Voting by Grouping Dependent Parts

Pradeep Yarlagadda, Antonio Monroy, and Björn Ommer

Interdisciplinary Center for Scientific Computing, University of Heidelberg, Germany
{pyarlaga,amonroy,bommer}@iwr.uni-heidelberg.de

Abstract. Hough voting methods efficiently handle the high complexity of multi-scale, category-level object detection in cluttered scenes. The primary weakness of this approach is however that mutually *dependent local* observations are *independently* voting for intrinsically *global* object properties such as object scale. All the votes are added up to obtain object hypotheses. The assumption is thus that object hypotheses are a sum of independent part votes. Popular representation schemes are, however, based on an overlapping sampling of semi-local image features with large spatial support (e.g. SIFT or geometric blur). Features are thus mutually dependent and we incorporate these dependences into probabilistic Hough voting by presenting an objective function that combines three intimately related problems: i) grouping of mutually dependent parts, ii) solving the correspondence problem conjointly for dependent parts, and iii) finding concerted object hypotheses using extended groups rather than based on local observations alone. Experiments successfully demonstrate that state-of-the-art Hough voting and even sliding windows are significantly improved by utilizing part dependences and jointly optimizing groups, correspondences, and votes.

1 Introduction

The two leading methods for detecting objects in cluttered scenes are voting approaches based on the Hough transform [19] and sliding windows (e.g. [33,12]). In the latter case, rectangular sub-regions of a query image are extracted at all locations and scales. A binary classifier is evaluated on each of these windows before applying post-processing such as non-max suppression to detect objects. The computational complexity of this procedure is critical although techniques such as interest point filtering, cascade schemes [33], or branch-and-bound [20] have been presented to address this issue. Rather than using a single, global descriptor for objects, Hough voting avoids the complexity issues by letting local parts vote for parametrized object hypotheses, e.g. object locations and scales. Generalizations of the Hough transform to arbitrary shapes, exemplar recognition [23], and category-level recognition [22,16,29,30,28,25,18] have successfully demonstrated the potential of this approach, and its wide applicability. Despite the current popularity of the method, Hough voting has two significant weaknesses that limit its performance: i) (semi-)local parts are *independently* casting their votes for the object hypothesis and ii) intrinsically global object properties such as object scale [28] have to be estimated locally. Consequently, current voting approaches to object detection, e.g. [22,16,25,18], are adding all local votes in a Hough accumulator and are, thus, assuming that *objects are a sum of their parts*. This assumption is against the fundamental conviction of Gestalt theory that the whole object is more than the sum of its

K. Daniilidis, P. Maragos, N. Paragios (Eds.): ECCV 2010, Part V, LNCS 6315, pp. 197–210, 2010.

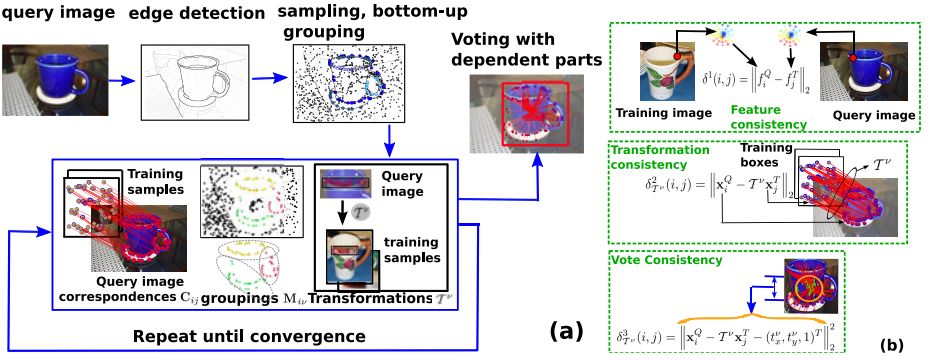

Fig. 1. a) Outline of the processing pipeline. b) The three terms of the cost function $d_{\mathcal{T}^\nu}$ from Eq. (7).

parts. And indeed, popular semi-local feature descriptors such as SIFT [23] or geometric blur [5] have a large spatial support so that different part descriptors in an image are overlapping and thus mutually dependent. To avoid missing critical image details, a recent trend has been to even increase sampling density which entails even more overlap. However, observing the same image region N times does not provide N independent estimates of the object hypothesis. Models with richer part dependencies (see section 2) such as constellation models [15] or pictorial structures [14] have been proposed to address these issues, however these methods are limited by their complexity (number of parts and the number of parameters per part). Without grouping, [5] transform a complete query image onto a training image. Therefore, this method is constrained to few distractors (e.g. little background clutter) and the presence of only one object in an image. In [16] Hough voting precedes the complex transformation of the complete object from [5] to limit the hypothesis space and reduce the influence of background clutter. However, the voting is limited by assuming independent part votes.

To establish reliable group votes, we incorporate dependencies between parts into Hough voting [22] by

- *grouping* mutually dependent parts,
- solving the *correspondence problem* (matching parts of the query image to model parts of training images) jointly for all dependent parts, thereby utilizing their information on each other,
- letting groups of dependent parts *vote* for concerted object hypotheses that all constituents of the group agree upon,
- integrating grouping, correspondence, and voting into a single objective function that is *jointly optimized*, since each subtask is depending on the remaining ones.

Outline of the Approach

Object detection in a novel image (c.f. Fig. 1) starts by first computing a probabilistic edge map (using [24]). A uniform sampling of edge pixels yields points where local features are extracted on a single scale (we use geometric blur features [5]). Each descriptor is mapped to similar features from training images. In standard Hough voting, all points

are then independently voting for an object hypothesis in scale space, i.e. object location and scale, before adding up all these votes in a Hough accumulator. Consequently, dependencies between points are disregarded and for each point, unreliable local estimates of global object properties such as object scale are required. To correctly model the dependencies between features, we group related points and estimate object hypotheses jointly for whole groups rather than independently for all of their constituents. This results in three intimately related problems: i) Grouping mutually dependent points, ii) letting groups of dependent points vote for a concerted object hypothesis, and iii) finding correspondences for each point in a group to training samples. We jointly find a solution to all of these three subtasks by formulating them in a single cost function and solving it using a single clustering algorithm. That way, all related points influence each others voting and correspondences and their voting influences their grouping, in turn. To obtain an initial grouping, we perform pairwise clustering of edge points. The necessary pairwise affinities are obtained by measuring the cooccurrence of points in different levels of the hierarchical segmentation of the initial probabilistic edge map from [24].

2 Voting Methods and Object Detection

Category-level object detection requires models that represent objects based on local measurements in an image. A broad variety of models with widely differing representation complexity have been proposed. These range from bag-of-features approaches [11] and latent topic models without spatial relationships [31] to richer spatial representations such as hierarchical models [7,17,2], k-fans [10], and latent scene models [32]. Complex spatial representations have been described by a joint model of all local parts (constellation model) [15], shape matching [5], pictorial structures [14], and by rigid template-like models [12,21]. The compositional nature of our visual world has been utilized by [27] to build hierarchical object representations.[26] describes a Tensor voting approach to form perceptually meaningful groups which can then be used for object recognition. The voting paradigm [22,16,28,25,18], which is central to this paper, effectively handles the complexity of large-scale part-based models.

2.1 Hough Voting with Independent Parts

Hough voting makes part-based object models with large numbers of parts feasible by letting all parts independently cast their votes for object hypotheses [22]. All these locally estimated object hypotheses are summed up in a Hough accumulator $\mathcal{H}^{\text{pnt}}(c, \mathbf{x}, \sigma)$ over scale space. Here, $\mathbf{x}$ and σ are the location and scale of an object hypothesis and c denotes its category. Moreover, a local part detected at location $\mathbf{x}_i^Q \in \mathbb{R}^2$ in a query image incorporates a feature vector $f_i^Q \in \mathbb{R}^N$ and a local estimate $\sigma_i^Q \in \mathbb{R}$ of object scale. The key assumption of Hough voting is that all parts are *independently* casting their votes for the object hypothesis so that the overall *object hypothesis is independently obtained from dependent parts*,

$$\mathcal{H}^{\text{pnt}}(c, \mathbf{x}, \sigma) \propto \sum_i p\big(\mathbf{x}, \sigma \big| c, f_i^Q, \mathbf{x}_i^Q, \sigma_i^Q\big) p\big(c \big| f_i^Q, \mathbf{x}_i^Q, \sigma_i^Q\big) \tag{1}$$

Let f_j^T denote the j-th codebook vector or the j-th training sample, depending on whether vector quantization or a nearest neighbor approach is used. Without loss of generality we can assume that the training object is centered at the origin so that the location $\mathbf{x}_j^T \in \mathbb{R}^2$ of f_j^T is the shift of the feature from the object center. Moreover, all training images are assumed to be scale normalized, i.e. they are rescaled so that objects are the same size. Summation over f_j^T and $\mathbf{x}_j^T$ then yields

$$\mathcal{H}^{\text{pnt}}(c, \mathbf{x}, \sigma) \propto \sum_{i,j} p\big(\mathbf{x} - [\mathbf{x}_i^Q - \sigma_i^Q \mathbf{x}_j^T], \sigma - \sigma_i^Q\big) \times p(c|f_j^T)\; p(f_j^T|f_i^Q) \tag{2}$$

Details of this derivation can be found in [22,28].

2.2 Key Points of Our Method

Hough voting methods (e.g. [22,16,28,25,18]) let all parts independently cast their votes for the object hypothesis, thereby neglecting part dependence. In contrast to this, our approach models the dependencies between parts by establishing groups and letting all parts in a group jointly find a concerted object hypothesis. In detail, we are differing from voting methods to detection in the following ways:

Grouping of Dependent Parts: Rather than considering all parts to provide independent votes (e.g. [22,16,28,25,18]), we segment a scene into groups of mutually dependent parts. Thus multiple strongly related features (e.g. due to overlapping descriptors) are not considered as providing independent information.

Joint Voting of Groups of Dependent Parts: Mutually dependent parts in a group assist each other in finding compatible correspondences and votes, rather than estimating these independently as in standard Hough voting. Thus groups yield votes with significantly less uncertainty than the individual part votes (c.f. Fig. 5). Intrinsically global parameters such as object scale are then obtained by global optimization rather than by local estimates (such as local scale estimation in [22,8]). [28] could only model the uncertainty of each local part. Based on a grouping of parts, we can however obtain reliable estimates.

Joint Optimization of Grouping, Voting, and Correspondences: Identifying and grouping dependent parts, computing joint votes for complete groups, and solving the part correspondence problem are mutually dependent problems of object detection. We tackle them jointly by iteratively optimizing a single objective function. Rather than letting each of these factors influence the others, [8] finds groups before using them to optimize correspondences in a model where parts are grouped with their k nearest neighbors. Estrada et al. [13] pursue the simpler problem of exemplar matching by only dealing with grouping and matching consecutively. Several extensions have been proposed to the standard Hough voting scheme, but the critical grouping of dependent parts has not been integrated into voting in any of those approaches. [29] extend the Implicit Shape Model by using curve fragments as parts that cast votes. Without incorporating a grouping stage into their voting, parts are still independently casting their votes. Amit et al. [3] propose a system limited to triplet groupings. In contrast to such rigid

groupings, our approach combines flexible numbers of parts based on their vote consistency and geometrical distortion. In contrast to hierarchical grouping approaches, where later groupings build on earlier ones, our method does not require any greedy decisions that would prematurely commit to groupings in earlier stages but rather optimizes all groupings at the same time.

Linear Number of Consistency Constraints: In contrast to Berg et al. [5] who need a quadratic number of consistency constraints between all pairs of parts, grouping reduces this to a linear number of constraints between parts and the group they belong to, see section 3.

Flexible Model vs. Rigid Template: Template-like descriptors such as HoG [12] or [21] have a rigid spatial layout that assumes objects to be box-shaped and non-articulated. Moreover, they require a computationally daunting search through hypothesis space although approximations such as branch-and-bound [20] have been proposed to deal with this issue. On the other end of the modeling spectrum are flexible parts-and-structure models [15,14]. However, the modeling of part dependencies in [15] becomes prohibitive for anything but very small number of points and [14] restrict the dependencies to a single, manually selected reference part. In contrast to this, we incorporate dependencies in the powerful yet very efficient Hough voting framework. Moreover, we do not rely on pixel accurate labeling of foreground regions as in [22] but only utilize bounding box annotations. In contrast to [16,5] who transform a query image onto training images using a complex, nonlinear transformation we decompose the object and the background into groups and transform these onto the training samples using individual, linear transformations. That way, unrelated regions do not interfere in a single, complex transformation and regions of related parts can be described by simpler and thus more robust, linear models.

3 Grouping, Voting, and Correspondences

Hough voting approaches to object detection let all local parts independently vote for a conjoint object hypothesis. However, there are direct mutual dependencies between features, e.g. due to their large spatial support and since interest point detection has a bias towards related regions in background clutter [6]. Thus, multiple related features yield dependent votes rather than independent evidence on the object. Rather than adding up all those duplicates as is common practice in Hough voting approaches (eg. [22,16,25,28]), a group of mutually dependent parts should actually jointly vote for a concerted object hypothesis. That way, the correspondence problem of matching features in a novel query image to features in training samples is jointly solved for a group of dependent parts.

3.1 Joint Objective Function for Grouping, Voting, and Correspondences

To solve the grouping, voting, and correspondence problem jointly, we have to i) match query features onto related training features, ii) find correspondences with low geometrical distortion, and iii) minimize the overall scatter of all votes within a group. Let us

now investigate each of these aspects in detail. Hough voting solves the correspondence problem by matching the i-th part of a query image, f_i^Q, to the training part or training codebook vector f_j^T that is most similar, i.e. for which

$$\delta^1(i,j) = \left\| f_i^Q - f_j^T \right\|_2 \tag{3}$$

is minimal. Boiman et al. [6] have demonstrated the deficiencies of quantization and codebook based representations. Therefore, we adopt a nearest neighbor approach, where query features are mapped onto training features rather than mapping them onto a quantized codebook. Let $\mathbf{C}_{ij} \in \{0,1\}$ denote a matching of the i-th query part to the j-th training part, where $\mathbf{C}_{ij}$ captures many-to-one-matchings, $\sum_j \mathbf{C}_{ij} = 1$. As discussed above, the correspondence problem has to be solved jointly for all mutually dependent parts, i.e. all related parts should undergo the same transformation $\mathcal{T}^\nu$ when being matched to the training samples, $\mathbf{x}_i^Q \stackrel{!}{=} \mathcal{T}^\nu \mathbf{x}_j^T$. This implies that related parts i and i' are clustered into the same group ν by computing assignments $\mathbf{M}_{i\nu}$ of parts to groups, $\mathbf{M}_{i\nu} \in \{0,1\}$,$\sum_\nu \mathbf{M}_{i\nu} = 1$.

Due to the relatedness of points in a group, transformations should be forced to be simple, eg. similarity transformations

$$\mathcal{T}^\nu = \begin{pmatrix} \sigma_x^\nu \cos(\theta) & -\sigma_y^\nu \sin(\theta) & t_x^\nu \\ \sigma_x^\nu \sin(\theta) & \sigma_y^\nu \cos(\theta) & t_y^\nu \\ 0 & 0 & 1 \end{pmatrix} \tag{4}$$

In effect, we are decomposing heterogeneous objects into groups of dependent parts so that piecewise linear transformations (one for each group) are sufficient rather than using a complex nonlinear transformation for the whole scene as in [5,16]. Let $G^\nu := \{i : \mathbf{M}_{i\nu} = 1\}$ denote all parts in a group ν and $|G^\nu| = \sum_i \mathbf{M}_{i\nu}$ denote the number of parts in the group. Then we have to find a transformation $\mathcal{T}^\nu$ that minimizes the distortion

$$\delta^2_{\mathcal{T}^\nu}(i,j) = \left\| \mathbf{x}_i^Q - \mathcal{T}^\nu \mathbf{x}_j^T \right\|_2 \tag{5}$$

for each part in the group.

(5) is penalizing the distortions of correspondences to yield minimal group distortion. The consistency of group votes is obtained by measuring the deviation of individual votes from the average vote of the group. Minimal group distortion does not necessarily guarantee consistent group votes. Hence we introduce a term that penalizes the scatter of the group vote.

$$\delta^3_{\mathcal{T}^\nu}(i,j) = \left\| \mathbf{x}_i^Q - \mathcal{T}^\nu \mathbf{x}_j^T - (t_x^\nu, t_y^\nu, 1)^T \right\|_2^2 \tag{6}$$

(6) is measuring the agreement of all parts in the group with respect to their object center estimate (summing over all parts i in a group yields the variance of the group vote).

This consistency constraint has a linear complexity in the number of image features in contrast to Berg et al. [5] who proposed pairwise consistency constraints with a quadratic complexity. This reduction in complexity is possible since dependent parts are combined in groups, so we can penalize the scatter of the entire group. Without the grouping, Berg et al. have to penalize the distortions of all pairs of parts under the transformation.

Joint Cost Function
Groupings $\mathbf{M}_{i\nu}$ of query parts, correspondences $\mathbf{C}_{ij}$ between query parts and training parts, and group transformations $\mathcal{T}^\nu$ are mutually dependent. Thus we have to combine them in a single cost function

$$d_{\mathcal{T}^\nu}(i,j) = \lambda_1 \delta^1(i,j) + \lambda_2 \delta^2_{\mathcal{T}^\nu}(i,j) + \lambda_3 \delta^3_{\mathcal{T}^\nu}(i,j) \tag{7}$$

that is jointly optimized for each of these unknowns. The weights $\lambda_1, \lambda_2, \lambda_3$ are adjusted by measuring the distribution of each distance term $\delta(.)$ in the training data. The weights are then set to standardize the dynamic range of each term to the same range. The cost for matching all the query parts i which belong to group ν to the corresponding training parts $j = \mathbf{C}(i)$ is given by

$$\mathcal{R}(G^\nu) = \frac{1}{|G^\nu|} \sum_i \mathbf{M}_{i\nu} \sum_j \mathbf{C}_{ij}\ d_{\mathcal{T}^\nu}(i,j) \tag{8}$$

3.2 Joint Optimization of Groups, Votes, and Correspondences

To find optimal groups, object votes, and correspondences, we need to minimize the overall cost of all groups $\sum_\nu \mathcal{R}(G^\nu)$. We seek optimal group assignments $\mathbf{M}^*$, correspondences $\mathbf{C}^*$, and transformations $\mathcal{T}^*$ that minimize the summation of costs over all the groups,

$$(\mathbf{M}^*, \mathbf{C}^*, \mathcal{T}^*) = \underset{\mathbf{M},\mathbf{C},\mathcal{T}}{\operatorname{argmin}} \sum_\nu \mathcal{R}(G^\nu)\ . \tag{9}$$

Since parts in a group are mutually dependent, each of these parameters depends on the other two. Therefore we incorporate an alternating optimization scheme. To find the optimal corresponding training part $j = \mathbf{C}(i)$ for query part i we have to minimize

$$\mathbf{C}(i) = \underset{j}{\operatorname{argmin}}\, d_{\mathcal{T}^\nu}(i,j)\ . \tag{10}$$

So for each i, we select the training part j with minimal cost. Optimal groupings are obtained by finding assignments $\nu = \mathbf{M}_{i\nu}(i)$ for each part i,

$$\mathbf{M}_{i\nu}(i) = \underset{\nu}{\operatorname{argmin}}\, d_{\mathcal{T}^\nu}(i, \mathbf{C}(i)) = \underset{\nu}{\operatorname{argmin}} \big[\lambda_2 \delta^2_{\mathcal{T}^\nu}(i, \mathbf{C}(i)) + \lambda_3 \delta^3_{\mathcal{T}^\nu}(i, \mathbf{C}(i))\big]. \tag{11}$$

Thus for each i, the group ν with minimal distortion is chosen. Finally, the transformation of each group from the query image onto the training images has to be estimated

$$\mathcal{T}^\nu = \underset{\mathcal{T}}{\operatorname{argmin}} \sum_i \mathbf{M}_{i\nu} \sum_j \mathbf{C}_{ij} \cdot \left[\lambda_2 \delta^2_{\mathcal{T}^\nu}(i, \mathbf{C}(i)) + \lambda_3 \delta^3_{\mathcal{T}^\nu}(i, \mathbf{C}(i))\right] . \tag{12}$$

Optimal $\mathcal{T}^\nu$ in (12) is obtained by Levenberg-Marquardt minimization. These three optimization steps are alternated until convergence. In our experiments, the optimization in Alg. 1 has usually converged after two or three iterations. We initialize ν by the output of a bottom-up grouping that is outlined in section 3.4. Initialization of $\mathbf{C}_{ij}$ for each query part i is obtained by a nearest neighbour search for j using the distance function $\delta^1(i, j)$. $\mathcal{T}^\nu$ is initialized with the transformation that aligns the centroid of group ν onto the centroid of the corresponding training parts.

3.3 Hough Voting with Groups

After finding optimal groupings, group transformations, and correspondences, the votes from all groups have to be combined. In standard Hough voting, the votes of all parts are summed up, thus treating them as being independent, c.f. the discussions in [34,1]. In our setting, all mutually dependent parts are combined in the same group. The joint optimization of correspondences and transformations forces these dependent parts to agree upon a joint overall vote.

$$(\mathbf{x}, \sigma)^\top = (\mathbf{x}_i^Q - \mathcal{T}^\nu \mathbf{x}_j^T \mathbf{C}(i) + t^\nu, \sigma^\nu)^\top \tag{13}$$

where t^ν and σ^ν are the translation and scaling component of $\mathcal{T}^\nu$. Evidently, all parts in a group are coupled by using the same transformation matrix $\mathcal{T}^\nu$ and the jointly optimized correspondences $\mathbf{C}_{ij}$. After jointly optimizing the votes of all dependent parts, the group vote can be obtained by averaging over the part votes. The Hough accumulator for the voting of groups is obtained by summing over independent groups rather than over dependent parts as in standard Hough voting. Since groups are mutually independent, their summation is justified. Analogous to (2) we obtain

$$\mathcal{H}^{\text{grp}}(c, \mathbf{x}, \sigma) \propto \sum_\nu \frac{1}{G^\nu \mathcal{R}(G^\nu)} \times \sum_{i \in G^\nu} \sum_j \mathbf{C}_{ij} \cdot P(\mathbf{x} - [\mathbf{x}_i^Q - \mathcal{T}^\nu \mathbf{x}_j^T + t^\nu], \sigma - \sigma^\nu) \tag{14}$$

where $P(\bullet)$ is obtained using the balloon density estimator [9] with Gaussian Kernel K, Kernel bandwidth b, and distance function in scale space $d : \mathbb{R}^3 \times \mathbb{R}^3 \mapsto \mathbb{R}$,

$$P(\mathbf{x} - [\mathbf{x}_i^Q - \mathcal{T}^\nu \mathbf{x}_j^T + t^\nu], \sigma - \sigma^\nu) = K\left(\frac{d\left[(\mathbf{x}, \sigma)^\top; (\mathbf{x}_i^Q - \mathcal{T}^\nu \mathbf{x}_j^T + t^\nu, \sigma^\nu)^\top\right]}{b(\sigma)}\right) \tag{15}$$

Algorithm 1. Voting with groups of dependent parts: Joint optimization of groupings, correspondences, and transformations.

Input: • parts from query image: f_i^Q, $\mathbf{x}_i^Q$,
• UCM-connectivity [4] $\bar{\mathbf{A}}_{ii'}$
• parts from all training images: f_j^T, $\mathbf{x}_j^T$
Init: • pairwise clustering on $\bar{\mathbf{A}}_{ii'} \rightarrow \mathbf{M}_{i\nu}()$
do
  $\mathbf{C}(i) \leftarrow \operatorname{argmin}_j d_{\mathcal{T}^\nu}(i,j)$
  $\mathbf{M}_{i\nu}(i) \leftarrow \operatorname{argmin}_\nu d_{\mathcal{T}^\nu}(i,\mathbf{C}(i))$
  $\mathcal{T}^\nu \leftarrow \operatorname{argmin}_{\mathcal{T}} \sum_i \mathbf{M}_{i\nu} \sum_j \mathbf{C}_{ij}(\lambda_2 \delta^2_{\mathcal{T}^\nu}(i,\mathbf{C}(i)) + \lambda_3 \delta^3_{\mathcal{T}^\nu}(i,\mathbf{C}(i)))$
  until convergence
$\mathcal{H}^{\text{grp}}(c,\mathbf{x},\sigma) \leftarrow$ Eq. (14)
$\left\{(\mathbf{x}^h,\sigma^h)^\top\right\}_h \leftarrow$ Local minima of $\mathcal{H}^{\text{grp}}$

3.4 Bottom-Up Grouping

Object detection in a query image starts by computing a probabilistic edge map [4] and uniformly sampling edge points. Next, we perform a bottom-up grouping on the probabilistic edges which serves as an initialization for ν in section 3.1. Two edge points i, i' are considered to be connected on level s of the hierarchical ultrametric contour map of [4], if they are on the boundary of the same region on this level. Let $1 = \mathbf{A}^s_{ii'} \in \{0,1\}$ denote this case. Averaging over all levels, $\bar{\mathbf{A}}_{ii'} \propto \sum_s \mathbf{A}^s_{ii'}$, yields a similarity measure between points and pairwise clustering (using Ward's method) on this similarity matrix produces a grouping $\mathbf{M}_{i\nu}$ which we use to initialize the optimization of (9).

3.5 Hypothesis Verification

Due to intra-class variations and noise, the votes of all parts in a group cannot be brought into perfect agreement. As is common practice in voting approaches, we employ a verification stage, where a SVM classifies histograms of oriented gradients (extracted on regular grids on 4 different resolutions and 9 orientations) using pyramid match kernels (PMK). To train the SVM, positive examples for a category are the groundtruth bounding boxes, rescaled to the average bounding box diagonal length of the class. Negative samples are obtained by running our group voting on the positive training samples and selecting false positive hypotheses, i.e. the most confused negative samples. In the verification stage, the SVM classifier is evaluated in a local 3×3 neighbourhood around each voting hypothesis. This local search refines the voting hypotheses from the groups.

4 Experiments

We evaluate our approach on ETHZ Shape and INRIA Horses Datasets. These two datasets feature significant scale changes, intra-class variation, multiple-objects per image, and intense background clutter. We use the latest experimental protocol of Ferrari et al. [16]: For ETHZ shape dataset, detectors are trained on half the positive samples of

a category. No negative training images are used and all remaining images are used for testing. For INRIA shape dataset, 50 horse images are used for training and the remaining 120 horse images plus 170 negative images are used for testing. In all experiments, the detection performance is measured using the PASCAL VOC criterion [16] (requiring the ratio of intersection and union of predicted and groundtruth bounding box to be greater than $.5$).

4.1 ETHZ Shape Dataset – Performance Analysis

Fig. 2 compares our approach with state-of-the-art voting methods on ETHZ. Voting with our groups of dependent parts outperforms all current voting based approaches. We achieve a gain of 27% over the Hough voting in [16], an improvement of 19% over [25], and 17% higher performance than [28], see Tab. 1. Even compared with the local sliding window classification in [28] (PMK re-ranking) we obtain a slightly higher performance (1.4%). The PMK re-ranking is a separate classifier that performs verification of votes. Thus our voting method alone not only improves current Hough voting approaches, but also produces results beyond those of the verification stage of some of the methods.

The primary focus of this paper is to improve Hough voting by modeling part dependence. Nevertheless, we also investigate the combined detector consisting of voting and a verification stage. The results are shown in Fig. 2. Our results compare favourably with sliding window classification in [28]. This approach has to search over 10^4 hypotheses whereas our approach produces on the order of 10 candidate hypotheses. Consequently, the gain in computational performance of our approach is between two and three orders of magnitude. Compared to preprocessing steps such as extraction of probabilistic edge maps and computation of geometric blur, our grouping, voting and correspondence optimization has insignificant running time. Nevertheless, we obtain a gain of 3.68% over sliding windows at 0.3 fppi. Compared to the best verification systems [25], we obtain a gain of 0.68% at 0.3 fppi.

Fig. 3 compares the supervised methods of [35] against our detector (which only needs training images with bounding boxes). Without requiring the supervision information of [35], we are dealing with a significantly harder task. [16] showed a performance loss of 15% at 0.4 fppi.Nevertheless, we perform better on 3 out of 5 categories. (actual values of [35] are unavailable).

Let us now compare the reliability of votes from individual parts with the reliability of object hypotheses produced by our groupings. Therefore, we map object query features (features from within the groundtruth bounding box) onto the positive training samples and we do the same for background query features. By comparing the matching costs we see how likely positive query features are mistaken to be background and vice versa. Then we are doing the same for groups, i.e. groupings (11) from the object and from the background are mapped onto positive training samples. Fig. 5 shows that groups have a significantly lower error rate $\mathcal{R}$ (30% vs. 77%) to be mapped onto wrong training samples. Thus group votes are significantly more reliable. Fig. 4 shows the voting of parts before and after optimization. Voting with groups produces concerted votes whereas independent parts(singleton groups) produce votes with significant clutter.

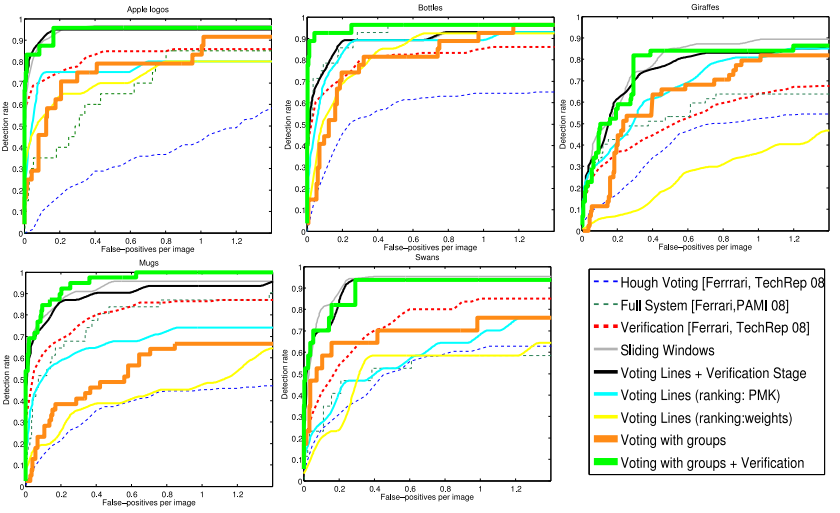

Fig. 2. Detection performance. On average our voting approach yields a 27% higher performance than standard Hough voting and improves line voting [28] by 17%.

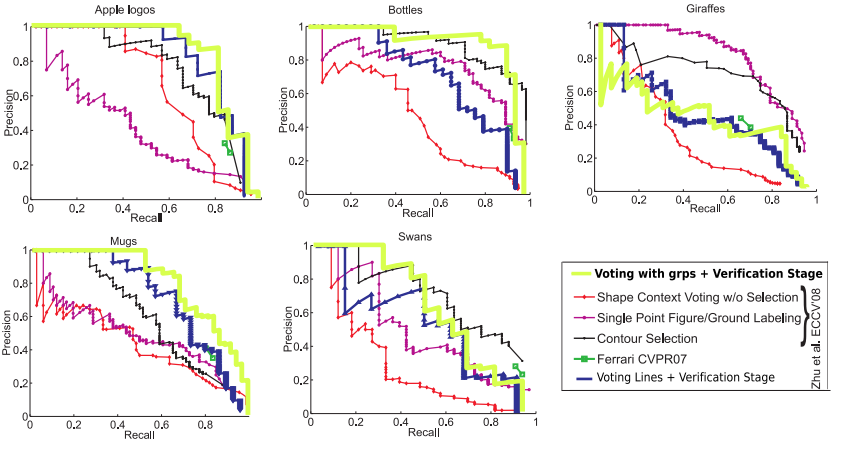

Fig. 3. Comparing our voting+verification with the supervised approach [35]. [16] has shown that our training scenario is significantly harder and yields 13% lower recall at .4 FPPI

4.2 INRIA Horse Dataset – Performance Analysis

Figure Fig. 6 shows the performance of voting with groups and the overall detector (voting + verification). Voting with groups significantly outperforms the best voting methods so far (M^2HT detector), e.g., roughly 12% gain at 3 fppi. In terms of overall performance, we have a detection rate of 87.3% at 1 fppi compared to the state of the art results of 85.27% for M^2HT + IKSVM and 86% for sliding windows (IKSVM).

Table 1. Comparing the performance of various methods. Detection rates (in [%]), PASCAL criterion .5 overlap. The approach of [25] use positive as well as negative samples for training whereas we use only positive samples for training. Our voting yields a 27% higher performance than the Hough voting in [16], 19% gain over max-margin Hough voting [25], and 17% gain over line voting [28], thus significantly improving the state-of-the-art in voting.

	Voting Stage (FPPI = 1.0)				*Verification Stage (FPPI = 0.3 / 0.4)*				
Cat	$\mathcal{H}^{\mathbf{grp}}$	Hough [16]	M^{2}HT [25]	voting [28]	$\mathcal{H}^{\mathbf{grp}}$ **voting+verif**	Full system [28]	Sliding Windows	Full syst [16]	M^{2}HT+ IKSVM [25]
Apples	*84.0*	43.0	85.0	80.0	*95.83 / 95.83*	95.0 / 95.0	95.8 / 96.6	77.7 / 83.2	95.0 / 95.0
Bottles	*93.1*	64.4	67.0	92.4	*96.3 / 96.3*	89.3 / 89.3	89.3 / 89.3	79.8 / 81.6	92.9 / 96.4
Giraffes	*79.5*	52.2	55.0	36.2	*81.82 / 84.09*	70.5 / 75.4	73.9 / 77.3	39.9 / 44.5	89.6 / 89.6
Mugs	*67.0*	45.1	55.0	47.5	*94.87 / 96.44*	87.3 / 90.3	91.0 / 91.8	75.1 / 80.0	93.6 / 96.7
Swans	*76.6*	62.0	42.5	58.8	*94.12 / 94.12*	94.1 / 94.1	94.8 / 95.7	63.2 / 70.5	88.2 / 88.2
Avg	*80.0*	53.3	60.9	63.0	*92.58 / 93.35*	87.2 / 88.8	88.9 / 90.1	67.2 / 72.0	91.9 / 93.2

Fig. 4. Left plot in panels (a) and (b) shows standard Hough voting which assumes mutual independence between features. Right plot in panels (a) and (b) shows the voting after joint optimization of correspondences, groups, and votes.

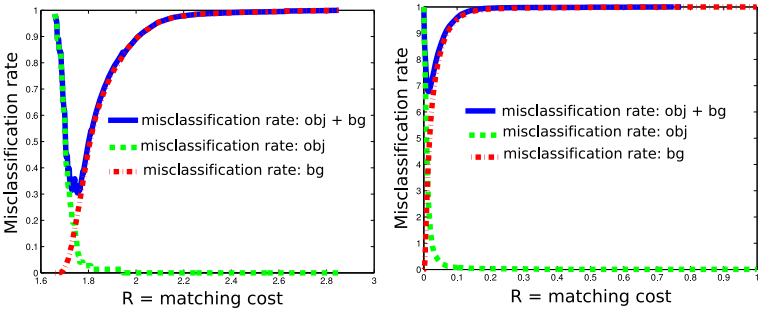

Fig. 5. Reliability of parts (singleton groups), left plot vs. groups, right plot. The plots show the misclassification rate of groups and parts for different matching cost $\mathcal{R}$. The optimal error rate for parts is 77%, for groups 30% thereby underlining the increased reliability of groups.

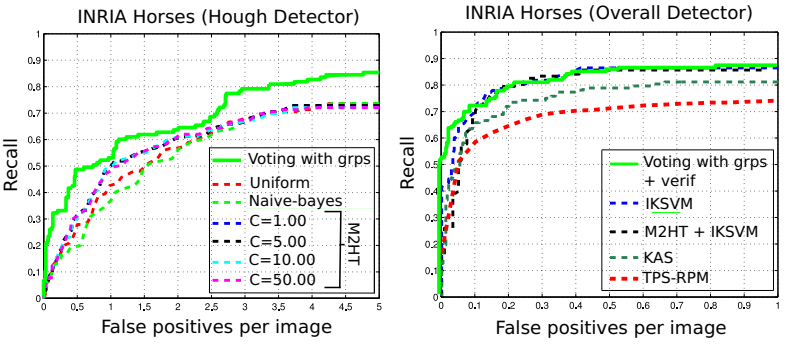

Fig. 6. Detection plots on INRIA Horses dataset. Left plot compares the M^2HT detector for different parameters with our group voting. Voting with groups is superior to all. Right plot compares the overall detection results obtained from voting with groups plus verification with sliding windows (IKSVM) and state-of-the-art methods. At 1.0 FPPI we achieve a detection rate of 87.3% compared to the state of the art result of 86% (IKSVM) [25]

5 Discussion

We have tackled the primary weakness of Hough voting methods, the assumption of part independence, by introducing the grouping of mutually dependent parts into the voting procedure. Therefore, we have formulated voting-based object detection as an optimization problem that jointly optimizes groupings of dependent parts, correspondences between parts and object models, and votes from groups to object hypotheses. Rather than using uncertain local votes from unreliable local parts we utilize their dependences to establish extended groups that reliably predict global object properties and are thus producing reliable object hypotheses. Compared to the sliding window paradigm, our voting approach reduces the number of candidate hypotheses by three orders of magnitude and improves its recall. Our model of part dependence in voting has demonstrated that it significantly improves the performance of probabilistic Hough voting in object detection.

Acknowledgements. This work was supported by the Excellence Initiative of the German Federal Government, DFG project number ZUK 49/1.

References

1. Lehmann, B.L.A., van Gool, L.: Prism principled implicit shape model. In: BMVC (2008)
2. Ahuja, N., Todorovic, S.: Connected segmentation tree: A joint representation of region layout and hierarchy. In: CVPR (2008)
3. Amit, Y., Geman, D.: A computational model for visual selection. Neural Computation (1999)
4. Arbelaez, P., Maire, M., Fowlkes, C., Malik, J.: From contours to regions: An empirical evaluation. In: CVPR (2009)
5. Berg, A.C., Berg, T.L., Malik, J.: Shape matching and object recognition using low distortion correspondence. In: CVPR, pp. 26–33 (2005)
6. Boiman, O., Shechtman, E., Irani, M.: In defense of nearest-neighbor based image classification. In: CVPR (2008)
7. Bouchard, G., Triggs, B.: Hierarchical part-based visual object categorization. In: CVPR, pp. 710–715 (2005)

8. Carneiro, G., Lowe, D.: Sparse flexible models of local features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3953, pp. 29–43. Springer, Heidelberg (2006)
9. Comaniciu, D., Ramesh, V., Meer, P.: The variable bandwidth mean shift and data-driven scale selection. In: ICCV, pp. 438–445 (2001)
10. Crandall, D.J., Felzenszwalb, P.F., Huttenlocher, D.P.: Spatial priors for part-based recognition using statistical models. In: CVPR, pp. 10–17 (2005)
11. Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: ECCV, Workshop Stat. Learn. in Comp. Vis. (2004)
12. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR, pp. 886–893 (2005)
13. Estrada, F.J., Fua, P., Lepetit, V., Susstrunk, S.: Appearance-based keypoint clustering. In: CVPR (2009)
14. Felzenszwalb, P.F., Huttenlocher, D.P.: Pictorial structures for object recognition. IJCV 61(1) (2005)
15. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In: CVPR, pp. 264–271 (2003)
16. Ferrari, V., Jurie, F., Schmid, C.: From images to shape models for object detection. IJCV (2009)
17. Fidler, S., Boben, M., Leonardis, A.: Similarity-based cross-layered hierarchical representation for object categorization. In: CVPR (2008)
18. Gall, J., Lempitsky, V.: Class-specific hough forests for object detection. In: CVPR (2009)
19. Hough, P.: Method and means for recognizing complex patterns. U.S. Patent 3069654 (1962)
20. Lampert, C.H., Blaschko, M.B., Hofmann, T.: Beyond sliding windows: Object localization by efficient subwindow search. In: CVPR (2008)
21. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR, pp. 2169–2178 (2006)
22. Leibe, B., Leonardis, A., Schiele, B.: Robust object detection with interleaved categorization and segmentation. IJCV 77(1-3), 259–289 (2008)
23. Lowe, D.: Object recognition from local scale-invariant features. In: ICCV (1999)
24. Maire, M., Arbelaez, P., Fowlkes, C., Malik, J.: Using contours to detect and localize junctions in natural images. In: CVPR (2008)
25. Maji, S., Malik, J.: Object detection using a max-margin hough transform. In: CVPR (2009)
26. Medioni, G., Tang, C., Lee, M.: Tensor voting: Theory and applications. In: RFIA (2000)
27. Ommer, B., Buhmann, J.: Learning the compositional nature of visual object categories for recognition. PAMI 32(3), 501–516 (2010)
28. Ommer, B., Malik, J.: Multi-scale object detection by clustering lines. In: ICCV (2009)
29. Opelt, A., Pinz, A., Zisserman, A.: Incremental learning of object detectors using a visual shape alphabet. In: CVPR, pp. 3–10 (2006)
30. Shotton, J., Blake, A., Cipolla, R.: Contour-based learning for object detection. In: ICCV (2005)
31. Sivic, J., Russell, B.C., Efros, A.A., Zisserman, A., Freeman, W.T.: Discovering objects and their localization in images. In: ICCV, pp. 370–377 (2005)
32. Sudderth, E.B., Torralba, A.B., Freeman, W.T., Willsky, A.S.: Learning hierarchical models of scenes, objects, and parts. In: ICCV, pp. 1331–1338 (2005)
33. Viola, P.A., Jones, M.J.: Robust real-time face detection. IJCV 57(2), 137–154 (2004)
34. Williams, C., Allan, M.: On a connection between object localization with a generative template of features and pose-space prediction methods. Technical report, University of Edinburg, Edinburg (2006)
35. Zhu, Q.H., Wang, L.M., Wu, Y., Shi, J.B.: Contour context selection for object detection: A set-to-set contour matching approach. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 774–787. Springer, Heidelberg (2008)