A 2-Fan Shape Model for Object Detection and Localization

Yuanqi Su†, Yuehu Liu‡, Xiao Huang‡ and Nanning Zheng‡

†Xi'an Jiaotong University, Xi'an, China
‡Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, China
Email: yuanqisu@mail.xjtu.edu.cn, liuyh@mail.xjtu.edu.cn

Abstract—For its simplicity and effectiveness, star model is popular in part-based object detection and localization. It finds the candidates for parts and then vote for object's configurations. However, it suffers from the loose geometric constraints since it neglects the connection among parts. In the paper, we reconsidered the connections and reduce the task of object detection and localization to a shape matching problem. We propose to use a global shape as a constraint to modify the star model. The resulting representation is then a 2-fan model that is subsequently optimized by searching for the candidates of parts and solving for the global constraint. Though the procedure is easily stuck at a local optimum, results show that is gives a comparable result with the state-of-art methods.

1. Introduction

Many problems in computer vision concerns matching a given shape template. It has served as the core part in many visual applications, such as object detection[8, 10, 12, 11], human pose estimation[9] and so on. On the other hand, shape matching is yet challenging; occlusion, cluttered backgrounds, and the non-rigid deformation all make it a relatively hard task.

A shape is usually represented by a set of contours; each is a point sequence $U : \{\mathbf{u}(t) = (xu(t), yu(t))'\}$ where t is the index starting from 1. For simplicity, the discussion focuses on U with a single sequence, and extension to that with multiple ones is straightforward. Throughout the paper, we use an *uppercase* letter for planar shape, the corresponding *bold lowercase* letter for its point and $|U|$ for the length. For example, the last point of planar shape U is $\mathbf{u}(|U|)$, whose x coordinate is $xu(|U|)$, and y coordinate is $yu(|U|)$.The second shape V comes from the given image. It contains a set of contour fragments. To distinguish the fragments in V, we introduce a vector $\mathbf{b}$ that stores the index range. The range for cth fragment is $[b_c, b_{c+1})$. We give an example of V in Fig.1, where fragments are discriminated by color.

When both U and V are represented by the contour fragments, shape matching is then to select a subset of contour fragments from V which can best explain the given U. However, selecting the contour fragments involves a combinatorial optimization[10, 12] that is usually NP-hard. To get rid of combinatorial explosion, we propose to localize the parts of a hypothesis via the local matching and use interconnection to force the selected parts along the same or adjacent contours, resulting in a 2-Fan shape model.

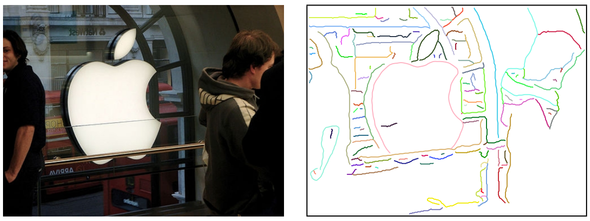


Figure 1: Linked contour fragments and the original image from ETHZ dataset[5].

The model is then formulated, optimized and analyzed in the following sections. In Section 2, we derive a brief reivew on the related works, then formulate the shape matching based on 2-fan shape model in Section 3. The subsequent Section 4 is then devoted to the optimization. Section 5 presents the experimental evaluations and analysis followed by conclusions drawn in Section 6.

2. Related Works

The proposed representation belongs to part-based model. Unlike the methods that treat the shape as a whole[2, 9], it represents the shape as a collection of parts. A part-based model usually involves the description of the local parts, the definition of connections among parts, and the use of bottom-up heuristics.

Taking the generalized hough transform (GHT)[1] for example, it is the pioneering work in part-based shape matching. In GHT, each boundary point corresponds to a part, that is described by the edge orientation. Parts are connected to a reference point, resulting in a star model.

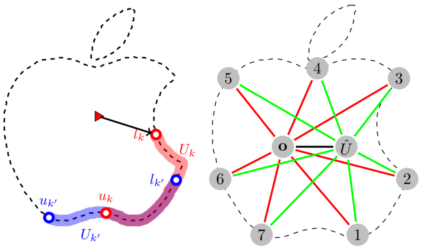
After decades of development, the star model used by GHT is still among the most popular ones for part-based shape matching. However, its voting style is substituted by a minimizing way, named the Hough-like voting. Usually, the voting scheme implies that an object is represented by a set of parts, and for each instance of the object, a part only accepts the candidate with minimum error. Model built by the assumption includes the implicit shape model (ISM)[8] and is adopted by many recent works[8, 13, 12]. Ferrari *et al.*[6, 5] proposed a family of scale invariant local shape descriptors which linked groups of adjacent segments of contour. The descriptors are then combined with the Hough-style voting scheme for object detection[5]. Praveen *et al.*[12] used shape context[3] to describe the part, and adopted the Hough-like voting for the hypotheses.

As for part description, there are many choices. Popular ones include shape context[3], distance transform[2] including the oriented chamfer distance[13, 9] and so on. A detailed survey can be found in [15].

The third aspect of part-based shape matching is about the bottom-up heuristics. The first heuristics that should be mentioned is the use of contour fragments. Majority of the the recent methods[12, 9, 6, 5] manipulate the contour fragments instead of edge points. Based on contour fragments, Ferrari *et al.*[6, 5] propose to use the spatial adjacency and Praveen *et al.*[12] requires that a contour fragment either belongs to an instance or is irrelevant to it.

3. Problem Formulation

The idea that lies that at the core of our methodology is to resolve the partial correspondence between both shapes. We split U into a set of overlapped pieces as shown in Fig.2(a). The kth piece is denoted by U_k, whose index range is $\mathcal{R}_k = [l_k, h_k]$. We select the start point of each piece as the anchor point and move the mass point of U to the origin.



(a) (b)

Figure 2: (a).Representation of U, (b).2-fan shape model.

The task that matching U to V should determine U's scale, position and orientation in the image. The combination of scale, position and orientation is referred to as the pose of an object.

In practice, we discretize the parameter space for object poses through sampling a set of scaling factors and rotation angles and using the grid locations of the image domain for the object center $\mathbf{o}$. For the sake of brevity, we neglects the influence of scale and rotation, and assumes that both factors have been excluded by scaling and rotating V correspondingly. Then the transformed coordinate is simplified as $\mathbf{u} + \mathbf{o}$ where $\mathbf{u}$ is a point from U. Next, we assume that the object center $\mathbf{o}$ is known, and use U for $U(\mathbf{o})$.

The partial correspondence between the kth piece of U and V is described by a function ω_k. For convenience, we call ω_k the selection function. It has the same index range as U_k, and produces a sequence from V,

$$\mathbf{v}(\omega_k(l_k)), \cdots, \mathbf{v}(\omega_k(h_k)). \qquad (1)$$

A feasible ω_k then determines a matching candidate for piece U_k.

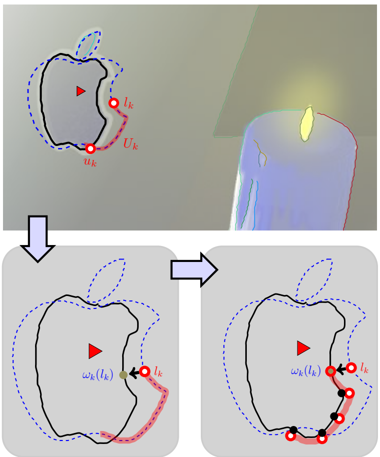


Figure 3: The two operations for warping U_k to V.

With the selection function, we define a dissimilarity measure that is derived from warping U_k to the point sequence in Eq.1. The warping undergoes two sequential operations. The first one displaces the segment such that its anchor point is aligned with $\mathbf{v}(\omega_k(l_k))$. The second one operates on the displaced segment, registering its points to the counterparts in Eq.1.

Energy for displacing the segment is given by,

$$E_{k,g}(\omega_k, \mathbf{o}) = \lambda_g |\mathcal{R}_k| \cdot \|\mathbf{v}(\omega_k(l_k)) - \mathbf{u}(l_k) - \mathbf{o}\|^2. \qquad (2)$$

where $\mathbf{v}(\omega_k(l_k)) - (\mathbf{u}(l_k) + \mathbf{o})$ is the offset required for aligning the two anchor points, and the unit λ_g changes the squared moving distance to the energy. After the operation, we get the displaced piece of U_k with each of its points displaced to $\mathbf{u}(t) - \mathbf{u}(l_k) + \mathbf{v}(\omega_k(l_k))$. Energy for the second operation calculates the total distance between the displaced piece and that in Eq.1.

$$E_{k,l}(\omega_k) = \sum_{t=l_k}^{u_k} \|\mathbf{u}(t) - \mathbf{u}(l_k) - \mathbf{v}(\omega_k(t)) + \mathbf{v}(\omega_k(l_k))\|_\rho \,. \qquad (3)$$

The metric $\|\cdot\|_\rho$ is the truncated Euclidean distance: $\|\cdot\|_\rho = \min(\|\cdot\|, \rho)$, where ρ bounds the Euclidean distance from above. We adopt the metric for suppressing noises and removing the outliers. In addition, we require that ω_k preserves the order of contour fragments.

The isolated way for defining selection functions breaks the interconnections among the overlapped segments. Therefore, we introduce an auxiliary shape $\hat{U}$ that connect the parts via the energy in Eq.4.

$$Z(\{\omega_k\}, \hat{U}) = \sum_k \sum_{t=l_k}^{h_k} \|\hat{\mathbf{u}}(t) - \mathbf{v}(\omega_k(t))\|^2. \qquad (4)$$

The auxiliary shape $\hat{U}$ has the same domain as U.Without the prior information, it is reasonable to ensure that shape

$\hat{U}$ is smooth. Assuming that the selection function for each segment is known, an optimal $\hat{U}$ can be inferred by fitting a smooth contour.

The objective function measures the similarities of all pieces, at the same times, penalizes the inconsistent case.

$$O(\{\omega_k\}, \hat{U}, \mathbf{o}) = \sum_k E_k(\omega_k, \mathbf{o}) + \lambda_s Z(\{\omega_k\}, \hat{U}), \tag{5}$$

where λ_s denotes the weight of the smoothing term.

With the graph representation, we get a 2-fan shape model as shown in Fig.2(b). In the model, parts are not connected to each other directly, while they are all linked to the pose parameter $\mathbf{o}$ and auxiliary shape $\hat{U}$. Given $\hat{U}$ and $\mathbf{o}$, parts are conditionally independent, and their selection functions can be optimized one by one with dynamic programming.

4. Dynamic Programming For Selection Function

In objective function, the terms involving ω_k are summarized as,

$$O_k(\omega_k, \hat{U}, \mathbf{o}) = E_{k,l}(\omega_k) + E_{k,g}(\omega_k, \mathbf{o}) + \lambda_s Z_k(\omega_k, \hat{U}) \tag{6}$$

where, $Z_k(\omega_k, \hat{U}) = \sum_{t=l_k}^{h_k} \|\hat{\mathbf{u}}(t) - \mathbf{v}(\omega_k(t))\|^2$. Fixed $\hat{U}$, the evaluation of ω_k is independent of others, hence, ω_k can be optimized on its own. Since $E_{k,g}(\omega_k, \mathbf{o})$ only depends on $\omega_k(l_k)$, thus the optimization for ω_k can be further decomposed shown as follows.

$$\min_j E_{k,g}(j, \mathbf{o}) + \min_{\omega_k \in \Omega_j} \left(E_{k,l}(\omega_k) + \lambda_s Z_k(\omega_k, \hat{U}) \right), \tag{7}$$

where j takes value from the indexes of V, $E_{k,g}(j, \mathbf{o}) = \lambda_g |\mathcal{R}_k| \cdot \|\mathbf{v}(j) - \mathbf{u}(l_k) - \mathbf{o}\|^2$; and Ω_j is the space of order preserved ω_k starting with j. The minimization with respect to j can be implemented by traversing the indecies of V, thus the left task is to optimize $E_{k,l}(\omega_k) + \lambda_s Z_k(\omega_k, \hat{U})$ with respect to $\omega_k \in \Omega_j$.

According to the definition of order preserving, the selection function goes either decreasingly or increasingly. The property ensures that the minimization can be solved by dynamic programming. For fast implementation, we restrict the optimization along the contour containing the starting point $\mathbf{v}(j)$, demanding that the selection function ω_k selects points from the same contour.

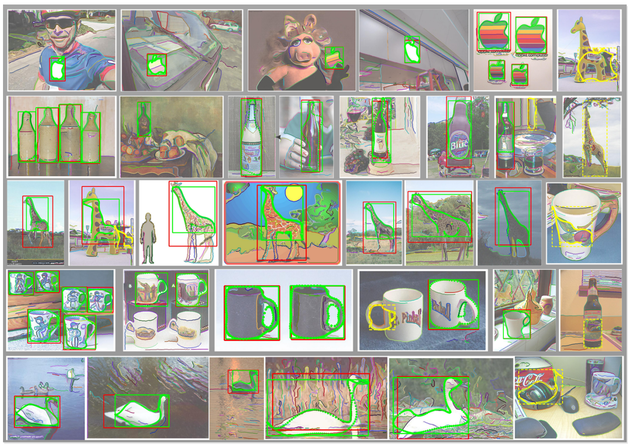
5. Experimental Evaluations

Experiments were conducted on the ETHZ shape dataset[5]. There are 255 test images from 5 categories in the dataset, and we followed the same experimental setup in [9] and used a single shape to detect and localized its instances. We use the method proposed in [14] to initialize a set of poses. For each pose, we refine the auxiliary shape and the partial correspondence in turn. Parameters for the matching were determined empirically; in fact, our method works for a wide range of parameters.

We compared our algorithm against method of oriented chamfer matching (OCM) [13], works by Ferrari *et al.*[6, 5], and fast direction chamfer distance matching (FDCM) by Ming-Yu *et al.*[9]. We show the false positive per image (FPPI) vs. detection rate (DR) in Fig.4. Our method achieved better performance than the others. Besides, some localization results are shown in Fig.5.

Table 1: Comparison for the detection rate for 0.3/0.4 FPPI on ETHZ shape classes.0 and 1 are the value of λ_s.

	Applelogos	Bottles	Giraffes
ours(0)	0.909/ 0.932	0.982/**1**	0.484/0.495
ours(1)	**1/ 1**	**1/1**	0.615/0.637
Praveen[12]	0.95/0.95	**1/1**	0.872/**0.896**
Maji[11]	0.95/0.95	0.929/0.964	**0.896/0.896**
Felz[4]	0.95/0.95	**1/1**	0.729/0.729
Ferrari[5]	0.777/0.832	0.798/0.816	0.399/0.445
Gu[7]	0.906/-	0.948/-	0.798/-
	Mugs	Swans	Mean
ours(0)	0.742/0.788	0.97/**1**	0.817/0.843
ours(1)	0.864/0.879	**1/1**	0.896/0.903
Praveen[12]	**0.936**/0.936	**1/1**	**0.952/0.956**
Maji[11]	**0.936/0.967**	0.882/0.882	0.919/0.932
Felz[4]	0.839/0.839	0.588/0.647	0.821/0.833
Ferrari[5]	0.751/0.8	0.632/0.705	0.671/0.72
Gu[7]	0.832/-	0.868/-	0.871/-



(a)

Figure 5: The localization results. (For top to bottom row: ”Applelogos”, ”Bottles”, ”Giraffes”, ”Mugs”, ”Swans”.) Last column shows some examples of false positives.

Besides, the detection rate (DR) at 0.3/0.4 FPPI is reported in Table.1. Our detection rates for ’Applelogos’, ’Bottles’, and ’Swans’ are the best among the comparison methods; for ’Mugs’, it is slight inferior, and is inferior for ’Giraffes’. Proposed method is only defers to the method

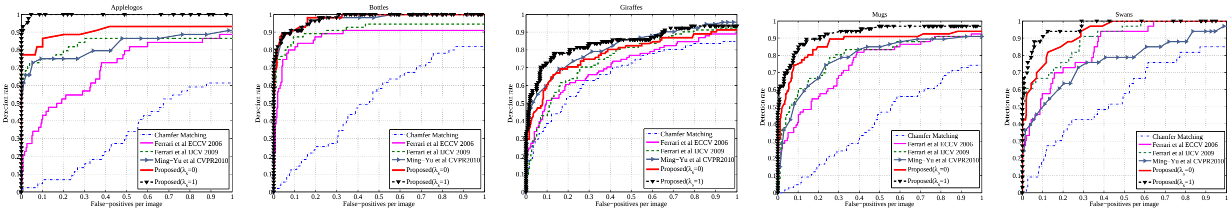


Figure 4: The FPPI vs DR curves for ETHZ dataset. All methods for comparison all used the 0.2 overlapping ratio.

of Praveen *et al*[12] and Maji *et al.*[11], which used half of the positive samples for training, and tested on the residual samples.

6. Conclusions

We proposed a 2-fan model for shape matching that makes the parts depend on both the pose parameter and a global auxiliary shape. Use of the auxiliary shape connects the parts, and makes the partial matching more resonable, that is verified by the experiments. In the further, we will extend the model to object segmentation, since auxiliary shape give a rough approximation to the object boundary.

Acknowledgments

The authors would like to thank NSFC for support under the grant number 91120009, 61328303 and 61305051.

References

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

[2] H. G. Barrow, J. M. Tenenbaum, R. C. Bolles, and H. C. Wolf. Parametric correspondence and chamfer matching: two new techniques for image matching. In *IJCAI'77: Proceedings of the 5th international joint conference on Artificial intelligence*, pages 659–663, San Francisco, CA, USA, 1977. Morgan Kaufmann Publishers Inc.

[3] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:509–522, 2002.

[4] Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:1627–1645, 2010.

[5] V. Ferrari, F. Jurie, , and C. Schmid. From images to shape models for object detection. *International Journal of Computer Vision*, 87(3):284–303, 2010.

[6] Vittorio Ferrari, Tinne Tuytelaars, and Luc Van Gool. Object detection by contour segment networks. In AleÅ Leonardis, Horst Bischof, and Axel Pinz, editors, *Computer Vision ECCV 2006*, volume 3953 of *Lecture Notes in Computer Science*, pages 14–28. Springer Berlin / Heidelberg, 2006.

[7] Chunhui Gu, J.J. Lim, P. Arbelaez, and J. Malik. Recognition using regions. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:1030–1037, 2009.

[8] Bastian Leibe, Ales Leonardis, and Bernt Schiele. Combined object categorization and segmentation with an implicit shape model. In *In ECCV workshop on statistical learning in computer vision*, pages 17–32, 2004.

[9] Ming-Yu Liu, O. Tuzel, A. Veeraraghavan, and R. Chellappa. Fast directional chamfer matching. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1696 –1703, 2010.

[10] C.E. Lu, L.J. Latecki, N. Adluru, X.W. Yang, and H.B. Ling. Shape guided contour grouping with particle filters. In *ICCV09*, pages 2288–2295, 2009.

[11] S. Maji and J. Malik. Object detection using a max-margin hough transform. pages 1038 –1045, jun. 2009.

[12] Qihui Zhu Praveen Srinivasan and Jianbo Shi. Many-to-one contour matching for describing and discriminating object shape. In *CVPR2010*, 2010.

[13] J. Shotton, A. Blake, and R. Cipolla. Multiscale categorical object recognition using contour fragments. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(7):1270 –1281, 2008.

[14] Yuanqi Su.and Yuehu Liu. A voting scheme for partial object extraction under cluttered environment. *Internatiional Journal of Pattern Recognition and Artificial Intelligence*, 27(2):1–37, 2013.

[15] Dengsheng Zhang and Guojun Lu. Review of shape representation and description techniques. *Pattern Recognition*, 37(1):1 – 19, 2004.