hough image for concurrent detection and recognition is formed from object parts observed on an image.

Let $\mathbf{e}$ denote an object part observed on the image. The appearance of $\mathbf{e}$ is matched against the codebook, and $\mathbf{e}$ activates N best matched codes from the trained codebook. Each code contains the appearance, its offset to the object center, and the class label. According to the N matched codes, $\mathbf{e}$ casts N votes. Each vote $V_{\mathbf{e}}$ is about the object center that generates $\mathbf{e}$. The position of the object center casted by V is denoted by $\mathbf{x}_V$, while the class label by l_V. Based on the N votes of $\mathbf{e}$, the probability that a position $\tilde{\mathbf{x}}$ is the center of an object with class label $\tilde{l}$ is given by,

$$p(\tilde{\mathbf{x}}, \tilde{l}|\mathbf{e}) = \sum_{i=1}^{N} p(\tilde{\mathbf{x}}, \tilde{l}|V_{\mathbf{e}}^i) p(V_{\mathbf{e}}^i|\mathbf{e}) \,. \tag{1}$$

Here $p(\tilde{\mathbf{x}}, \tilde{l}|V_{\mathbf{e}}^i)$ is the likelihood of $\tilde{\mathbf{x}}$ being an object center of class $\tilde{l}$ based on $V_{\mathbf{e}}^i$. And $p(V_{\mathbf{e}}^i|\mathbf{e})$ is the prior of $V_{\mathbf{e}}^i$, given $\mathbf{e}$.

The idea of the proposed method is that, the prior term, $p(V_{\mathbf{e}}^i|\mathbf{e})$, is defined by the motion grouping results of all the object parts.

The likelihood term is defined as,

$$p(\tilde{\mathbf{x}}, \tilde{l}|V) = \begin{cases} 0 & \text{if } l_V \neq \tilde{l} \text{ or } |\tilde{\mathbf{x}} - \mathbf{x}_V| > d \\ G(\tilde{\mathbf{x}}; \mathbf{x}_V, \sigma) & \text{otherwise} \end{cases} . \tag{2}$$

Here $G(\tilde{\mathbf{x}}; \mathbf{x}_V, \sigma)$ is a Gaussian function that fixes the spacial gap between $\tilde{\mathbf{x}}$ and $\mathbf{x_V}$.

Let M be the total number of object parts on the image, and assume the object parts are mutually independent. Then by marginalizing over all the object parts, the probability of $\tilde{\mathbf{x}}$ being the center of a $\tilde{l}$-class object is given by,

$$\begin{aligned} p(\tilde{\mathbf{x}}, \tilde{l}) &= \sum_{j=1}^{M} p(\tilde{\mathbf{x}}, \tilde{l}|\mathbf{e}_j) p(\mathbf{e}_j) \\ &= \sum_{j=1}^{M} \sum_{i=1}^{N} p(\tilde{\mathbf{x}}, \tilde{l}|V_{\mathbf{e}_j}^i) p(V_{\mathbf{e}}^i|\mathbf{e}_j) p(\mathbf{e}_j) \,. \end{aligned} \tag{3}$$

Usually, a uniform prior is assumed for each object part, and $p(\mathbf{e}_j) = \frac{1}{M}$. Then by considering $p(\tilde{\mathbf{x}}, \tilde{l})$ as the evaluation score of the Hough space $(\tilde{\mathbf{x}}, \tilde{l})$, the task of concurrent detection and recognition converts to finding and then validating the local maxima of the Hough image.

(a) (b) (c) (d)

Figure 1. Effect of the proposed prior. The motion grouping results are given in (a), and different motion groups are marked with different colors. The object centers voted according to the 7 best matched codes are given in (b). Red circles are centers of pedestrians while blue circles are centers of bicycle riders. The voted centers given in (c) are the voted centers of the 7 votes with the highest defined priors from 35 votes. The object centers in (d) are voted by votes with priors higher than 0.1.

3. Common Fate Hough Transform

Since the codebook is trained by multi-class object parts and the training images have a clustered background, the best matched codes here cannot guarantee the traceability of the Hough image. This section describes how, by considering the motion grouping results of the object parts, different priors are assigned to the votes of each object part.

Let $\gamma = \{\mathbf{g}\}$ denote the grouping results, where $\mathbf{g}$ is a group of object parts, i.e., $\mathbf{e}_m \in \mathbf{g}$ and $\mathbf{e}_n \in \mathbf{g}$. Those votes of $\mathbf{e_m}$ which are more "agreeable" by the votes of the other objects in $\mathbf{g}$ are assigned higher priors.

Towards this end, the relationship between the votes of $\mathbf{e_m}$ and the votes of $\mathbf{e_n}$ needs to be given in advance. This relationship is named support. The support from $V_{\mathbf{e}_n}$ to $V_{\mathbf{e}_m}$ is defined by that based on $V_{\mathbf{e}_n}$, the possibility $V_{\mathbf{e}_m}$'s voted center is correct, as,

$$S(V_{\mathbf{e}_n} \to V_{\mathbf{e}_m}) = p(\mathbf{x}_{V_{\mathbf{e}_m}}, l_{V_{\mathbf{e}_m}} | V_{\mathbf{e}_n}) \, , n \neq m \, .$$

Here $p(\mathbf{x}_{V_{\mathbf{e}_m}}, l_{V_{\mathbf{e}_m}} | V_{\mathbf{e}_n})$ is defined in (2). This measures the coherence of the two votes from different object parts.

Then, the support from $\mathbf{e}_n$ to $V_{\mathbf{e}_m}$ is defined by that based on $\mathbf{e}_n$, the possibility $V_{\mathbf{e}_m}$'s voted center is correct, as,

$$\begin{aligned} S(\mathbf{e}_n \to V_{\mathbf{e}_m}) &= p(\mathbf{x}_{V_{\mathbf{e}_m}}, l_{V_{\mathbf{e}_m}} | \mathbf{e}_n) \\ &= \sum_{i=1}^{N} p(\mathbf{x}_{V_{\mathbf{e}_m}}, l_{V_{\mathbf{e}_m}} | V^i_{\mathbf{e}_n}) p(V^i_{\mathbf{e}_n} | \mathbf{e}_n) \\ &= \sum_{i=1}^{N} S(V^i_{\mathbf{e}_n} \to V_{\mathbf{e}_m}) p(V^i_{\mathbf{e}_n} | \mathbf{e}_n) \, , n \neq m \, . \end{aligned}$$

And the support from $\mathbf{g}$ to $V_{\mathbf{e}_m}$ is defined by the possibility that $V_{\mathbf{e}_m}$'s voted center is correct based on the votes of all the other object parts but its belonging object part in $\mathbf{g}$, as,

$$\begin{aligned} S(\mathbf{g} \to V_{\mathbf{e}_m}) &= \sum_{\mathbf{e}_i \in \mathbf{g} - \{\mathbf{e}_m\}} p(\mathbf{x}_{V_{\mathbf{e}_i}}, l_{V_{\mathbf{e}_m}} | \mathbf{e}_i) p(\mathbf{e}_i) \\ &= \frac{1}{M} \sum_{\mathbf{e}_i \in \mathbf{g} - \{\mathbf{e}_m\}} S(\mathbf{e}_i \to V_{\mathbf{e}_m}) \, . \end{aligned}$$

Assuming all object parts in the same motion group are from the same object, which means motion grouping gives good results. The center position and the class label given by one vote shall be consistent with that given by the motion group. Thus for a particular vote of $\mathbf{e}_m$, i.e., $\tilde{V}_{\mathbf{e}_m}$, a prior is assigned to it by considering its consistence with $\mathbf{g}$ and the consistence of $\mathbf{e}_m$'s other votes with $\mathbf{g}$, as:

$$\begin{aligned} p(\tilde{V}_{\mathbf{e}_m} | \mathbf{e}_m) &= \frac{S(\mathbf{g} \to \tilde{V}_{\mathbf{e}_m}) + \frac{\Delta}{N}}{\sum_{i=1}^{N} S(\mathbf{g} \to V^i_{\mathbf{e}_m}) + \Delta} \\ &= \frac{\sum_{\mathbf{e}_j \in \mathbf{g} - \{\mathbf{e}_m\}} S(\mathbf{e}_j \to \tilde{V}_{\mathbf{e}_m}) + \frac{M\Delta}{N}}{\sum_{i=1}^{N} \sum_{\mathbf{e}_j \in \mathbf{g} - \{\mathbf{e}_m\}} S(\mathbf{e}_j \to V^i_{\mathbf{e}_m}) + M\Delta} \\ &= \frac{\sum_{\mathbf{e}_j \in \mathbf{g} - \{\mathbf{e}_m\}} \sum_{k=1}^{N} S(V^k_{\mathbf{e}_j} \to \tilde{V}_{\mathbf{e}_m}) p(V^k_{\mathbf{e}_j} | \mathbf{e}_j) + \frac{M\Delta}{N}}{\sum_{i=1}^{N} \sum_{\mathbf{e}_j \in \mathbf{g} - \{\mathbf{e}_m\}} \sum_{k=1}^{N} S(V^k_{\mathbf{e}_j} \to V^i_{\mathbf{e}_m}) p(V^k_{\mathbf{e}_j} | \mathbf{e}_j) + M\Delta} \, . \end{aligned} \tag{4}$$

Here, Δ is a small constant for preventing zeros. Notice, $p(\tilde{V}_{\mathbf{e}_m} | \mathbf{e}_m)$ is defined using $p(V^k_{\mathbf{e}_j} | \mathbf{e}_j)$, the priors of the votes of the other object parts in $\mathbf{g}$. In order to give $p(\tilde{V}_{\mathbf{e}_m} | \mathbf{e}_m)$, uniform priors are firstly assigned to the votes of each object part in $\mathbf{g}$, i.e., $p(V^k_{\mathbf{e}_j} | \mathbf{e}_j) = \frac{1}{N}$. Then new priors are calculated based on the uniformly assigned priors. The priors of votes to form the Hough image are priors converged in iterations.

The grouping results $\gamma = \{\mathbf{g}\}$, can be replaced by other grouping results of the voting elements. The proposed

method uses motion to group the voting elements. The extended Hough transform with motion grouping results is called the common fate Hough transform. The voted centers voted by votes according to the best matched codes and the centers voted by votes with the highest priors are shown in Figure 1. The centers voted by votes with highest priors defined here are more concentrated to the true object centers than the centers voted according to the best matched codes.

3.1. Motion Grouping

In order to group the object parts by their motion patterns. The object parts are tracked through frames before and after the current frame to generate trajectories. The object parts in this method are in the form of keypoint descriptors. The Harris Corner [8] feature is chosen to represent the object part, while for appearance, the region covariance [19] of the image patch around the keypoint is used.

For each object part, a trajectory is generated by tracking its corresponding Harris Corner by the KLT tracker [18].

To group the trajectories, pairwise similarities are firstly defined. Let $T_{\mathbf{e}_m}$ and $T_{\mathbf{e}_n}$ denote two trajectories corresponding to $\mathbf{e}_m$ and $\mathbf{e}_n$. The first similarity between two trajectories is defined as,

$$D_1(T_{\mathbf{e}_m}, T_{\mathbf{e}_n}) = \max_{i=1\ldots L} (Euclid(\mathbf{x}^i_{T_{\mathbf{e}_m}}, \mathbf{x}^i_{T_{\mathbf{e}_n}})) \, .$$

Here, $Euclid$ calculates the Euclidean distance, i is the frame index, and L is the length of the overlapping part of the two trajectories. To define the second similarity, the ith directional vector of T is firstly defined as, $\mathbf{d}^i_T = \mathbf{x}^{i+3}_T - \mathbf{x}^i_T$. Let $\mathbf{a}_i = \mathbf{d}^i_{T_{\mathbf{e}_m}}$, $\mathbf{b}_i = \mathbf{d}^i_{T_{\mathbf{e}_n}}$, $a_i = \frac{\mathbf{a}_i \cdot \mathbf{b}_i}{\mathbf{a}_i \cdot \mathbf{a}_i}$, and $b_i = \frac{\mathbf{a}_i \cdot \mathbf{b}_i}{\mathbf{b}_i \cdot \mathbf{b}_i}$. Then the second similarity is defined as,

$$D_2(T_{\mathbf{e}_m}, T_{\mathbf{e}_n}) = \max_{i=1\ldots L-3} (\max(|\mathbf{a}_i - a_i\mathbf{a}_i|, |\mathbf{b}_i - b_i\mathbf{b}_i|)) \, .$$

To group the trajectories, the static points are firstly excluded. Inspired by [3], a minimal spanning tree of the trajectories is built upon D_1, and split by cutting edges larger then a threshold, D^1_{th}. For each element of the splitting results, a minimal spanning tree is built upon D_2 and split by cutting the edges larger than a threshold, D^2_{th}. This hierarchical procedure ensures that trajectories in the same group have both small D_1 and D_2.

Each trajectory corresponds to an object part, and the grouping results of the trajectories correspond to grouping results of the object parts.

3.2. Codebook

For training, Harris corners are extracted from the training images with the object center and the class label annotated. In this method, region covariance is chosen to represent the appearance, which is defined as,

$$\mathbf{r} = \frac{1}{K-1} \sum_{i=1}^{K} (\mathbf{z}_i - \mu)(\mathbf{z}_i - \mu)^T \, .$$

Here, K is the number of pixels in the region, and $\mathbf{z}_i$ is a 7-dimensional vector regarding the (x, y) coordinate of the pixel, while μ is the mean of $\mathbf{z}_i$. And $\mathbf{z}(x, y)$ contains the RGB color of the pixel and intensity gradient of the pixel, as: $r(x, y)$, $g(x, y)$, $b(x, y)$, $|\frac{\partial I(x,y)}{\partial x}|$, $|\frac{\partial I(x,y)}{\partial y}|$, $|\frac{\partial^2 I(x,y)}{\partial x^2}|$, and $|\frac{\partial^2 I(x,y)}{\partial y^2}|$.

The appearance similarity between $\mathbf{r}_m$ and $\mathbf{r}_n$ is given by,

$$\rho(\mathbf{r}_m, \mathbf{r}_n) = \sqrt{\sum_{i=1}^{7} \ln^2 \lambda_i} \, .$$

Here, λ_i is the generalized eigenvalue by solving the generalized eigenvalue problem, $\lambda_i \mathbf{r}_m \mathbf{u}_i = \mathbf{r}_n \mathbf{u}_i, \mathbf{u}_i \neq \mathbf{0}$, with $\mathbf{u}_i$ the eigenvector.

A square image patch around each keypoint is used to represent the appearance of an object part. Six region covariances are generated for each image patch by using the pixels of the top-left, the top-right, the bottom-left, the bottom-right, the central, and all of the image patch. Then besides the offset and the class label, a code contains six region covariances. When an object part is matched against the codebook, the similarity between the image patch of the object part and a code is defined by the smallest similarity of the corresponding region covariances.

4. Detection and Recognition

After forming the Hough image, the detection and recognition hypotheses are validated. Let $\mathbf{h} = \{H\}$ be the points in the Hough space which are evaluated by $p(\mathbf{x}_H, l_H)$ and have $p(\mathbf{x}_H, l_H) > 0$. Inspired by [2], the hypotheses are validated by an optimizing procedure. Let O be the number of the points in $\mathbf{h}$. let $u_i = 1$ or 0 indicate H_i being a true object center or not. The problem is:

$$\arg\max_{u_i} \prod_{i=1}^{O} p^{u_i}(H_i) \Longleftrightarrow \arg\max_{u_i} \sum_{i=1}^{O} u_i \ln(p(H_i)) \, .$$

Let $v_{ij} = 1$ or 0 indicate e_j belongs to H_i or not, then

$$\begin{aligned} p(H_i) &= \sum_{j=1}^{M} p(\mathbf{x}_{H_i}, l_{H_i} | \mathbf{e}_j) p(\mathbf{e}_j) \\ &= \frac{1}{M} \sum_{j=1}^{M} v_{ij} p(\mathbf{x}_{H_i}, l_{H_i} | \mathbf{e}_j) \, , \end{aligned}$$

and by assuming one object part belongs to and only belongs to one hypothesis, the problem is,

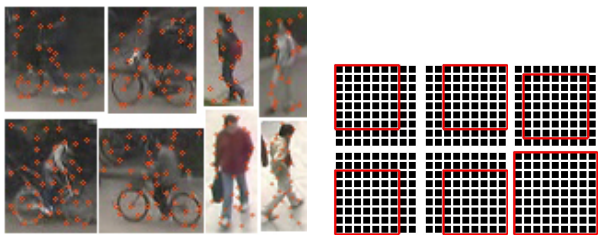
$$\arg\max_{u_i,v_{ij}} \sum_{i=1}^{O} u_i \ln(\sum_{j=1}^{M} v_{ij} p(\mathbf{x}_{H_i}, l_{H_i}|\mathbf{e}_j))$$
$$s.t.: \; u_i = 0 \text{ or } u_i = 1, \forall\, i$$
$$v_{ij} = 0 \text{ or } v_{ij} = 1, \forall\, i, \forall\, j$$
$$\sum_{i=1}^{O} v_{ij} = 1, \forall\, j$$
$$\sum_{j=1}^{M} v_{ij} <= u_i, \forall\, i\,.$$

Following [2], the optimal result for the problem is inferred by greedy maximization. As described in Algorithm 1, the largest local maximum of all the local maxima is chosen to be the center of a true object and then the object parts belonging to the chosen object center are excluded from the object part set. A new Hough image where new objects are found is formed using the remaining object parts. This procedure ends when the object part set is empty or the confidence of the chosen object is lower than a threshold.

Algorithm 1 Greedy Maximization

Let ε be the set of object parts, p_{th} be the low confidence threshold to accept detection responses, and $\hat{\mathbf{h}}$ be the local maxima of $\mathbf{h}$

while $\varepsilon \neq \emptyset$ **do**
  Form $\mathbf{h}$ with ε
  Generate $\hat{\mathbf{h}}$ and select $H_i \in \hat{\mathbf{h}}$, $u_i = 0$ and $\forall H^{'} \in \hat{\mathbf{h}}, p(\mathbf{x}_{H_i}, l_{H_i}) >= p(\mathbf{x}_{H^{'}}, l_{H^{'}})$
  if $p(\mathbf{x}_{H_i}, l_{H_i}) >= p_{th}$ **then**
    $u_i \leftarrow 1$
    for $\mathbf{e}_j \in \varepsilon$ **do**
      if $\forall H^{'} \in \hat{\mathbf{h}}, p(\mathbf{x}_{H_i}, l_{H_i}|\mathbf{e}_j) >= p(\mathbf{x}_{H^{'}}, l_{H^{'}}|\mathbf{e}_j)$ **then**
        $v_{ij} \leftarrow 1$
        $\varepsilon \leftarrow \varepsilon - \{\mathbf{e}_j\}$
      end if
    end for
  else
    for $\mathbf{e}_j \in \varepsilon$ **do**
      $v_{1j} \leftarrow 1$
    end for
    $\varepsilon \leftarrow \emptyset$
  end if
end while
return all $u_i, u_i = 1$



(a) (b)

Figure 2. Examples of training images are given in (a). The red circles are the extracted Harris corners. Some keypoints fall on the background. The six rectangles in (b) mark the pixels of the 9×9 image patch used for the six region covariances.

5. Experimental Results and Evaluation

In the experiments, the enhancement of the Hough image's traceability by the method is verified in terms of detection accuracy and recognition accuracy. The method is firstly tested on the *P-campus* dataset with [2] as benchmark, and then tested on a dataset of similar animals.

5.1. Campus Objects Detection and Recognition

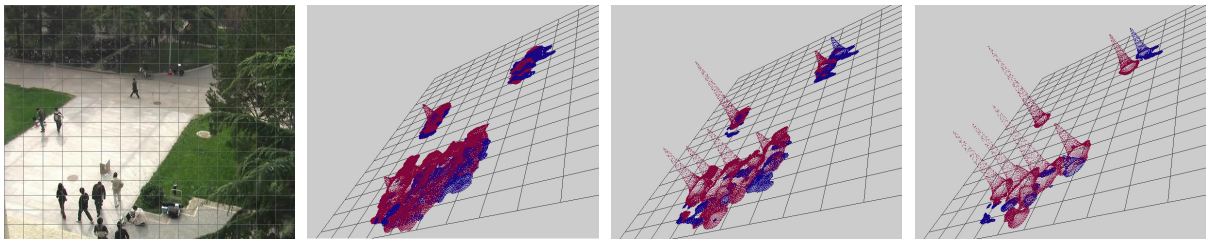
Dataset The *P-campus* dataset contains two primary classes of foreground objects: pedestrians and bicycle riders. The frame size is 720×576. Among all the 401 continuous frames, 633 different-class ground truth bounding boxes are annotated on 79 frames. In this dataset, pedestrians and bicycle riders have in common the upper human body, and pedestrians appear in front, back, and side views.

Implementation Settings For training, 52 images of bicycle riders and 171 images of pedestrians are randomly selected. Harris corners are generated on the image, and training image examples are given in Figure 2(a). For appearance, six region covariances are generated for each keypoint using the 9×9 image patch around it as shown in Figure 2(b). The appearance, the offset to the image (object) center, and the label of the training image are encoded, and the code is inserted into a codebook. The final codebook contains 5502 codes.

For motion grouping, each keypoint is tracked through 10 frames before and through 10 frames after the current frame. The similarity of two 21-point trajectories is defined using only the overlapping part. To set the two thresholds for motion grouping, D_1 and D_2 are measured for keypoint pairs of different objects. D_{th}^1 is set that it is larger than only 10% of the measured D_1s, and so is D_{th}^2. By doing so, keypoints belonging to different objects are not likely to be grouped together. So, in one motion group, the keypoints are very likely to belong to the same object, as shown in Figure 3.

To form the Hough image, 35 best matched codes are chosen from the codebook for each object part. In (3), d and σ need to be given. The precision-recall curves are based on

Figure 3. Motion grouping results.



(a) (b) (c) (d)

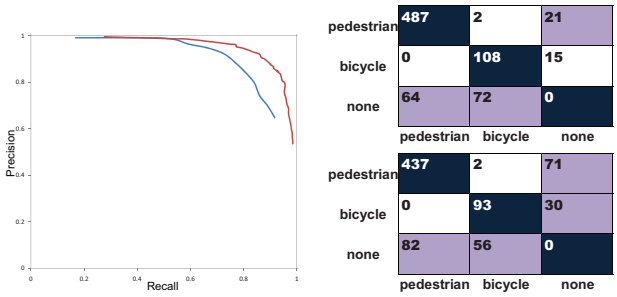
Figure 4. Example Hough images. Grids in (b), (c), and (d) correspond to the grids on the image coordinate. Red indicates pedestrians, while blue indicates bicycle riders. Hough image of (b) is formed by votes with uniform priors, Hough image of (c) is formed by votes with priors after 5 iterations, and Hough image in (d) is formed with converged priors.

σ, while d is set to be 10. Here σ is the most important parameter, while changing d from 5 to 16 only results in a 1% difference in the detection rate. For Algorithm 1, p_{th} is set to be 0, which means all local maxima found in Algorithm 1 are accepted.

Comparisons For comparison, concurrent detection and recognition are done on the Hough images formed with and without motion grouping results. The same codebook and the same parameter settings are used for forming and searching over both Hough images. The votes of each object part are assigned uniform priors in the benchmark method, while priors defined in (4) are assigned in the proposed method. Example Hough images are shown in Figure 4. With the noisy codebook, peaks of different objects on the original Hough image not only mix up in position but also in class label.

The precision-recall curves are shown in Figure 5(a). An object is considered as correctly detected only if the distance from the ground truth to it is less than 10 pixels. In Figure 5(a), the correctly detected but wrongly recognized objects are considered as true positives, aiming at verifying the detection ability of the proposed method.

The confusion matrices are given in Figure 5(b). For clarity of the comparisons, the proposed method is compared with the benchmark method when they have nearly equal number of false alarms. To evaluate the concurrent detection and recognition ability, a class of "none" to represent missed detections and false alarms is manually added. For example, in Figure 5(b), 487 pedestrian instances are correctly detected and recognized by the proposed method,



(a) (b)

Figure 5. Precision-recall curves (red: the proposed method, blue: the benchmark method) and confusion matrices (upper: the proposed method, down: the benchmark method). In (b), pedestrian and bicycle miss detections are considered as wrongly recognized to be none. Pedestrian or bicycle false alarms are considered as belonging to the none class while wrongly recognized as pedestrians or bicycles.

2 are wrongly recognized to be bicycle riders, and 21 are miss-detected.

The proposed method is limited by its relying on reliable motion information. As shown in Figure 6, failures occur when objects enter the scene.

5.2. Big "Cats" Detection and Recognition

Dataset To further verify the proposed method, a mini dataset is built upon leopards and tigers of the family Felidae. The dataset contains 6 video clips of 9 leopards and 4 tigers. The frame size is 640×480. Both the animals are in the side view.

Figure 6. Campus objects detection and recognition results. Red rectangles and blue rectangles mark correctly detected and recognized pedestrians and bicycle riders. Yellow rectangles mark missed detections. White rectangles mark correctly detected but not correctly recognized objects. Green rectangles mark false alarms. Black rectangles mark static objects, which are beyond the verification for the method.

Figure 7. Effect of the proposed prior. Red circles are voted center for leopards, while blue ones are voted centers for tigers. On the top are the motion grouping results. In the middle are the voted centers according to the best matched codes. On the bottom are the voted centers voted by votes with highest priors.

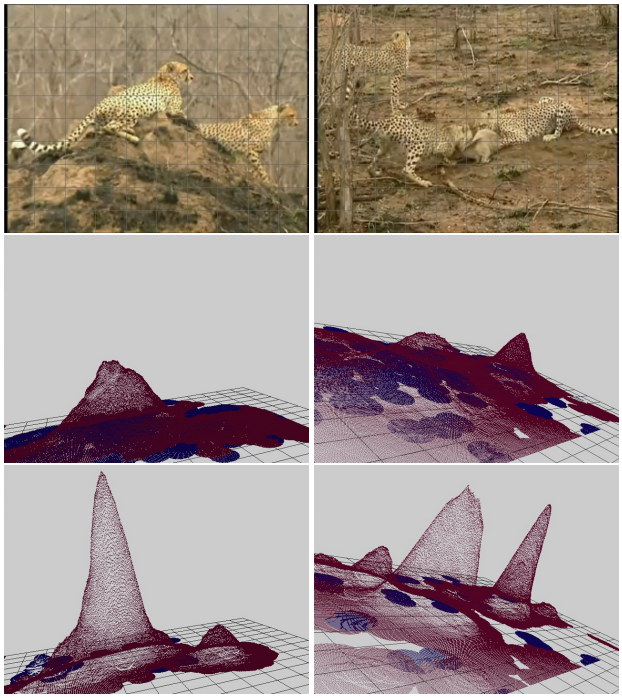


Figure 8. Example Hough images. On the top are the original images. In the middle are the Hough images formed by votes with uniform priors. On the bottom are the Hough images formed by votes with the proposed priors. Red indicates leopards, and blue indicates tigers. Note for the two leopards, there is no peak corresponding to the right one on the benchmark Hough image. For the three leopards, there is also no peak corresponding to the leopard in behind on the benchmark Hough image.

Implementation settings Most implementation settings are the same with the settings for campus object detection and recognition. For training, 5 leopards and 2 tigers are used. The size of the image patch around each keypoint is 27×27.

Comparisons In Figure 7, the motion grouping results

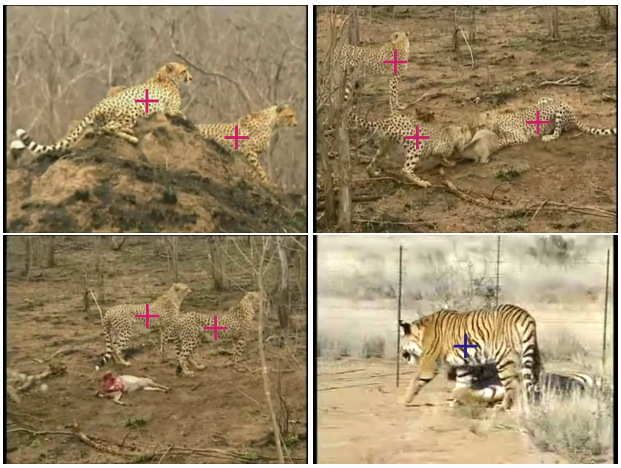


Figure 9. Big "cats" detection and recognition results. Red crosses mark the centers for leopards and blue crosses mark the centers for tigers.

and how the voted centers are affected are given. Since parts from different positions of the leopard are very similar, the true center of a leopard is difficult to find from the voted centers of the object parts. In Figure 8, example Hough images are given to show the merit of the proposed prior by the ability to detect leopards. In Figure 9, the detection and recognition results are given. The proposed method successfully detects and recognizes all the leopards and tigers, while the benchmark method miss-detects three leopards.

6. Discussion and Conclusion

This paper proposes a method for concurrent detection and recognition by extending the probabilistic Hough transform with motion. The underlying assumption is that object parts moving coherently are likely to belong to the same object. Before voting, all the object parts are grouped by their motion patterns. The votes of each object part are given different priors according to the motion grouping results. In this manner, the proposed method enhances the traceability of the Hough image formed. Experiments show the merit of the method in terms of detection and recognition accuracy.

References

[1] D. Ballard. Generalizing the hough transform to detect arbitrary shapes. *Pattern Recognition*, 13(2):111–122, 1981. 1

[2] O. Barinova, V. Lempitsky, and P. Kohli. On detection of multiple object instances using hough transforms. In *CVPR*, pages 2233–2240, 2010. 1, 4, 5

[3] G. Brostow and R. Cipolla. Unsupervised bayesian detection of independent motion in crowds. In *CVPR*, pages I: 594–601, 2006. 4

[4] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1):55–79, January 2005. 1

[5] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, pages II: 264–271, 2003. 1

[6] J. Gall and V. Lempitsky. Class-specific hough forests for object detection. In *CVPR*, pages 1022–1029, 2009. 1

[7] C. Gu, J. Lim, P. Arbelaez, and J. Malik. Recognition using regions. In *CVPR*, pages 1030–1037, 2009. 1

[8] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, pages 147–152, 1988. 4

[9] J. Knopp, M. Prasad, G. Willems, R. Timofte, and L. Van Gool. Hough transform and 3d surf for robust three dimensional classification. In *ECCV*, pages VI: 589–602, 2010. 1

[10] C. Lampert, M. Blaschko, and T. Hofmann. Beyond sliding windows: Object localization by efficient subwindow search. In *CVPR*, pages 1–8, 2008. 1

[11] B. Leibe, A. Leonardis, and B. Schiele. Robust object detection with interleaved categorization and segmentation. *IJCV*, 77(1-3):259–289, May 2008. 1

[12] B. Leibe and B. Schiele. Interleaved object categorization and segmentation. In *BMVC*, pages 759–768, 2003. 1

[13] S. Maji and J. Malik. Object detection using a max-margin hough transform. In *CVPR*, pages 1038–1045, 2009. 1

[14] S. Mohottala, S. Ono, M. Kagesawa, and K. Ikeuchi. Fusion of a camera and a laser range sensor for vehicle recognition. In *OTCBVS*, pages 16–23, 2009. 1

[15] K. Ohba and K. Ikeuchi. Detectability, uniqueness, and reliability of eigen windows for stable verification of partially occluded objects. *PAMI*, 19(9):1043–1047, September 1997. 1

[16] R. Okada. Discriminative generalized hough transform for object dectection. In *ICCV*, pages 2000–2005, 2009. 1

[17] M. Sun, G. Bradski, B. Xu, and S. Savarese. Depth-encoded hough voting for joint object detection and shape recovery. In *ECCV*, pages V: 658–671, 2010. 1

[18] C. Tomasi and T. Kanade. Detection and tracking of point features. Technical report, Carnegie Mellon University, April 1991. 4

[19] O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classification. In *ECCV*, pages II: 589–600, 2006. 4

[20] M. Wertheimer. Laws of organization in perceptual forms (partial translation). In W. B. Ellis, editor, *A Sourcebook of Gestalt Psychology*, pages 71–88. Harcourt, Brace, 1938. 2

[21] A. Yao, J. Gall, and L. Van Gool. A hough transform-based voting framework for action recognition. In *CVPR*, pages 2061–2068, 2010. 1

[22] T. Yeh, J. Lee, and T. Darrell. Fast concurrent object localization and recognition. In *CVPR*, pages 280–287, 2009. 1