

A Shape Reconstructability Measure of Object Part Importance with Applications to Object Detection and Localization

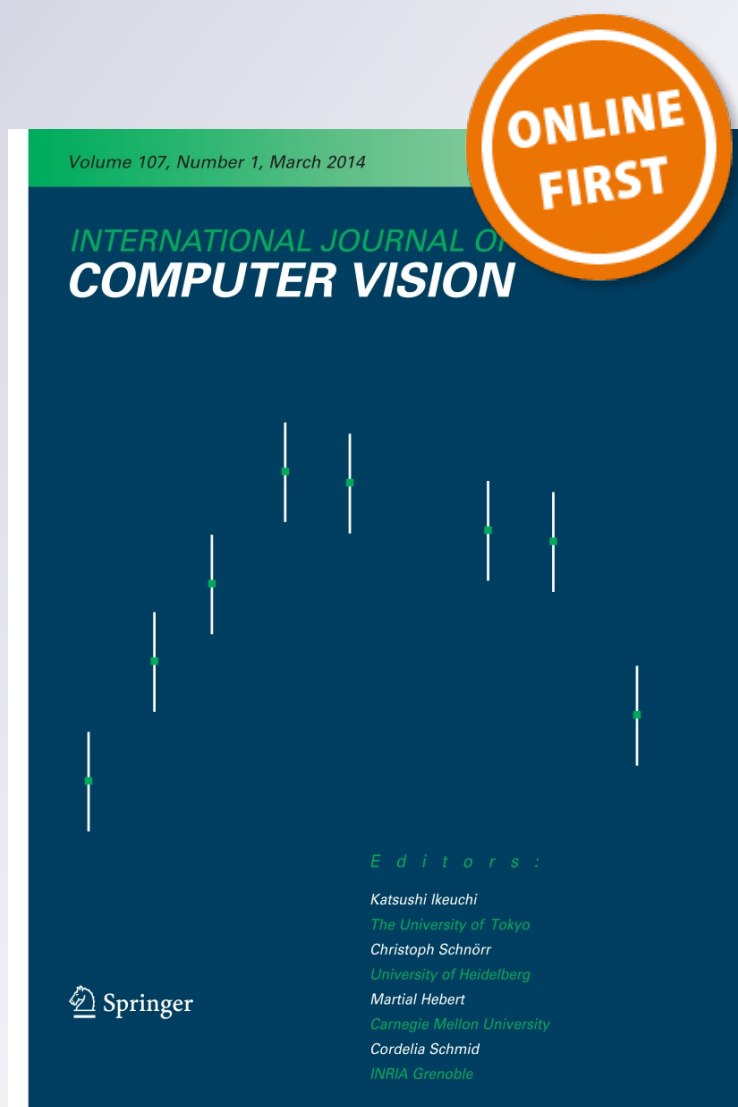
**Ge Guo, Yizhou Wang, Tingting Jiang,
Alan L. Yuille, Fang Fang & Wen Gao**

**International Journal of Computer
Vision**

ISSN 0920-5691

Int J Comput Vis

DOI 10.1007/s11263-014-0705-9



Your article is protected by copyright and all rights are held exclusively by Springer Science +Business Media New York. This e-offprint is for personal use only and shall not be self-archived in electronic repositories. If you wish to self-archive your article, please use the accepted manuscript version for posting on your own website. You may further deposit the accepted manuscript version in any repository, provided it is only made publicly available 12 months after official publication or later and provided acknowledgement is given to the original source of publication and a link is inserted to the published article on Springer's website. The link must be accompanied by the following text: "The final publication is available at link.springer.com".

A Shape Reconstructability Measure of Object Part Importance with Applications to Object Detection and Localization

Ge Guo · Yizhou Wang · Tingting Jiang · Alan L. Yuille · Fang Fang · Wen Gao

Received: 25 March 2013 / Accepted: 18 February 2014
© Springer Science+Business Media New York 2014

Abstract We propose a computational model which computes the importance of 2-D object shape parts, and we apply it to detect and localize objects with and without occlusions. The importance of a shape part (a localized contour fragment) is considered from the perspective of its contribution to the perception and recognition of the global shape of the object. Accordingly, the part importance measure is defined based on the ability to estimate/recall the global shapes of objects from the local part, namely the part's "shape reconstructability". More precisely, the shape reconstructability of a part is determined by two factors—part variation and part uniqueness. (i) Part variation measures the precision of the global shape reconstruction, i.e. the consistency of the reconstructed global shape with the true object shape; and (ii) part unique-

ness quantifies the ambiguity of matching the part to the object, i.e. taking into account that the part could be matched to the object at several different locations. Taking both these factors into consideration, an information theoretic formulation is proposed to measure part importance by the conditional entropy of the reconstruction of the object shape from the part. Experimental results demonstrate the benefit with the proposed part importance in object detection, including the improvement of detection rate, localization accuracy, and detection efficiency. By comparing with other state-of-the-art object detectors in a challenging but common scenario, object detection with occlusions, we show a considerable improvement using the proposed importance measure, with the detection rate increased over 10 %. On a subset of the challenging PASCAL dataset, the Interpolated Average Precision (as used in the PASCAL VOC challenge) is improved by 4–8 %. Moreover, we perform a psychological experiment which provides evidence suggesting that humans use a similar measure for part importance when perceiving and recognizing shapes.

Communicated by M. Hebert.

G. Guo · Y. Wang (✉) · T. Jiang · W. Gao
Nat'l Engineering Lab for Video Technology, Key Lab. of Machine Perception (MoE), Peking University, Beijing, China
e-mail: Yizhou.Wang@pku.edu.cn

G. Guo
e-mail: ggao@jdl.ac.cn

T. Jiang
e-mail: ttjiang@pku.edu.cn

W. Gao
e-mail: wgao@pku.edu.cn

A. Yuille
Department of Statistics, University of California, Los Angeles (UCLA), Los Angeles, CA, USA
e-mail: yuille@stat.ucla.edu

F. Fang
Department of Psychology and Key Lab. of Machine Perception (MoE), Peking-Tsinghua Center for Life Sciences, PKU-IDG/McGovern Institute for Brain Research, Peking University, Beijing, China
e-mail: ffang@pku.edu.cn

Keywords Shape part · Part importance · Shape reconstruction · Object recognition and detection

1 Introduction

Many convincing psychological evidences suggested that object parts play a significant role in object perception and recognition e.g. (Hoffman and Richards 1984; Siddiqi BK and Tresness 1996; Biederman 1987; Biederman and Cooper 1991). These research results motivated a lot of studies on part-based object representation e.g. (Epshtein and Ullman 2007; Dubinskiy and Zhu 2003; Gopalan et al. 2010), and part-based models have been widely used and con-

firmed to be successful in many computer vision applications, such as object recognition and classification e.g. (Zhu et al. 2010; Crandall and Huttenlocher 2006; Mikolajczyk et al. 2004).

Part models can be generally classified into two major categories, the appearance-based e.g. (Bouchard and Triggs 2005; Felzenszwalb and Huttenlocher 2005; Schneiderman and Kanade 2004) and shape-based models e.g. (Shotton et al. 2008; Opelt et al. 2008; Sala and Dickinson 2010). In this paper we study the shape-based part models, in particular *2D contour parts*. Although 2D shape and shape part models have attracted much attention and made great progress in a broad range of fields, such as psychology, neuroscience and computer vision, there are still some questions which deserve further study, in particular:

“Are shape parts of equal importance to a certain visual task? And how to quantitatively measure the importance of different parts?”

To pursue the answers to these questions, we investigate the role that parts play in the tasks of shape perception and object recognition. Psychological studies have discovered that shapes are perceived to be generated in terms of their constituent parts (Hubel and Wiesel 1962) and different shape parts provide different retrieval cues for shape perception; their abilities to recall object contours are quite different (Bower and Glass 2011). In addition, even with a partial shape (for example, due to occlusion), the human visual system has a powerful reconstruction ability to complete the global shape (Rensink and Enns 1998). Biederman (Biederman 1987) further demonstrated that object recognition is implemented by the Principle of Componential Recovery, i.e. objects can be quickly recognized by certain parts; and the parts *prime* (facilitate or speed up) the recognition process.

This motivates us to propose a mathematical model of part importance from a new perspective – a part's *shape reconstructability*, i.e. the ability of a part to recall, or recover, the global object shape. This measure of part importance is applied to a range of vision problems in object detection and representation. In particular, it offers an approach to the unsolved problem of *identifying partially occluded objects*.

In order to compute the *shape reconstructability* of a part, we propose an efficient shape reconstruction algorithm from a local 2D contour fragment under the Bayesian framework. The part provides a local observation (which gives a partial constraint on the global shape) and a class-specific shape model is learned as a prior model (i.e. a global constraint) which can be combined to estimate the global shape.

The shape reconstructability of a part is determined by two factors, the shape variation of the part and the uniqueness of the part with respect to the other parts of the object class.

(i) Part variation decides the reconstruction quality, i.e. the consistency of a recovered global shape with the object shape model. As shown in Fig. 1a and 2a, the heads of swans have less variation compared with the tails. When using a head part to recover a whole swan shape, we obtain a much better reconstruction than that when using a tail part (as shown in Fig. 1b). (ii) The part uniqueness, i.e. the ambiguity of matching the part to the object contour, which is higher if the part can be matched at several different locations along the object contour. This factor determines the *uncertainty of the shape reconstruction* from this part. For example, as shown in Fig. 1a, the flat part is much less unique compared with the head part. In consequence, it generates a greater number of high-quality reconstructions at different matching locations along the object contour compared with the head part (Fig. 1c). This leads to a larger reconstruction uncertainty. In summary, a shape reconstruction from a part with higher quality and less uncertainty suggests that the part is more important. Therefore, both factors are embodied by a conditional entropy formulation, and the part importance measure is defined accordingly.

We do extensive experiments on object detection/ recognition which demonstrate the advantage of integrating the proposed part importance measure into current approaches to object detection. In the voting-based object detection framework, part importance is used to weigh the votes for object candidates. For object candidate verification, we use part importance in two ways. Firstly, in backward shape matching of localizing object boundaries (as in (Ferrari et al. 2009)), the learned part importance is used to infer matching correspondences of shape parts, as well as to weigh the matching costs (e.g., matching distance and non-correspondence penalties). Secondly, we train the weights of an SVM classifier e.g. (Riemenschneider et al. 2010)), using an “importance kernel” whose design is based on part importance. Experimental results show that both methods using part importance improve the object detection rate and localization accuracy. Especially on a subset of the challenging PASCAL dataset, the Interpolated Average Precision (Everingham et al. 2010) (as used in the PASCAL VOC challenge) is improved by 4 ~ 8 %.

In particular, there is a considerable improvement of the detection rate, over 10 %, when using part importance in a particularly challenging real scenario – object detection with occlusion. We test on two types of occlusion datasets. The first contains images with naturally occluded objects collected from the Internet. The second is an artificial occlusion dataset obtained by adding occluding masks to images from conventional object detection datasets. This artificial dataset is used to evaluate the model performance under controlled amounts of occlusion conditions.

In addition, we show that using part importance for object detection helps improve computational efficiency. One can

Fig. 1 Shape reconstruction from parts. (a) Three part instances (in black) and their matching locations (in blue) on an object shape (the green contours). (b) The reconstructed shapes (in pink) from matching the black part instances in (a) to the corresponding blue segments on a swan shape. It uses this local matching as local constraint and the learned object model as global prior. (c) Reconstruction scores at different matching locations. The peaks are pointed to at which matching locations the shapes are reconstructed (Color figure online)

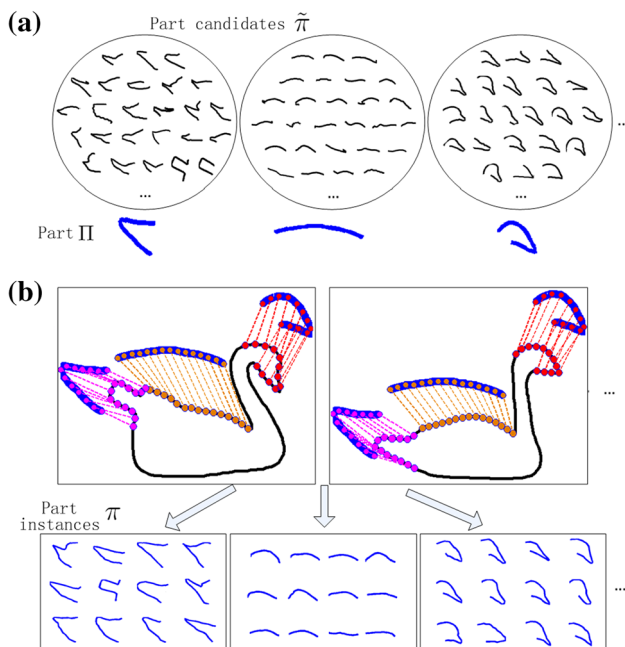
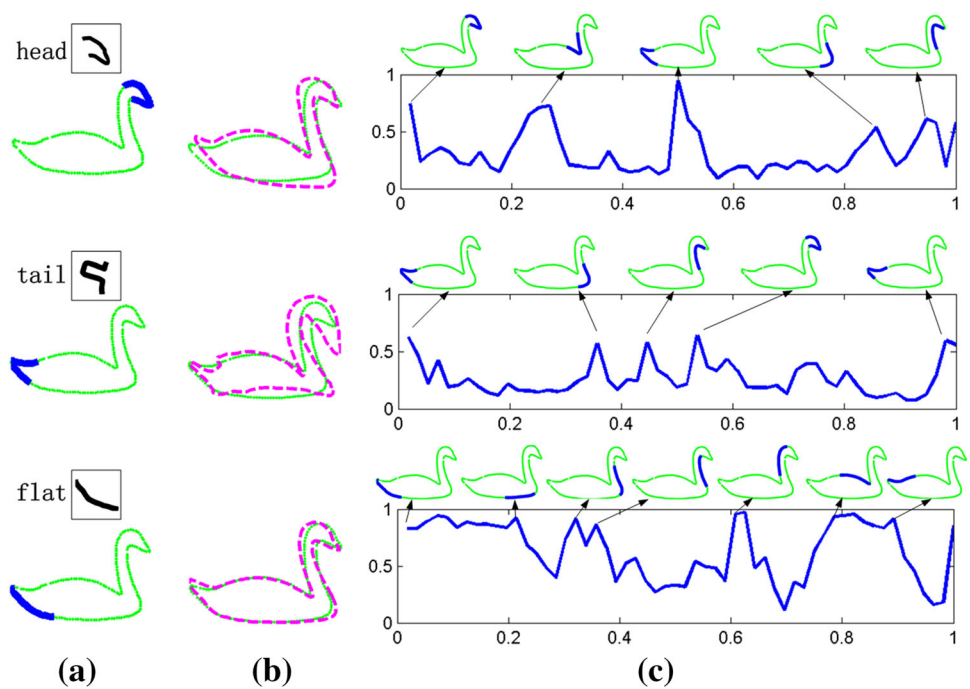


Fig. 2 Learning parts and locating part instances on objects. (a) Part candidates (in black) are extracted from aligned and normalized object contours. They are clustered into canonical parts (in bold blue). (b) Matching canonical parts to object instances (black contours) in order to extract part instances. The dots in different colors denote the contour points of different parts, and the dotted lines indicate the matching. The bottom rectangles contain examples of extracted part instances (Color figure online)

select a set of important parts to accelerate the candidate voting process and reduce the computational cost without losing much detection accuracy.

Finally, we perform a psychological experiment to evaluate whether humans use the part importance measure. Our experiment is inspired by previous studies (Biederman 1987). A contour part is shown before the complete object contour is displayed. The presentation of the contour part is used to prime the shape recognition process. The response time of recognition is used to measure part importance of the human subjects. It is demonstrated that our method for calculating part importance are much more consistent with the human data than previous computational models.

This paper stresses the importance of parts within one object class. For example, we compare the importance of the swan neck *versus* the swan back. In other words, we consider the contributions of parts to shape reconstruction/perception/recognition of their own category, rather than trying to distinguish between different object classes. By contrast, the discriminative learning approaches much studied in the computer vision literature, reviewed in Sect. 1.1.3, always involve more than one object class, and hence can be heavily affected by the training datasets. For instance, the importance of parts for differentiating between a zebra and a giraffe can be quite different from that for differentiating between a zebra and a car. Thus, the importance of parts for discrimination will depend on the set of object classes being considered. By contrast, our approach is based on the perspective of single object classes and so it is less sensitive to variations on the dataset. We argue that this is important for dealing with practical computer vision applications such as object recognition/detection for a very large number of classes.

1.1 Related Work

1.1.1 Contour-Based Object Recognition

Shape/contour-based object detection is a classical problem and remains a very active research topic in computer vision.

In most of state-of-the-art methods, one of the key issues is to study effective shape features/representations for object recognition. In the literature, there have been many popular shape features and representations, e.g. shape context (Belongie et al. 2002), the k -Adjacent-Segment (k AS) features (Ferrari et al. 2008) and the medial axis based representations (Sharvit et al. 1998; Bai and Latecki 2008). A part-based shape model of k -segment groups was proposed in (Ravishankar et al. 2008), in which the curve segments were generated by cutting at high curvature points. In (Shotton et al. 2008) and (Opelt et al. 2008) the authors learned class-specific codebooks of local contour fragments as the part-based representation of objects. In (Luo et al. 2010) contour segments were quantized with three types of distance metrics (procrustes, articulation and geodesic distance metrics), and spanned into a number of part manifolds. A user-defined vocabulary of simple part models was proposed in (Sala and Dickinson 2010) to group and abstract object contours in the presence of noise and within-class shape variation. (Lin et al. 2012) employed shape structure learning based on the and-or Tree representation. (Wang et al. 2012) suggests a Fan shape model in which contour points were modeled as flexible rays or slats from a reference point. In (Yarlagadda and Ommer 2012), the codebook contours were generated by clustering based on the contour co-activation (considering both the contour similarity and the matching locations), and then the co-placements of all the codebook contours were learned by max-margin multiple instance learning to obtain a discriminative object shape model.

Another crucial issue in contour-related methods concerns developing efficient shape matching algorithms. A set-to-set matching strategy is adopted in (Zhu et al. 2008) to utilize shape features with large spatial extent and capture long-range contextual information. Also, a many-to-one contour matching from image contours to object model was proposed in (Srinivasan et al. 2010). Partial shape matching is especially important in real and clutter images. In (Riemenschneider et al. 2010) an efficient partial matching schema was introduced, using a new shape descriptor of the chord angles. The method was further improved in (Ma and Latecki 2011) by the developed shape descriptor and the maximum clique inference to group the partial matching hypotheses.

Different detection framework have been suggested as well. One of the most popular strategies is the voting-based methods, such as in (Ommer and Malik 2009; Yarlagadda et al. 2010; Ferrari et al. 2009). Some others proposed to solve

the detection problem by contour grouping, e.g. (Lu et al. 2009).

In the literature, there has been limited research concerning the roles that different contour parts play in object detection. One notable exception is the work (Maji and Malik 2009) which proposed a discriminative Hough transform for object detection in which the a set of importance weights are learned in a max-margin framework. However, one of the main differences of our work compared with (Maji and Malik 2009) was that our generative part importance derived from the shape perception viewpoint, focused on the parts' different contributions to the class-specific object model, rather than their discriminative abilities across different classes. Besides, the advantages of our model include that, the proposed generative part importance is more robust to the variation of training data sets. Hence, the generative part importance has greater generalization ability. In addition, it is favored by psychological experimental results, which is shown in Sect. 5.

1.1.2 Toward Shape Part Importance

Although previous researches have made some progress toward shape part importance, the computational aspect of shape part importance is still understudied.

Some measures were proposed based on simple local geometric characteristics of shapes. For example, the curvature variation measure (CVM) for 3D shape parts (Sukumar et al. 2006), in which the entropy of surface curvature was proposed, and parts with large entropy were considered informative. Another measure based on the edgelet orientation distributions was suggested in (Renninger et al. 2007) to model the information of each location along the 2D shape contours. The edge orientations (discretized in eight orientations) of a local part were computed, the histogram (or probability distribution after normalization) of different edge orientations was calculated, and the entropy of the edge orientation distribution was adopted to measure the informativeness of this location. Nevertheless, the measures based on local characteristics might have some limitations, e. g., the repetitive parts might not be important even if they were of high entropy of local curvature variation or edgelet orientation distributions. The limitations were greatly improved in our model which embodies both the local shape variation and the uniqueness of a part.

Additionally researchers suggested that global information should be considered. For example, the authors (Hoffman and Singh 1997) proposed three factors that determined the shape part importance – the relative size, protrusion degree and boundary strength. However, specific computational model was lacking in the work.

1.1.3 Feature Weighting

In a broader view, part importance evaluation is closely related to the literature on feature weight learning. For example, the minimax entropy framework (Zhu et al. 1998) learned a generative model of texture by selecting features weights by learning-by-sampling. Kersten et al. (Kersten et al. 2004) suggested a general principle to determine feature weights based on a Bayesian framework. Features with more reliable information had higher weights attributed to their corresponding prior constraint. Our observation is that, in addition to Kersten et al.'s principle, (which corresponds to the first factor of part variation and reconstruction quality), the feature/part uniqueness factor should also be taken into consideration.

By contrast in the discriminative paradigm, feature weighting mechanisms are often automatically embodied. For instance, the RELIEF algorithm weighs features according to the information gain based on the nearest "hit" and nearest "miss" (the two nearest neighbors of the positive class and negative class) (Kira and Rendell 1992). Boosting and its variants learn weights of weak classifiers (taken as features) according to classification error rate in an iterative procedure. In maximum margin based models, kernels are implicit features, and the notion of margin corresponds to the weights of the implicit features e.g. (Cai et al. 2010). Feature weights are formulated in the potential functions in conditional random field (CRF) models e.g. (Schnitzspan et al. 2010). Additionally, there are discriminative feature selection methods using simple criteria such as feature statistics e.g. (Ullman 2007) or certain utility functions e.g. (Freifeld et al. 2010). But it is known that the discriminative methods are subject to the positive and negative classes. Hence model generalization ability is generally inferior to that of generative models.

The rest of the paper is organized as follows. In Sect. 2 we present the shape part importance formulation, which is based on a proposed shape reconstruction approach in Sect. 3. Then we show how to apply the proposed part importance to object detection (Sect. 4). Section 5 demonstrates psychological experiments which support our model. Finally, Sect. 6 concludes the paper.

2 Shape Part Importance

We introduce the problem formulation and computation of part importance in this section.

2.1 Problem Formulation

We focus on the importance evaluation of 2D object shape parts, i.e. the contour-based parts of an object category. The specific object and part representations are introduced later in Sect. 2.2.1.

The importance of a part is measured by its reconstructability of the object shape, which is determined by (i) part variation, which decides the reconstruction quality; and (ii) the part uniqueness, which depends on whether a part can be matched to the object at different locations on the object contour. Taking these two issues into consideration, we define a conditional entropy formulation which describes the uncertainty of shape reconstruction induced by the matching location ambiguities and the reconstruction problem itself.

$$H_{\Pi}(Y, L|X_{\Pi}; \mathcal{O}) = - \int \int \int p(X_{\Pi}; \mathcal{O}) p(Y, L|X_{\Pi}; \mathcal{O}) \cdot \log p(Y, L|X_{\Pi}; \mathcal{O}) dX_{\Pi} dY dL, \quad (1)$$

where X_{Π} represents the shape of a part, \mathcal{O} denotes the shape model of an object category, Y denotes recovered object shapes, and L denotes the matching location of the given shape part on the object shape. $p(X_{\Pi}; \mathcal{O})$ is the prior of the part.

$$p(Y, L|X_{\Pi}; \mathcal{O}) = p(Y|X_{\Pi}, L; \mathcal{O}) p(L|X_{\Pi}; \mathcal{O}) \quad (2)$$

where $p(Y|X_{\Pi}, L; \mathcal{O})$ represents the reconstruction probability given part X_{Π} when it is matched to a location, and $p(L|X_{\Pi}; \mathcal{O})$ is the probability of such matching.

The reconstruction problem is formulated as a Maximum a Posteriori (MAP) estimation. Given the contour part X_{Π} as the observation, and the object shape model \mathcal{O} , the goal is to infer the most probable reconstructed shape. The specific solution is introduced in Sect. 3.1.

In the above conditional entropy formulation, the part variation affects the reconstruction quality and the probability $p(Y|X_{\Pi}, L; \mathcal{O})$. Meanwhile the part uniqueness greatly determines the uncertainty of the reconstruction and the entropy H . Even if a part has low variability and is able to recover an object shape nicely, but if it does not have a unique match and produces many good reconstructions at different correspondence locations, such as the flat fragment in Fig. 1, the uncertainty of the shape reconstruction is still high so that the part cannot be considered to be "important". In contrast, an important part generates only a very small number of good reconstructions.

2.2 Implementation

Here we go into the details of our computations of Eq. 1. First of all, we should define the object models and part-based representation. For 2D shapes, there are some key problems involved, such as shape variation (deformation and transformation), viewpoint and articulation. In this paper, we mainly focus on shape variation, and learn different object and part models with respect to different viewpoints, just as the popular way in the literature. Articulation is not modeled in the current version; it is a challenging problem and discussed in some literature e.g. (Ion et al. 2011).

2.2.1 Object and Part Models

To learn the shape model of an object category, a set of object contour instances of the category is utilized as training data, e.g. the labeled object outlines from the ETHZ dataset (Ferrari et al. 2006). An object contour Y is represented by a set of contour points (the object center as the origin), first, the object contour instances of the category $\{Y_i\}, i \in \{1, \dots, n\}$ are aligned in location and orientation, and normalized in scale. The alignment is based on the TPS-RPM shape matching algorithm (Chui and Rangarajan 2003), which infers the point correspondences between object contours. Consequently, the object shape model is learned by the ASM method (Cootes et al. 1995), assuming Y follows a normal distribution

$$p(Y; \mathcal{O}) \sim \mathcal{N}(\mathcal{T}_{\mathcal{O}}, \Sigma_{\mathcal{O}}), \tag{3}$$

where $\mathcal{T}_{\mathcal{O}}$ is the mean object shape, $\Sigma_{\mathcal{O}}$ is the covariance matrix of the contour points.

A part refers to a localized fragment on the object contour. Considering different shape variations of a part, we make use of all the part instances extracted from the object instances. There are two key issues of part-based representation: (I) how to learn object parts, and (II) how to use the learnt part models to extract part instances of an object. In the following, we introduce these two issues in details.

(I) A number of part candidates are extracted from object contours; then, the candidates are clustered into several groups, each group corresponds to a ‘‘part’’.

To extract part candidates from object contours, many successful methods in literature can be adopted, e.g., (Shotton et al. 2008; Opelt et al. 2008). To demonstrate that our part importance model is generally applicable to the contour-based representation (or not limited to a particular part generation method), we take two approaches as examples. One is the k AS (k -Adjacent Segments) detector (Ferrari et al. 2008), which finds short line segments and generates local shape configurations by combining k adjacent segments ($k = 1, 2, \dots$). We extract k AS-based parts on the objects of the ETHZ (Ferrari et al. 2006) and INRIA-horses (Jurie and Schmid 2004) datasets. The other is the convex shape decomposition (Liu et al. 2010), which cuts a shape into segments under concavity constraints. We extract this type of parts on the objects of the MPEG-7 and PASCAL datasets (Everingham et al. 2010). For both methods, part candidates are extracted from the aligned and normalized object contours.

Denote the obtained part candidates as $\tilde{\pi}_i = (\tilde{l}_i, \tilde{X}_i)$, $i \in \{1, \dots, m\}$, where \tilde{l} and \tilde{X} are the relative location of the part center and part contour points respectively, in the coordinate system with the object center as the origin. Then the part candidates are clustered (as in (Ferrari et al. 2009; Luo et al. 2010)) according to their similarities in shape and relative position. Consequently, a set of canonical parts are generated from each cluster, $\Pi_i = (l_i, X_i), i \in \{1, \dots, N\}$,

where $l_i = \frac{1}{|C_i|} \sum_{j \in C_i} \tilde{l}_i$, and $X_i = \frac{1}{|C_i|} \sum_{j \in C_i} \tilde{X}_j$ (C_i is a cluster of part candidates). Some extracted part candidates and the clustering results are shown in Fig. 2. Notice that there is no limitation to the lengths of parts.

(II) The instance of part Π is extracted by matching the contour points X of Π to each object contour instance at the part’s canonical location l , using the method in (Riemenschneider et al. 2010). The best locally matched piece X_{Π} is taken as the contour fragment of the part instance $\pi = (l, X_{\Pi})$. Some extracted part instances are shown in Fig.2 (b).

Note that part instances are identified on the normalized object contours, from which the part models are learned, thus there is no scale ambiguity. This process will be used in the next step to measure part importance.

2.2.2 Computing the Conditional Entropy

To make Eq. 1 computationally feasible, we discretize the formulation:

$$\begin{aligned} & H_{\Pi}(Y, L | X_{\Pi}; \mathcal{O}) \\ &= - \sum_i \sum_l \sum_j p(X_{\Pi,i}; \mathcal{O}) p(Y_{lj}, L_l | X_{\Pi,i}; \mathcal{O}) \\ & \quad \cdot \log p(Y_{lj}, L_l | X_{\Pi,i}; \mathcal{O}) \end{aligned} \tag{4}$$

where $\{X_{\Pi,i}\}, i = 1, \dots, n$ is the set of contour fragments of part Π ’s instances extracted from the training dataset of a studied class using the method described in Sect. 2.2.1. $p(X_{\Pi,i}; \mathcal{O})$ is the prior of each part instance, and it is chosen to be uniform in this paper. Take each part instance as an observation to reconstruct the complete object shape. The part is matched to any location of the object model. Here a discrete set of normalized locations $L_l \in [0, 1]$ is uniformly sampled along the whole object contour with equal spacing. Y_{lj} is one reconstructed shape from $X_{\Pi,i}$, when the segment is matched to the object contour at location L_l .

To further reduce the computation, at each matching location, instead of summing over all possible reconstructed shapes, we estimate the most probable one Y_l^* and use $p(Y_l^*, L_l | X_{\Pi,i}; \mathcal{O}) \log p(Y_l^*, L_l | X_{\Pi,i}; \mathcal{O})$ as an approximation to $\sum_j p(Y_{lj}, L_l | X_{\Pi,i}; \mathcal{O}) \log p(Y_{lj}, L_l | X_{\Pi,i}; \mathcal{O})$ in Eq. 4. Then the conditional entropy formulation is reduced to

$$\begin{aligned} & H_{\Pi}(Y, L | X_{\Pi}; \mathcal{O}) \\ &= - \sum_i p(X_{\Pi,i}; \mathcal{O}) \sum_l p(Y_l^*, L_l | X_{\Pi,i}; \mathcal{O}) \\ & \quad \times \log p(Y_l^*, L_l | X_{\Pi,i}; \mathcal{O}) \\ &= - \sum_i p(X_{\Pi,i}; \mathcal{O}) \sum_l p(Y_l^* | X_{\Pi,i}; \mathcal{O}) p(L_l | X_{\Pi,i}; \mathcal{O}) \\ & \quad \cdot \log p(Y_l^* | X_{\Pi,i}; \mathcal{O}) p(L_l | X_{\Pi,i}; \mathcal{O}) \end{aligned} \tag{5}$$

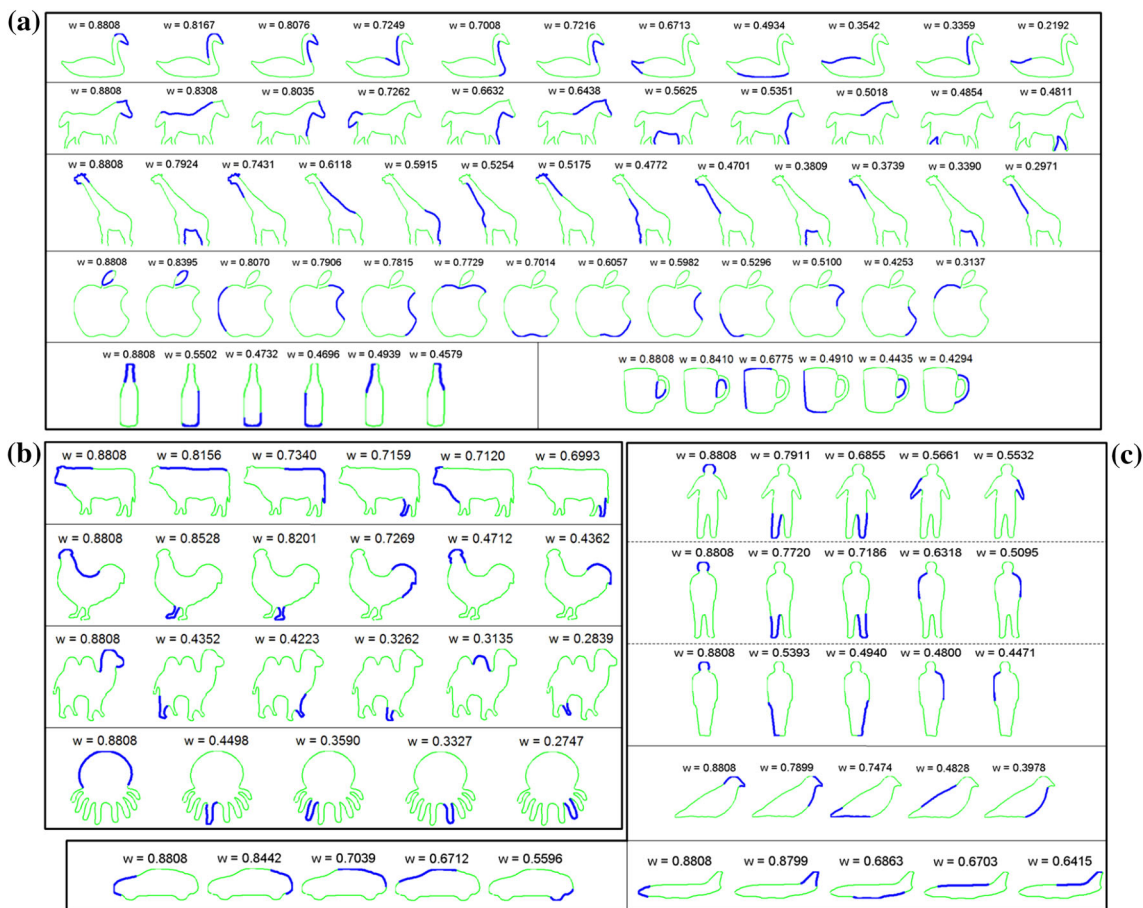


Fig. 3 Ranked importance of parts (*highlighted in blue*) on (a) the ETHZ (Ferrari et al. 2006) and INRIA-horses datasets (Jurie and Schmid 2004), (b) the MPEG-7 shape classes and (c) the PASCAL datasets

(Everingham et al. 2010). The numbers denote the part importance values which are functions of the conditional entropy, see Eq. 8. The green contours are the learned mean shapes (Color figure online)

where $p(L_l|X_{\Pi,i}; \mathcal{O})$ is considered uniform. And

$$p(Y_l^*|X_{\Pi,i}; \mathcal{O}) \propto p(X_{\Pi,i}|Y_l^*; \mathcal{O})p(Y_l^*; \mathcal{O}), \quad (6)$$

where $p(X_{\Pi,i}|Y_l^*; \mathcal{O})$ represents the consistency between the part instance $X_{\Pi,i}$ and the corresponding part of the reconstructed shape Y_l^* ; and this consistency is measured based on the L1-norm distance of the two corresponding parts. $p(Y_l^*; \mathcal{O})$ describes how much the MAP shape agrees with the object shape model. Based on Eq. 3, there is

$$p(Y_l^*; \mathcal{O}) \propto \exp\{-(Y_l^* - \mathcal{T}_{\mathcal{O}})^T \Sigma_{\mathcal{O}}^{-1} (Y_l^* - \mathcal{T}_{\mathcal{O}})\}. \quad (7)$$

2.2.3 The Part Importance Measure

The part of lower reconstruction uncertainty is considered of higher reconstructability, and hence is important.

Accordingly, we define the *part importance* as

$$w_{\Pi} = \frac{1}{1 + \exp\{c_1 H_{\Pi} + c_2\}}. \quad (8)$$

The definition is a non-linear transformation, the logistic function, where c_1 and c_2 are two parameters selected to ensure that the values of the linear term in the exponential fall into the range $(-2, 2)$ for all object categories. Notice that the tuning of these parameters will only change the absolute importance values, but not the relative rankings of the parts.

Figure 3a shows the ranked parts of the ETHZ shape classes (Ferrari et al. 2006) and INRIA-Horses dataset (Jurie and Schmid 2004) according to the part importance computed Eq. 8. It worth noticing that some groundtruth outlines are not complete shapes, e.g. the legs are pruned away from the outlines of the giraffes. Figure 3b, c show examples on the MPEG-7 and PASCAL datasets (Everingham et al. 2010) respectively.

According to Eq. 5, a part importance is computed based on the reconstruction qualities of all the part instances at all matching positions along the object contours. Therefore, for an object category with N parts, the computational complexity is $N \times n \times k$, where n and k are the number of each part's instances and the number of the sampled matching positions.

In the following section we will present a detailed shape reconstruction method, which gives an efficient algorithm for the implementation of shape reconstruction from partial observations.

3 Shape Reconstruction

There have been many previous work on shape completion, for example, amodal completion (Kanizsa and Gerbino 1982), which makes use of heuristics such as local continuities, proximity and global regularities; and curve completion (Kimia et al. 2003) under the rules such as smoothness or curvature-based constraints. However, these generic priors are usually only successful in bridging small gaps of shape contours; in case of severe occlusions, they are often unable to recover reasonable estimates of the object shapes. Therefore, we propose a new shape reconstruction method, which leverages global shape priors in order to estimate complete object outlines from a curve fragment (i.e. a part).

3.1 Our Solution

In the 2D space of planer shapes (denoted by $\Omega_{\mathcal{R}^2}$), given a contour fragment X_{Π} of certain part, assuming its correspondence location l on a object shape is known, the goal is to infer the most probable complete shape contour Y_l^* with respect to the object class model \mathcal{O} . It is formulated as a MAP estimation problem under the Bayesian framework.

$$Y_l^* = \operatorname{argmax}_{Y \in \Omega_{\mathcal{R}^2}} p(Y|X_{\Pi}, l; \mathcal{O}) \tag{9}$$

$$= \operatorname{argmax}_{Y \in \Omega_{\mathcal{R}^2}} p(Y; \mathcal{O}) p(X_{\Pi}|Y, l) / p(X_{\Pi}; \mathcal{O})$$

where $p(X_{\Pi}; \mathcal{O})$ is independent of Y and hence does not affect Y_l^* .

We implement this MAP estimation as an energy minimization,

$$Y_l^* = \operatorname{argmax}_{Y \in \Omega_{\mathcal{R}^2}} E_G + \lambda E_P \tag{10}$$

where E_G is the energy term of the global shape prior $p(Y; \mathcal{O})$, and it is used to constrain Y to follow the shape model of the class as much as possible. Recall that the global shape prior was described in Sect. 2.2.1 and is a Gaussian distribution $\mathcal{N}(\mathcal{T}_{\mathcal{O}}, \Sigma_{\mathcal{O}})$. Let \mathbf{b} be the projection vector of Y onto the principle components of object class shape space. Y can be approximated by

$$Y \approx \mathcal{T}_{\mathcal{O}} + \Phi \mathbf{b}. \tag{11}$$

where Φ is the eigenvectors of the covariance $\Sigma_{\mathcal{O}}$. And the energy term is

$$E_G(\mathbf{b}) = \mathbf{b}^T \Sigma_{\mathcal{O}}^{-1} \mathbf{b} \tag{12}$$



Fig. 4 Shape completion examples. The blue contour segments are the object boundaries not been occluded. The red shapes denote the reconstructed contours by the proposed shape reconstruction method. The dotted green curves are the occluded ground truth object outlines. The regions of chessboard pattern are the occlusion masks (Color figure online)

E_P is the energy term of the likelihood $p(X_{\Pi}|Y, l)$. This term enforces a partial constraint, which imposes a good matching between the observation X_{Π} and its corresponding segment on the reconstructed shape Y_{Π} .

$$E_P(\mathbf{b}) = |Y_{\Pi}(\mathbf{b}) - X_{\Pi}| \tag{13}$$

$$Y_{\Pi}(\mathbf{b}) \approx \mathcal{T}_{\Pi} + \Phi_{\Pi} \mathbf{b}. \tag{14}$$

Equation 14 is derived from the decomposed form of Eq. 11,

$$Y = \begin{bmatrix} Y_{\Pi} \\ Y_{\Pi'} \end{bmatrix} \approx \begin{bmatrix} \mathcal{T}_{\Pi} \\ \mathcal{T}_{\Pi'} \end{bmatrix} + \begin{bmatrix} \Phi_{\Pi} \\ \Phi_{\Pi'} \end{bmatrix} \mathbf{b}. \tag{15}$$

where \mathcal{T}_{Π} and Y_{Π} are the corresponding parts at the matching location l on the mean shape and the reconstructed shape separately, and Φ_{Π} is corresponding submatrix of Φ . $\mathcal{T}_{\Pi'}$, $Y_{\Pi'}$ and $\Phi_{\Pi'}$ are the rest parts in the decomposition.

It remains to find the optimal \mathbf{b} by performing MAP estimation which reduces to minimizing the following energy function,

$$E(\mathbf{b}) = E_G(\mathbf{b}) + \lambda E_P(\mathbf{b}) \tag{16}$$

where λ is set to 10 to balance the two constraints of the global prior and local observation. We solve this optimization problem using the method described in (Duchi et al. 2008). The MAP shape Y_l^* is then reconstructed using Eq. 11.

Figure 4 shows shape reconstruction examples for completing occluded object contours. It demonstrates that although the observed parts have some deformation, the object contours are still well recovered by taking advantage of the global shape prior. This demonstrates the strong reconstruction power of the proposed method under shape deformation and severe occlusion.

4 Object Detection with Part Importance

In this section we show that, by exploiting the proposed part importance in object detection, we can improve the performance of object detectors; and in particular, we show this for the challenging case of severely occluded objects. We note that important parts can be used to find objects efficiently and so reduce computation cost.

4.1 Part Importance in Object Detection

We adapt the popular voting-based method for object detection. There are two steps, (1) proposing object candidates using a Hough-style voting scheme for estimating object location and scale; (2) hypothesis verification to determine whether a candidate is a target object. We adopt two verification techniques. The first uses a non-rigid shape matcher to localize the exact object boundaries (adapted from (Ferrari et al. 2009)). The second uses support vector machines (SVM) to verify whether an object is present in a bounding box containing the object candidate.

Part importance is applied to both steps. In the voting step, part importance is used to weigh parts during the voting. In the verification step, firstly, in the shape-matching-based method, part importance is used to infer correspondences in the matching and also used to weigh the total matching score. The intuition is that important parts should be matched in priority, otherwise the matching cost will be high. Second, we introduce an *importance kernel* into the SVM framework, which is a function of the part importance. The followings are the implementation details.

4.1.1 The Voting Step

Given a testing image, we first extract contour segments from the image edge maps (Martin et al. 2004) (the threshold of PB edge maps is set to $0.1 \times$ the maximum value of a PB edge map). Then, each contour segment is matched to the learned shape parts X_{Π} to vote for the object center and scale. The voting score v is determined by the part importance w_{Π} in addition to the (standard) similarity $\alpha_{\Pi,i}$ of the matching pair, the strength of the image contour (denoted by e_i , i is the index of a contour segment):

$$v_{\Pi,i} = w_{\Pi} \cdot e_i \cdot \alpha_{\Pi,i}. \quad (17)$$

The importance weight emphasizes the votes of the important parts so that the effect from distracters in the background clutters and other noises is reduced.

The above procedure was adapted from the method of the literature (Ferrari et al. 2009). The similarity of shape parts are measured based on the k AS feature distances.

4.1.2 Object Hypothesis Verification

(I) The Shape Matching Method

The verification stage based on shape matching is implemented by back-matching the object model to each object hypothesis. The shape matching method is adapted from the TPS-RPM algorithm (Chui and Rangarajan 2003). A deterministic annealing process is adopted to jointly estimate the correspondence and transformation between two shapes. The objective is to minimize the following energy function,

$$E_{\text{TPS-RPM}} = \sum_i \sum_j w_i m_{ij} \|z_j - \psi(x_i)\|^2 - \sum_i \sum_j w_i m_{ij} + t \sum_i \sum_j m_{ij} \log m_{ij} + \lambda_1 \text{trace}(W^T K_{\text{TPS}} W) + \lambda_2 |A - I|. \quad (18)$$

where x_i and z_j denote two points from a shape part in the model and the candidate object contour in the test image respectively. For the point-based representation in this shape matching, the importance w_i of x_i is inherited from the part to which x_i belongs. m_{ij} represents the correspondence of two points and ψ is the shape transformation, where $\psi(\mathbf{x}) = \mathbf{x}A + K_{\text{TPS}}W$; A , W and K_{TPS} denote the affine transformation, warping deformation and the TPS kernel (Chui and Rangarajan 2003) respectively. λ_1 and λ_2 are two parameters balancing the amount of wrapping deformation and affine transformation, $\lambda_1 = 0.1, \lambda_2 = 0.5$. t is the annealing temperature initialized as 5 and decreases at the rate 0.8, till t goes below 0.001.

In the adapted energy function, the first term in Eq. 18 is a weighted matching distance to penalize large matching cost of important parts. The second term encourages more important parts to be matched. The third term is an entropy barrier function with the temperature parameter t as defined in the TPS-RPM algorithm. The fourth and the last term penalize the total amount of warping deformation and affine transformation respectively. I is the identity matrix.

For object candidate verification, we compute the total matching cost of a candidate object as:

$$E_s = \beta_1 \sum_i \sum_j w_i m_{ij} \|z_j - \psi(x_i)\|^2 + \beta_2 \frac{\sum_i w_i \cdot \mathbf{1}(m_i < \theta)}{\sum_i w_i} + \beta_3 \text{trace}(W^T K_{\text{TPS}} W) + \beta_4 |A - I|. \quad (19)$$

If $\exp(-E_s)$ is above a threshold (set to be 0.2 in our experiments), the object candidates is accepted as a detected object. In Eq. 19, the terms are similarly defined as in Eq. 18) except for the second one. Here m_i is the maximum value of m_{ij} , $j \in \{1, \dots, K\}$, where K is the number of the candidate points. $\mathbf{1}(\cdot)$ is an indicator function, and $\theta = \frac{1}{M}$ is a matching

threshold (M is the number of points of the object shape model). $(\beta_1, \beta_2, \beta_3, \beta_4) = (8, 0.2, 0.2, 0.5)$ for all the object classes in the experiments.

(II) The Importance Kernel in SVMs

A second verification approach is based on a trained SVM classifier, which is applied to a window containing an object candidate. We propose an ‘‘importance kernel’’ for the SVMs. The kernel is designed as $\mathbf{K}_I(f_i, f_j) = f_i^T \mathbf{W}_I f_j$, where f is a feature and $\mathbf{W}_I = \mathbf{w}_I \mathbf{w}_I^T$, \mathbf{w}_I is a column vector, in which each entry is the importance value of a shape part computed by Eq. 8. It is obvious that \mathbf{W}_I is positive semi-definite, so that the kernel is valid and satisfies Mercer’s condition (Shawe-Taylor and Cristianini 2004).

The feature f_i is related to a descriptor in candidate window. Each contour fragment in the window is matched with all the parts. The matching score α is computed as in Sect. 4.1.1. We calculate a histogram where each bin counts the total accumulation of the matching scores for each part. This histogram is adopted as the descriptor f_i . Then, instead of using the standard linear kernel (Ferrari et al. 2008), we apply the importance kernel to the SVM classifier.

4.2 Experimental Results

4.2.1 Results on Standard Datasets

On the ETHZ shape dataset (Ferrari et al. 2006), the most popular benchmark for contour-based object detection, we use half of the images of each object class for training and the others for testing, just as the traditional way on this benchmark. Experiments were conducted using the proposed importance-based object detection with two kinds of approaches for object candidate verification.

In the hough voting stage (as in Sect. 4.1.1), The votes are weighted by part importance (Eq. 17) and an accumulated voting score above a preset threshold ($0.3 \times \#$ object parts) signifies a candidate object hypothesis. We compare this bottom-up process with the one without considering part importance (or identical part weights) in Table 1. Experimental results show that weighted voting improves the detection

rate. We use the most popular criteria, i.e. the detection rate at 1.0 FPPI, and also compare with other state-of-the-art voting-based methods as shown in Table 1. The average performance is comparable to (Riemenschneider et al. 2010). Notice that two types of ranking process are adopted in (Riemenschneider et al. 2010), one is based on the coverage score of the matched reference contour; the other is the PMK score based on a SVM classifier.

For the *shape-matching verification* (as introduced in Sect. 4.1.2 (I)), it is shown in Fig. 5 that the detection rate vs. false positives per image (DR/FPPI) is improved by using part importance compared with that without part importance. For thorough comparisons with the other state-of-the-arts detection approaches, Table 2 and 3 show the detection rate at 0.3/0.4 FPPIs and the Interpolated Average Precision (AP) as in the PASCAL VOC challenge (Everingham et al. 2010) respectively. It illustrates that our detection rate and the Average Precision of the ETHZ object classes achieve a comparable result to the state-of-the-arts e.g. (Lin et al. 2012) and (Yarlagadda and Ommer 2012) or even better performances than many recent work. It should be point out that among the state-of-the-arts shape-based methods, some rely on labeled contours for training, such as (Wang et al. 2012) and ours, and some only need weekly supervision such as (Yarlagadda and Ommer 2012) and (Srinivasan et al. 2010). The more supervision could be more helpful to improve the performances.

Figure 6 compares the localization of object boundaries by the proposed method using part importance (in green) and those without part importance (in red). It shows that the method which exploits part importance locates object boundaries much more accurately than the method which do not (an adapted version of (Ferrari et al. 2009) – the differences in the part-based object model learning is discussed in Sect. 4.1.1). Table 4 shows the boundary coverage/precision criteria as in (Ferrari et al. 2009). The coverage is the percentage of the ground-truth outline points that have been detected, and the precision is the percentage of the true positive object boundary points. The localization accuracies of our shape-matching-based verification with part importance are also competitive to the state-of-the-arts, except that (Toshev et al. 2012) obtained more accurate localization than ours for

Table 1 Comparison of the detection rates of the voting step at 1.0 FPPI

ETHZ classes	Ferrari et al. (2009)	Ommer and Malik (2009) (w_{ac} / PMK ranking)	Maji and Malik (2009)	Riemenschneider et al. (2010) (Cov/PMK)	Ours without importance	Ours with importance
Applelogos	0.430	0.800 / 0.800	0.850	0.904 / 0.904	0.739	0.783
Bottles	0.644	0.924 / 0.893	0.670	0.844 / 0.964	0.778	0.815
Giraffes	0.522	0.362 / 0.809	0.550	0.500 / 0.788	0.778	0.867
Mugs	0.451	0.475 / 0.742	0.550	0.323 / 0.614	0.763	0.790
Swans	0.620	0.588 / 0.686	0.425	0.901 / 0.886	0.688	0.813
Mean	0.533	0.630 / 0.786	0.609	0.694 / 0.832	0.749	0.814

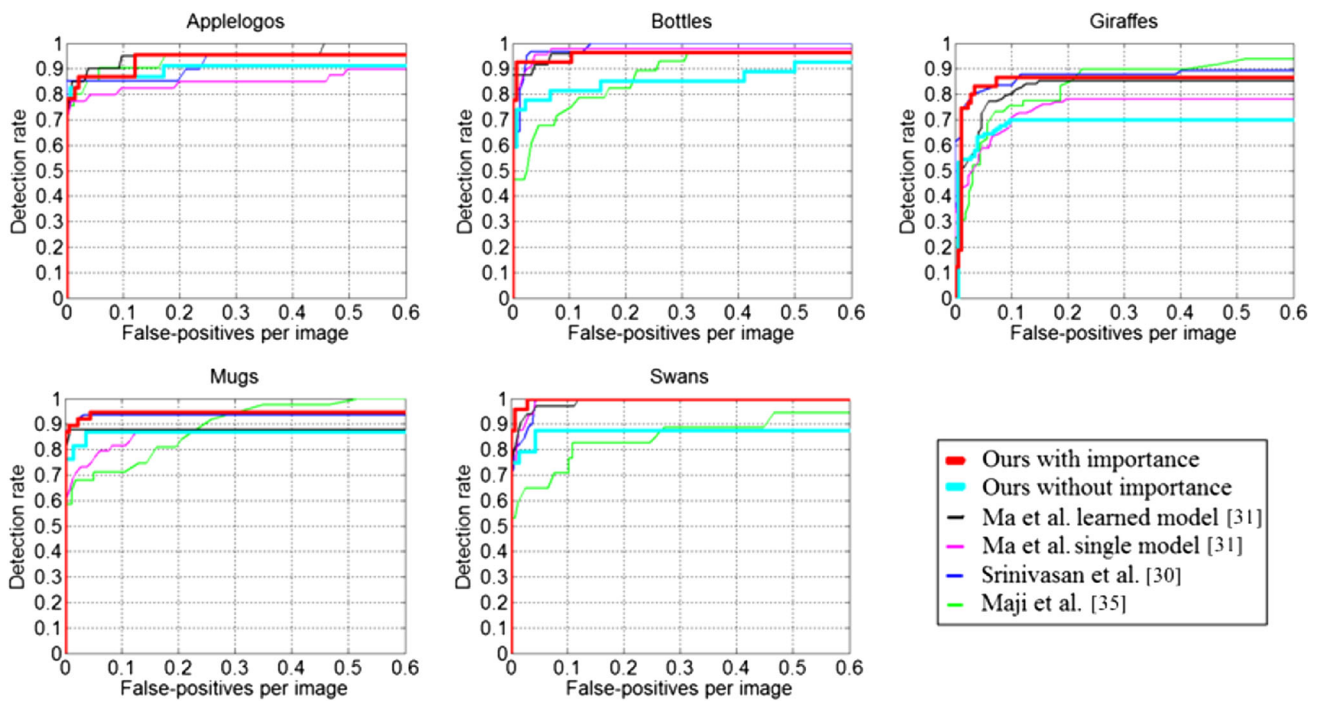


Fig. 5 Performance comparison of detection rate vs. false positives per image (DR/FPPI) by the different object detectors. In this figure our method uses voting followed by shape-matching verification (Color figure online)

Table 2 Comparison of the detection rates at 0.3/0.4 FPPI.

	Applelogos	Bottles	Giraffes	Mugs	Swans	Mean
Ours with importance	0.957 / 0.957	0.963 / 0.963	0.867 / 0.867	0.947 / 0.947	1 / 1	0.947 / 0.947
Ours without importance	0.913 / 0.913	0.852 / 0.852	0.701 / 0.701	0.868 / 0.868	0.875 / 0.875	0.842 / 0.842
(Yarlagadda and Ommer 2012)	0.95 / 0.95	1 / 1	0.913 / 0.913	0.967 / 0.967	1 / 1	0.965 / 0.965
(Wang et al. 2012)	0.90 / 0.90	1 / 1	0.92 / 0.92	0.94 / 0.94	0.94 / 0.94	0.94 / 0.94
(Toshev et al. 2012)	0.864 / 0.886	0.964 / 0.964	0.813 / 0.868	0.727 / 0.773	0.939 / 0.939	0.861 / 0.886
(Ma and Latecki 2011)	0.92 / 0.92	0.979 / 0.979	0.854 / 0.854	0.875 / 0.875	1 / 1	0.926 / 0.926
(Srinivasan et al. 2010)	0.95 / 0.95	1 / 1	0.872 / 0.896	0.936 / 0.936	1 / 1	0.952 / 0.956
(Maji and Malik 2009)	0.95 / 0.95	0.929 / 0.964	0.896 / 0.896	0.936 / 0.967	0.882 / 0.882	0.919 / 0.932
(Felzenszwalb et al. 2009)	0.95 / 0.95	1 / 1	0.729 / 0.729	0.839 / 0.839	0.588 / 0.647	0.821 / 0.833
(Lu et al. 2009)	0.9 / 0.9	0.792 / 0.792	0.734/0.77	0.813/0.833	0.938 / 0.938	0.836 / 0.851
(Riemenschneider et al. 2010)	0.933 / 0.933	0.970 / 0.970	0.792 / 0.819	0.846 / 0.863	0.926 / 0.926	0.893 / 0.905
(Ferrari et al. 2009)	0.777 / 0.832	0.798 / 0.816	0.399 / 0.445	0.751 / 0.8	0.632 / 0.705	0.671 / 0.72
(Zhu et al. 2008)	0.800 / 0.800	0.929 / 0.929	0.681 / 0.681	0.645 / 0.742	0.824 / 0.824	0.776 / 0.795

Our methods here are voting + shape-matching verification

most ETHZ classes. This is mainly due to their employment of figure/ground segmentation (Table 4).

Next we evaluate the *SVM verification* method (introduced in Sect. 4.1.2 (II)). We compare the performance of the traditional linear kernel (Ferrari et al. 2008) with the proposed importance kernel. As shown in Fig. 7 and Table 5, both the detection rate and localization accuracy are improved by using the importance kernel. In Table 5 the Bounding Box Accuracy (BB Accuracy) refers to the average area rate of the intersection vs. union of the ground-truth bounding boxes

and the detected windows. The result showed that the detection rate of the SVM-based verification with part importance was lower than that of the shape matching verification, as demonstrated by the detection rate at 0.3 / 0.4 FPPI in Table 5 vs. 2, and also the DR / FPPI curves in Fig. 7 vs. 5.

Besides, we test on the PASCAL dataset (Everingham et al. 2010), which is a much more challenging dataset for object detection. The proposed part importance measure is based on two assumptions: (i) Objects should have stable shape priors. Several object classes are not considered here,

Table 3 Comparison of interpolated average precisions (AP).

	Apple-logos	Bottles	Giraffes	Mugs	Swans	Mean	
Our method with importance	0.861	0.902	0.796	0.902	0.953	0.883	
Our method without importance	0.856	0.808	0.617	0.813	0.798	0.778	
(Lin et al. 2012)	0.909	0.898	0.811	0.893	0.964	0.895	
(Wang et al. 2012)	0.866	0.975	0.832	0.843	0.828	0.869	
(Ma and Latecki 2011)	0.881	0.920	0.756	0.868	0.959	0.877	
(Srinivasan et al. 2010)	0.845	0.916	0.787	0.888	0.922	0.872	
(Maji and Malik 2009)	0.869	0.724	0.742	0.806	0.716	0.771	
(Felzenszwalb et al. 2009)	0.891	0.950	0.608	0.721	0.391	0.712	
Our methods here is voting with shape-matching verification	(Lu et al. 2009)	0.844	0.641	0.617	0.643	0.798	0.709



Fig. 6 Comparison results of object detection and boundary localization by the method without (Ferrari et al. 2009) (in red) and with (in green) part importance by the shape-matching verification (Color figure online)

Table 4 Comparison of boundary localization accuracies using the coverage/precision criteria (Ferrari et al. 2009)

	Apple-logos	Bottles	Giraffes	Mugs	Swans
Our method with importance	0.928 / 0.947	0.876 / 0.893	0.715 / 0.797	0.866 / 0.863	0.828 / 0.849
Our method without importance	0.910 / 0.925	0.858 / 0.878	0.686 / 0.750	0.841 / 0.834	0.805 / 0.823
(Toshev et al. 2012)	0.918 / 0.975	0.903 / 0.925	0.768 / 0.824	0.865 / 0.905	0.858 / 0.876
(Ma and Latecki 2011)	0.923 / 0.948	0.845 / 0.903	0.456 / 0.784	0.735 / 0.803	0.848 / 0.909
(Ferrari et al. 2009)	0.916 / 0.939	0.836 / 0.845	0.685 / 0.773	0.844 / 0.776	0.777 / 0.772

Our methods here are voting + shape-matching verification

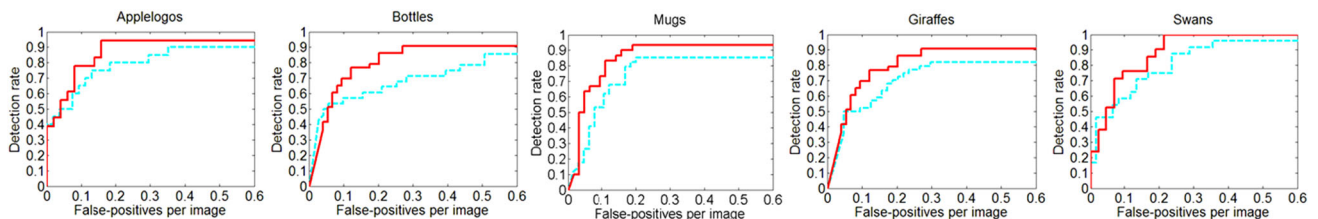


Fig. 7 Performance of detection rate vs. false positives per image. Here the SVM verification with linear kernel (dotted cyan) (Ferrari et al. 2008) and importance kernel (solid red) are compared. The importance kernel is significantly better (Color figure online)

for example, the *potted plants* which are basically free form objects with too large intra-class shape variations, and the *TV-monitors* and *trains* with too simple shapes. (ii) The shape model is view-dependent, i.e. we learn shape models for dif-

ferent viewpoints and different poses separately. For instance, three models are learned for the *person* class (Fig. 3c). This assumption is consistent with the organization of the dataset. Considering the above assumptions, we test the proposed part

Table 5 Comparisons of detection performances of the linear kernel as in (Ferrari et al. 2008) and the proposed importance kernel in the SVM verification

	BB Accuracy		DR at 0.3/0.4 FPPI	
	Linear kernel	Importance kernel	Linear kernel	Importance kernel
Applelogos	0.832	0.849	0.870 / 0.870	0.957 / 0.957
Bottles	0.834	0.907	0.704 / 0.704	0.889 / 0.889
Giraffes	0.769	0.820	0.822 / 0.822	0.911 / 0.911
Mugs	0.808	0.835	0.816 / 0.816	0.921 / 0.921
Swans	0.816	0.832	0.938 / 0.938	1 / 1

importance model on several chosen object classes, which represent the typical challenges in object detection, e.g. relatively large intra-class shape variation and scale variation, a considerable degree of occlusion, articulation, and background clutter. Currently, we have collected images of *aeroplane*, *bird*, *bottle*, *car*, *cow*, *horse* and *person*. For the *bottle* and *person* classes, images of the frontal view are collected, and for the other classes there are images of the left and right views according to the annotations of PASCAL dataset. In all, the numbers of images of the training / validation sets are as follows: 120 / 187 for *aeroplane*, 82 / 170 for *bird*, 200 / 224 for *bottle*, 291 / 318 for *car*, 102 / 125 for *cow*, 151 / 182 for *horse* and 136 / 380 for *person*.

On this derived dataset from PASCAL dataset, the detection performances are improved by the method with part importance (Fig. 8), and the interpolated AP is increased by 4 ~ 8 % by using the proposed part importance and shape-matching-based verification method (Table 6). Notice that most existing tests on the challenging PASCAL dataset always make use of other features besides shape contour, and to the best of our knowledge most contour-based methods have not tested on PASCAL dataset. For thorough comparisons, we show the results of the DPM method (Felzenszwalb and Girshick 2010), one state-of-the-arts but not pure shape-based method (Table 6). Notice that DPM is quite different from ours in the overall detection framework as well as the features used. One advantage of DPM could be the utilization of the HoG features, resulting better performances. However, our method adopts well-trained shape models with part importance, and some results are better than DPM, such as the person class (three models are learned as shown in Fig. 3c).

4.2.2 Part Importance for Occlusion

Occlusion occurs frequently in natural images. It can defeat most of the state-of-the-art object detection and recognition methods. Here we show that by integrating part importance into object detectors greatly improves detection performance even with severe occlusion.



Fig. 8 Comparison results of object detection on a subset of PASCAL dataset. The top / bottom row showing the results by the method without / with part importance (Color figure online)

Table 6 Performances on the PASCAL dataset (Everingham et al. 2010)

	Aeroplane	Bird	Bottle	Car	Cow	Horse	Person
AP (ours without / with importance)	0.418 / 0.487	0.351 / 0.395	0.363 / 0.402	0.380 / 0.422	0.354 / 0.441	0.416 / 0.508	0.343 / 0.414
AP (DPM Felzenszwalb and Girshick (2010))	0.652	0.398	0.428	0.464	0.425	0.539	0.309
BB accuracy (ours without / with importance)	0.791 / 0.825	0.811 / 0.818	0.783 / 0.797	0.796 / 0.822	0.813 / 0.837	0.772 / 0.806	0.788 / 0.807

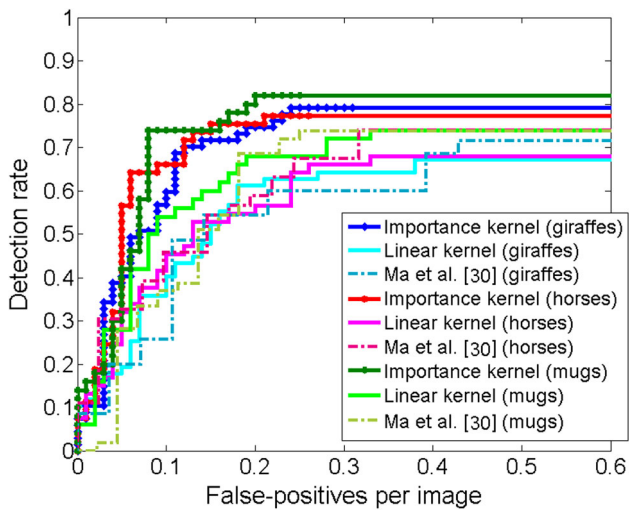


Fig. 9 DR/FPPI comparison on the natural occlusion image datasets (Color figure online)

Due to lack of datasets for occlusion cases in the literature, we construct two types of occlusion datasets. The first one includes images with natural occlusion, downloaded from the Internet. Currently three object classes: horses, giraffes and mugs are collected (30 ~ 50 images for each class). The second is an artificial occlusion dataset in which we synthesize object occlusions by placing occluding bands around the object outlines. The bands are of lengths $a * LEN$, where $a = 0.1, 0.2, \dots, 0.7$, and LEN is the total curve lengths of the object outlines. The images in this dataset are selected from the standard dataset. From each image, we generate 10 occlusion images for every a , with the position of the bands evenly distributed around the object outlines.

We just plug in the learned part importance to the object detectors obtained from standard datasets (with few occlu-

sion) and apply them to the collected occlusion datasets. In this way, we justify the advantage of the proposed importance measure. There is no specific training procedure for occlusion.

Figure 9 and 10 show the results on the natural occlusion dataset. We can see that by using part importance the detection performance is significantly improved compared with that of state-of-the-art methods without part importance such as (Ma and Latecki 2011) and an adapted version of (Ferrari et al. 2008). In Fig. 10, we only detect objects whose pose orientations are the same as the object model (as in Fig. 3), e.g. the horses whose head on the right and the giraffes whose head on the left. Figure 11 shows the Interpolated AP changes in different degrees of occlusion. All these experiments demonstrate the great advantage of exploiting the shape part importance to handle occlusion.

The running time The time costs of part importance learning for the ETHZ classes are as follows, applelogos – 0.51 h, bottles – 0.25 h, giraffes – 0.47 h, mugs – 0.33 h, and swans 0.52 h. The mean value is 0.42 h. For object detection, the average running time is 95.1 s per image (shape-matching based method), where the voting stage and post-processing cost 32.7 and 62.4 s per image respectively. The code is running in 32-bit Matlab (R2009b) on Core (TM) 2 Duo CPU @ 3.0 GHz.

4.2.3 Efficiency Improvement

In the section we show that we can achieve fast detection by using only a small set of important parts. We rank the parts of each object category according to their learned importance. The parts participate one by one in the process of Hough voting for object candidates. Once the voting scores go beyond a predefined threshold, the voting process stops.

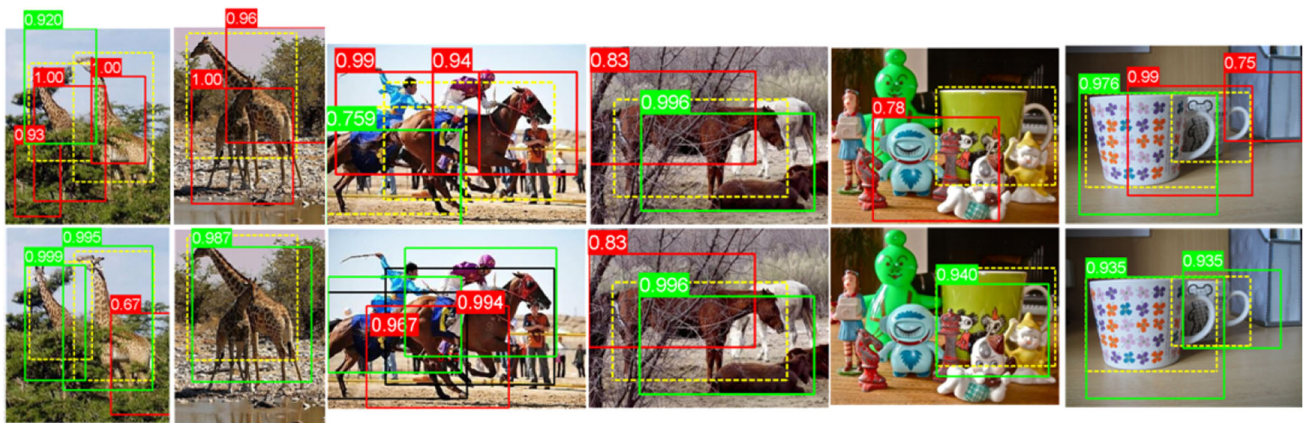
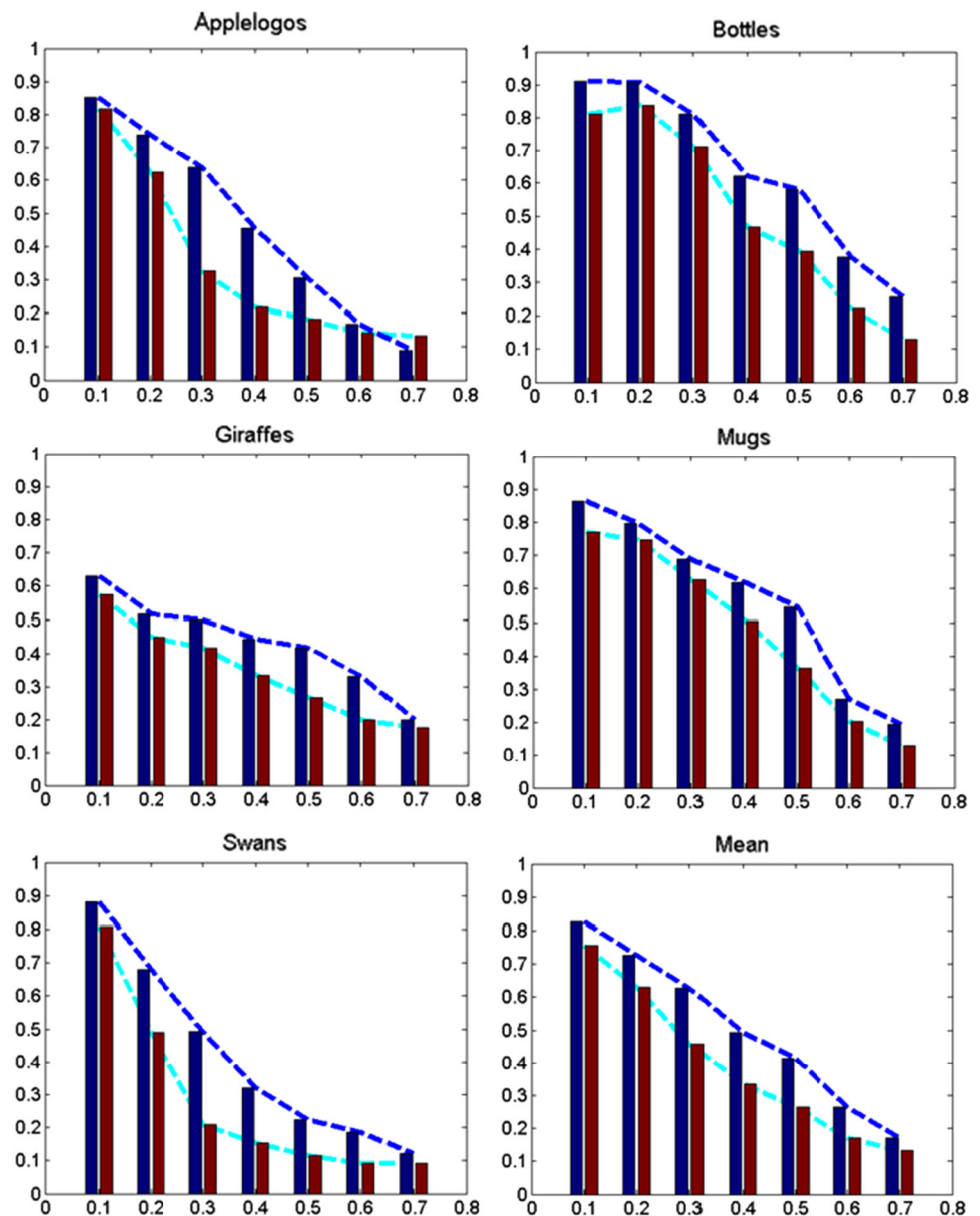


Fig. 10 Object detection & localization on the natural occlusion dataset (using the SVM verification): the top and bottom rows show the results of the method without part importance (Ferrari et al. 2008) and using part importance respectively. The green, red and yellow rectan-

gles denote the true positives, false positives and ground-truth bounding boxes respectively. The numbers on the top-left corner of boxes are the output detection scores from SVM (Color figure online)

Fig. 11 Detection performance vs. the amount of occlusion. The blue and red bars represent the Interpolated APs of the methods using and not using importance (Ferrari et al. 2008) respectively. The envelop curves show the trend of the Interpolated APs changing with different degrees of occlusion (dotted blue/cyan: with/without importance) (Color figure online)



The number of parts used are recorded, which indicates how much the computational cost can be saved in finding potential objects. We start at a relatively large threshold; in this case the detection rate may be very low. Then we gradually lower the threshold, until the performance achieves a comparable level that using the whole set of parts.

Table 7 shows the efficiency improvement obtained by using the ranked parts. We can see that for some object categories more than half of the parts can be removed without significant loss of performance. By comparison, if the parts are added in a random order into the voting process, the efficiency improvement is quite limited. This shows another way that part importance can benefit object detection.

5 Psychological Experiment

We now perform a psychological experiment to see if humans use part importance. Our experiment is based on the work described in (Biederman 1987). In the experiment, a contour part is displayed before the whole object contour is shown. Subjects' task is to name the object as soon as possible. The contour part is used to prime (i.e. speed up) the recognition as so called "primal access" (Biederman 1987). We measure the human subject's part importance by their response time. The shorter the response, the more important the part.

The experimental procedure is described in Fig. 12. We use the ground-truth object outlines in the ETHZ and INRIA-Horses datasets, and also the part instances as described in

Table 7 Efficiency comparison: using parts ordered by importance *versus* a random order.

	Applelogos	Bottles	Giraffes	Mugs	Swans	Mean
Number of parts	26	16	20	17	24	20.6
Ordered by importance (Proportions of parts used / I.A.P. achieved)	0.470 / 0.852	0.684 / 0.875	0.419 / 0.834	0.706 / 0.868	0.547 / 0.913	0.565 / 0.868
Random order (Proportions of parts used / I.A.P. achieved)	0.661 / 0.852	0.767 / 0.875	0.721 / 0.801	0.871 / 0.868	0.706 / 0.903	0.745 / 0.860

I.A.P. stands for Interpolated Average Precision

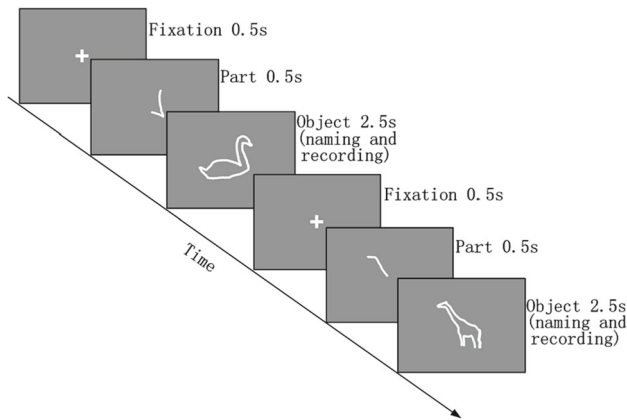


Fig. 12 Illustration of the psychology experiment procedure. The numbers denote the duration of displaying each figure

earlier sections. The stimuli from all the object categories are put together and mixed in a random order. The complete object contour is displayed in the center of the screen. The priming part, which is extracted from the object contour, is allowed to have a small spatial jitter from its original position. The part is presented for 0.5 s and the whole object outlines for 2.5 s. There are 30 subjects (for males, the average ages and standard deviation are 22.7 and 3.2 respectively, and 25.0 and 2.8 for females respectively). It is not necessary to show the stimuli to subjects in advance, but those common objects are supposed to be familiar to them. Subjects need to verbally name the object that is presented. The response time is measured through voice analysis. Then the importance score of part i is derived by $x_i = 1 - t_i / t_{max}$, where t_i is the response time of part i (the mean value of different subjects, notice that the values beyond the 95 % confidence interval are ignored), and t_{max} is the maximal value of the t_i s.

We compare different part importance measures using the above psychological results as the baseline. The importance scores of all parts within the object category form a distribution, $D_0 = (x_i / \sum_i x_i, i = 1, \dots, N)$ (N is the number of the parts). Let D_1 be the distribution from the part importance measure we propose in this paper, and D_2 be that according to the entropy of local edge orientation distributions by Renninger et al. (Renninger et al. 2007), in which high entropies indicate important parts (Both measures are normalized in

Table 8 Comparison of different part importance measures with respect to human performance

KL dist.	Uniform	(Renninger et al. 2007)	Ours
Apple logos	1.9918	1.5654	0.57511
Bottles	0.71266	0.62263	0.42895
Giraffes	2.6431	1.7164	0.748
Mugs	1.2278	0.8110	0.4592
Swans	2.2567	2.2096	1.9676
Horses	1.4104	1.2498	0.93664



Fig. 13 Heat map illustrations of part importance from (a) psychological part importance scores, (b) our model and (c) the measure of edge orientation distribution (Renninger et al. 2007). The more red the more important the part (Color figure online)

the same way as D_0). In addition, if the parts are considered to be of equal importance then it corresponds to a uniform distribution $D_3 = 1/N$. The differences between the three measures to the human subject data are shown in Table 8, where we evaluated by the KL divergences between D_0 and D_1, D_2, D_3 respectively. Figure 13 shows the heat maps based on the different measures. The results exhibit that our model is more coherent to human perception of shape parts comparing with that based on local features as in (Renninger et al. 2007).

6 Conclusion

In this paper, a novel method to measure shape part importance is proposed according to the “shape reconstructability”

of a part. It is successfully applied to a variety of object detection and localization tasks and, in particular, in the presence of severe occlusion. We also perform psychological experiments which show that our model is roughly consistent with human performance.

Our current implementation of the proposed method still has some limitations. For example, it is not very robust to articulation or to large changes in the viewing angles. These cause large part pose variations of the global shape, rather than deformations of shape parts. In future work, we shall augment our model by introducing pose variables to discount these problems. Additionally, the currently proposed part importance relies on contour-based models and requires labeled contours for learning. This may bring about some laborious work compared with the weakly supervised methods.

Acknowledgments We'd like to thank for the support from the following research grants 973-2011CBA00400, NSFC-61272027, NSFC-61121002, NSFC-61231010, NSFC-61210005, NSFC-61103087, NSFC-31230029, and Office of Naval Research N00014-12-1-0883.

References

- Bai, X., & Latecki, L. (2008). Path similarity skeleton graph matching. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 30(7), 1282–1292.
- Belongie, S., Malik, J., & Puzicha, J. (2002). Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 24(4), 509C522.
- Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological Review*, 92, 115–147.
- Biederman, I., & Cooper, E. E. (1991). Priming contour-deleted images: Evidence for intermediate representations in visual object recognition. *Cognitive Psychology*, 23, 393–419.
- Bouchard, G., & Triggs, B. (2005). Hierarchical part-based visual object categorization. In *IEEE Conference on Computer Vision and Pattern Recognition*.
- Bower, G. H., & Glass, A. L. (2011). Structural units and the redintegrative power of picture fragments. *Journal of Experimental Psychology*, 2, 456–466.
- Cai, H., Yan, F., & Mikolajczyk, K. (2010). Learning weights for codebook in image classification and retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*.
- Chui, H., & Rangarajan, A. (2003). A new point matching algorithm for non-rigid registration. *Computer Vision and Image Understanding*, 89(2–3), 114–141.
- Cootes, T. F., Taylor, C. J., Cooper, D. H., & Graham, J. (1995). Active shape models—their training and application. *Computer Vision and Image Understanding*, 61(1), 38–59.
- Crandall, D. J., & Huttenlocher, D. (2006). Weakly supervised learning of part-based spatial models for visual object recognition. In *European Conference on Computer Vision*.
- Dubinskiy, A., & Zhu, S. C. (2003). A multi-scale generative model for animate shapes and parts. In *Proceedings of IEEE International Conference on Computer Vision*.
- Duchi, J., Shalev-Shwartz, S., Singer, Y., & Chandra, T. (2008). Efficient projections onto the l_1 -ball for learning in high dimensions. In *International Conference on Machine Learning*.
- Epshtein, B., & Ullman, S. (2007). Semantic hierarchies for recognizing objects and parts. In *IEEE Conference on Computer Vision and Pattern Recognition*.
- Everingham, M., Gool, L. V., Williams, C. K. I., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International Journal of Computer Vision (IJCV)*, 88(2), 303–338.
- Felzenszwalb, P. F., & Huttenlocher, D. P. (2005). Pictorial structures for object recognition. *International Journal of Computer Vision (IJCV)*, 61(9), 55–79.
- Felzenszwalb, P. F., Girshick, R. B., McAllester, D., & Ramanan, D. (2009). Weakly supervised learning of part-based spatial models for visual object recognition. In *Computer Vision and Pattern Recognition (CVPR)*.
- Ferrari, V., Tuytelaars, T., & Gool, L. V. (2006). Object detection by contour segment networks. In *European Conference on Computer Vision (ECCV)*, dataset. www.vision.ee.ethz.ch/~calvin/datasets.html.
- Ferrari, V., Fevrier, L., Jurie, F., & Schmid, C. (2008). Groups of adjacent contour segments for object detection. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*.
- Ferrari, V., Jurie, F., & Schmid, C. (2009). From images to shape models for object detection. *International Journal of Computer Vision (IJCV)*, 104, 2–3.
- Freifeld, O., Weiss, A., Zuffi, S., & Black, M. J. (2010). Contour people: A parameterized model of 2d articulated human shape. In *IEEE Conference Computer Vision and Pattern Recognition*.
- Gopalan, R., Turaga, P., & Chellappa, R. (2010). Articulation-invariant representation of non-planar shapes. In *European Conference on Computer Vision*.
- Hoffman, D. D., & Richards, W. (1984). Parts of recognition. *Cognition*, 18, 65–96.
- Hoffman, D. D., & Singh, M. (1997). Saliency of visual parts. *Cognition*, 63, 29–78.
- Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cats visual cortex. *Journal of Neurophysiology*, 160, 106–154.
- Ion, A., Artner, N. M., Peyre, G., Kropatsch, W. G., & Cohen, L. D. (2011). Matching 2d and 3d articulated shapes using the eccentricity transform. *Journal of Experimental Psychology*, 115(6), 817–834.
- Jurie, F., & Schmid, C. (2004). Scale-invariant shape features for recognition of object categories. In *IEEE Conference on Computer Vision and Pattern Recognition*. dataset: lear.inrialpes.fr/data.
- Siddiqi BK, K., & Tresness, K. (1996). Parts of visual form: Psychophysical aspects. *Perception*, 25, 399–424.
- Kanizsa, G., & Gerbino, W. (1982). Amodal completion: Seeing or thinking? In J Beck (Ed). *Organization and representation in perception*. Hillsdale, NJ: Lawrence Erlbaum Associates, (pp. 167–190).
- Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. In *Annual Review of Psychology*.
- Kimia, B., Frankel, I., & Popescu, A. (2003). Euler spiral for shape completion. *International Journal of Computer Vision*, 54(1/2), 157–180.
- Kira, K., & Rendell, L. A. (1992). A practical approach to feature selection. In *The 9th International Conference on Machine Learning*.
- Lin, L., Wang, X., Yang, W., & Lai, J. (2012). Learning contour-fragment-based shape model with and-or tree representation. In *IEEE Conference Computer Vision and Pattern Recognition*.
- Liu, H., Liu, W., & Latecki, L. J. (2010). Convex shape decomposition. In *IEEE Conference on Computer Vision and Pattern Recognition*.
- Lu, C., Latecki, L. J., Adluru, N., Yang, X., & Ling, H. (2009). Shape guided contour grouping with particle filters. In *Proceedings of IEEE International Conference on Computer Vision*.
- Luo, P., Lin, L., & Chao, H. (2010). Learning shape detector by quantizing curve segments with multiple distance metrics. In *European Conference on Computer Vision*.
- Ma, T., & Latecki, L. J. (2011). From partial shape matching through local deformation to robust global shape similarity for object detec-

- tion. In *IEEE Conference on Computer Vision and Pattern Recognition*.
- Maji, S., & Malik, J. (2009). A max-margin hough transform for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*.
- Martin, D., Fowlkes, C., & Malik, J. (2004). Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 26(5), 530–549.
- Mikolajczyk, K., Schmid, C., & Zisserman, A. (2004). Human detection based on a probabilistic assembly of robust part detectors. In *European Conference on Computer Vision*.
- Ommer, B., & Malik, J. (2009). Multi-scale object detection by clustering lines. In *International Conference on Computer Vision*.
- Opelt, A., Pinz, A., & Zisserman, A. (2008). Learning an alphabet of shape and appearance for multi-class object detection. *International Journal of Computer Vision*, 80(1), 45–57.
- Felzenszwalb, P., McAllester, D., & Girshick, R. (2010). Cascade object detection with deformable part models. In *IEEE Conference on Computer Vision and Pattern Recognition*.
- Ravishanker, S., Jain, A., & Mittal, A. (2008). Multi-stage contour based detection of deformable objects. In *European Conference Computer Vision*.
- Renninger, L. K., Verghese, P., & Coughlan, J. (2007). Where to look next? Eye movements reduce local uncertainty. *Journal of Vision*, 7(3), 1–17.
- Rensink, R. A., & Enns, J. T. (1998). Early completion of occluded objects. *Vision Research*, 38, 2489–2505.
- Riemenschneider, H., Donoser, M., & Bischof, H. (2010). Using partial edge contour matches for efficient object category localization. In *European Conference Computer Vision*.
- Sala, P., & Dickinson, S. (2010). Contour grouping and abstraction using simple part models. In *European Conference on Computer Vision*.
- Schneiderman, H., & Kanade, T. (2004). Object detection using the statistics of parts. *International Journal of Computer Vision*, 60(2), 135–164.
- Schnitzspan, P., Roth, S., & Schiele, B. (2010). Automatic discovery of meaningful object parts with latent CRFs. In *IEEE Conference on Computer Vision and Pattern Recognition*.
- Sharvit, D., Chan, J., Tek, H., & Kimia, B. B. (1998). Symmetry-based indexing of image databases. *Journal of Visual Communication and Image Representation*, 9(4), 366–380.
- Shawe-Taylor, J., & Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge: Cambridge University Press.
- Shotton, J., Blake, A., & Cipolla, R. (2008). Multi-scale categorical object recognition using contour fragments. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*.
- Srinivasan, P., Zhu, Q., & Shi, J. (2010). Many-to-one contour matching for describing and discriminating object shape. In *IEEE Conference Computer Vision and Pattern Recognition*.
- Sukumar, S.R., Page, D. L., Koschan, A. F., Gribok, A. V., & Abidi, M. A. (2006). Shape measure for identifying perceptually informative parts of 3d objects. In *Proceeding of 3rd International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT)*.
- Toshev, A., Taskar, B., & Daniilidis, K. (2012). Shape-based object detection via boundary structure segmentation. *International Journal of Computer Vision (IJCV)*, 99(2), 123–146.
- Ullman, S. (2007). Object recognition and segmentation by a fragment based hierarchy. *Trends in Cognitive Sciences*, 11, 58–64.
- Wang, X., Bai, X., Ma, T., Liu, W., & Latecki, L. J. (2012). Fan shape model for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*.
- Yarlagadda, P., & Ommer, B. (2012). From meaningful contours to discriminative object shape. In *European Conference on Computer Vision*.
- Yarlagadda, P., Monroy, A., & Ommer, B. (2010). Voting by grouping dependent parts. In *European Conference on Computer Vision*.
- Zhu, L., Chen, Y., Yuille, A., & Freeman, W. (2010). Latent hierarchical structural learning for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*.
- Zhu, Q., Wang, L., Wu, Y., & Shi, J. (2008). Contour context selection for object detection: A set-to-set contour matching approach. In *European Conference on Computer Vision*.
- Zhu, S. C., Wu, Y. N., & Mumford, D. B. (1998). Filters, random field and maximum entropy(frame): Towards a unified theory for texture modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*.