JOINT LEARNING OF FOREGROUND REGION LABELING AND DEPTH ORDERING

Young-Joo Seo, Jongmin Kim, Hoyong Jang, Tae-Ho Kim and Chang D. Yoo

Department of EE, Korea Advanced Institute of Science & Technology
youngjoo_seo, waterboys, hoyong.jang, ktho22@kaist.ac.kr, cdyoo@ee.kaist.ac.kr

ABSTRACT

This paper considers a joint learning algorithm of foreground region labeling and depth ordering for 3D scene understanding. Given an object-level segmentation, the proposed algorithm classifies each region as either foreground or background while simultaneously infers the relative depth orders between every adjacent region pairs. For this, we consider a graph where regions are considered as nodes while boundaries between adjacent regions as edges, and the problem is formulated as jointly assigning binary labels to every nodes and edges via maximizing a unified linear discriminant function, under the constraints that make the resulting depth order to be always physically plausible. Instead of inferring region and edge labels separately, we infer them jointly by grouping them as a single variable referred to as *triplet*. Then, the problem is reformulated as multi-class triplet prediction to penalize the inconsistent labeling of regions and edges in a soft manner. As the discriminant function is linear, the parameters can be learned with structured support vector machine(S-SVM), and efficient inference using linear programming relaxation is possible. Experimental results show that the proposed joint inference algorithm improves both foreground region labeling and depth ordering performances.

Index Terms— Figure/ground, Depth ordering, Foreground region labeling

1. INTRODUCTION

Depth estimation is one of the great challenges of computer vision. For the past years, the study of depth estimation has been focused on inferring the exact depth value based on multi-view images [21] and motion cues[2, 17]. However, with only a single image, this task is difficult. Fortunately, a single image depth cues such as occlusion and geometric information make it possible to infer the relative depth-order among the objects. It has been reported that relative depth information is very useful to handle high-level vision tasks, such as salient object detection [15, 8] and 3D reconstruction for scene understanding [6, 9].

Most previous approaches have been focused on figuring out which side owns the boundary by using variety of local cues from the contour and the T-junction structure. This cues are determined by convexity, lower region, parallelism, etc. [7, 18, 16, 14, 12, 5] However, estimating the depth order from junction or boundary has natural flaws without considering characteristic of foreground-background configuration. One of the trivial characteristic of relative depth order is that background region such as sky, ground, etc., is always located at backmost of the image. Geometric information of a certain region can help to understand the relative depth configuration between adjacent regions. Meanwhile, when estimating the region label, the pairwise relative depth order between adjacent regions enforces its region label to be inferred correctly. Thus, depth ordering and foreground region labeling create synergy when they are conducted simultaneously.

Related works: Among the previous works for depth ordering, Hoiem *et al* [9] has recieved considerable attention. They analyze the effect of surface layout confidences on inferring relative depth order among the objects. The performance get a significant improvement with the help of geometric confidence cues. However, once geometric confidence is estimated, the result of relative depth ordering strongly depends on the geometric confidence accuracy. As aforementioned, relative depth ordering and region labeling can assist each other, so the performance of each task is expected to be improved when the two tasks are conducted jointly.

Contribution: To address this issue, we adopt joint learning framework to reason about depth ordering and foreground region labeling with the hope of enhancing the performance by sharing correlated information between them while maintaining physically reasonable configuration. The problem can be interpreted as assigning binary labels on every boundary and region. It is naturally formulated as an integer programming with the constraints that make the predicted labels to conserve the global consistency between depth order and region labels. However, it is difficult to design such constraints between region and boundary labels as linear inequalities, which makes LP not available. To cope with the above problem, we introduce an algorithm where we group three variables correspond to an adjacent region pairs and a boundary in between, and jointly assign binary labels to them. For this, we define a *triplet* variable that consists of above three variables, and reformulate the problem as multi-class triplet prediction on ev-

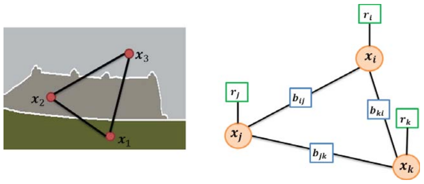

Fig. 1. Undirected graph on segmentation.

ery boundary. Instead of manually defining hard-constraints between region and boundary labels, the proposed algorithm with triplet prediction allows us to automatically enforce the consistency existing in the training data during the learning process. The labeling process is formulated as maximizing a linear discriminant function which can be solved efficiently by LP relaxation. In addition, the linear discriminant function takes an advantage of large-margin parameter training using structured support vector machine(S-SVM) [20]. It efficiently learns optimal parameters for multi-class triplet classifier offering good generalization.

2. JOINT FRAMEWORK VIA MULTICLASS TRIPLET PREDICTION

Given a segmented image that conserves occlusion boundary, the proposed algorithm simultaneously infers physically-plausible-depth-ordering and foreground-region-labeling. For this, we consider an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ as shown in Fig. 1. Here the nodes are the regions and the edges are the depth order which needs to be predicted. From a segmented image $\mathbf{X}$ composed of N regions $\{x_i\}_{i=1}^{N}$ and boundaries between regions such that $\mathcal{E} = \{(i,j)|\forall j \in \mathcal{N}(i) \wedge i = 1,\dots,N\}$, the relative depth order $\mathbf{B} = \{b_{ij}|\forall(i,j) \in \mathcal{E}\} \in \{0,1\}^{|\mathcal{E}|}$ and the region label $\mathbf{R} = \{r_i|i = 1,\dots,N\} \in \{0,1\}^{N}$ need to be determined such that $b_{ij} = 0$ when x_i is located in front of x_j; otherwise, $b_{ij} = 1$. When x_i is in the background then $r_i = 0$; otherwise, $r_i = 1$. A linear discriminant function to measure the quality of the depth-order and the geometric relationship in an image is defined as follows:

$$\begin{aligned} F(\mathbf{X},\mathbf{B},\mathbf{R};\mathbf{w},\mathbf{v}) &= \sum_{(i,j)\in\mathcal{E}} D(x_i,x_j;\mathbf{w})b_{ij} + \sum_{i=1}^{N} E(x_i;\mathbf{v})r_i \\ &= \sum_{(i,j)\in\mathcal{E}} \langle \mathbf{w}, \phi_{ij}(x_i,x_j)\rangle b_{ij} + \sum_{i=1}^{N} \langle \mathbf{v}, \psi_i(x_i)\rangle r_i \\ &= \langle \boldsymbol{\theta}, \Phi(\mathbf{X},\mathbf{B},\mathbf{R})\rangle \end{aligned} \qquad (1)$$

Here, $D(x_i,x_j;\mathbf{w})$ is parameterized by $\mathbf{w}$ and measures which region is front such that a large negative value means region x_i is front while a large positive value means region x_j is front. while $E(x_i;\mathbf{v})$ is parameterized by $\mathbf{v}$ and becomes negative when x_i is foreground, positive when x_i is background region. ϕ_{ij} and ψ_i represent the edge feature and region feature vector respectively.

segment		Relative depth order b_{ij}							
x_i	x_j	**inval**	val	val	val	val	val	val	**inval**
x_1	x_2	**0**	0	0	0	1	1	1	**1**
x_2	x_3	**0**	0	1	1	0	0	1	**1**
x_3	x_1	**0**	1	0	1	0	1	0	**1**

Table 1. Label validity for cyclic depth order from the graph in Fig. 1. '0' means i_{th} segment is front and '1' means j_{th} segment is front.

Each binary label in a given graph must satisfy the physically-plausible condition. As shown in Table 1, there exists valid cases for proper figure/ground relationship between three adjacent regions for physically-plausible depth ordering. The constraint between edge labels b_{ij} can be mathematically formulated as follows:

Definition 1 (Valid depth order relationship) *Given a junction $\mathcal{J}$ composed of three nodes x_i, x_j and x_k such that any of its pair is a nearest neighbors of one another, in other words $\mathcal{J} = \{(i,j),(j,k),(k,i)\} \subset \mathcal{E}$ then the following cycle inequalities on the boundary labels for all junctions $\mathbf{J}$ should be satisfied as follows :*

$$1 \leq \sum_{(i,j)\in\mathcal{J}} b_{ij} \leq 2, \quad \forall\mathcal{J} \in \mathbf{J}. \qquad (2)$$

Aside from the constraint formulated in Eq.(2), region labels must be consistent with the relative depth order such that a foreground region remains in front of a background region. Therefore, constraint on $\mathbf{R}$ must also be considered to be consistent with Eq.(2). Therefore $\mathbf{R}$ and $\mathbf{B}$ must be jointly estimated. Instead of training separate classifiers for region and edge labeling, we train a single classifier which maps a *joint feature* to a joint label for $\mathbf{R}$ and $\mathbf{B}$. For this, we define a *triplet* label set $\mathbf{T} = \{t_{ij} = (b_{ij}, r_i, r_j)|\forall i,j \in \mathcal{E}\}$. Since both r_i and b_{ij} are binary, t_{ij} can take on 8 different values, and its inference can be considered as a multi-class classification problem with eight classes. Now, a discriminant function involving t_{ij} is defined as follows:

$$\begin{aligned} F(\mathbf{X},\mathbf{T};\mathbf{W}) &= \sum_{(i,j)\in\mathcal{E}} U(t_{ij};x_i,x_j,\mathbf{W}) \\ &= \sum_{(i,j)\in\mathcal{E}} \langle \mathbf{W}, \phi_{ij}^{joint}(x_i,x_j)\rangle. \end{aligned} \qquad (3)$$

Here, $U(t_{ij};x_i,x_j,\mathbf{W})$ is a linear discriminative function that is parameterized by $\mathbf{W}$. The joint feature ϕ^{joint} is a concatenation of edge and region features such that $\phi_{ij}^{joint} = [\phi_{ij}, \psi_i, \psi_j]$.

Since the depth ordering must be physically plausible, the constraints over triplet variables should obey Eq.(2). Therefore, the optimization problem for finding the optimal triplet labels can be formulated as follows:

$$\begin{aligned}&\mathbf{T}^{*} = \arg\max_{\mathbf{T}\in\mathcal{T}} F(\mathbf{X},\mathbf{T};\mathbf{W}),\\ &s.t.\;\; 1 \leq \sum_{(m,n)\in\mathcal{J}} b_{mn}(t_{mn}) \leq 2,\;\; \forall\mathcal{J}\in\mathbf{J}\end{aligned} \tag{4}$$

Here,

$$b_{mn}(t_{mn}) = \begin{cases} 0 & \text{if } t_{mn} \geq 4 \\ 1 & \text{otherwise} \end{cases} \tag{5}$$

By LP-relaxation, Eq.(4) can be solved as follows:

$$\begin{aligned}&Z^{*} = \arg\max_{Z} \sum_{(i,j)\in\mathcal{E}} \sum_{t_{ij}=0}^{7} \langle \mathbf{W}, \phi_{ij}^{joint}(x_i,x_j)\rangle z_{ij}^{t_{ij}},\\ &s.t. \sum_{t_{ij}=0}^{7} z_{ij}^{t_{ij}} = 1, \quad 0 \leq z_{ij}^{t_{ij}} \leq 1,\\ &1 \leq \sum_{(m,n)\in\mathcal{J}} \left[\sum_{t_{mn}\geq 4} z_{mn}^{t_{mn}} + \sum_{t_{mn}<4} (1 - z_{mn}^{t_{mn}}) \right] \leq 2,\\ &\forall\mathcal{J}\in\mathbf{J}, t_{ij} = 0, ..., 7.\end{aligned} \tag{6}$$

Here,

$$Z = \begin{bmatrix} z_{12}^{0} & \cdots & z_{12}^{7} \\ \vdots & \ddots & \vdots \\ z_{N-1,N}^{0} & \cdots & z_{N-1,N}^{7} \end{bmatrix}. \tag{7}$$

When the optimal Z^{*} is inferred tightly, Z^{*} is the optimal triplet label $\mathbf{T}^{*}$. Since there is one-to-one correspondance between t_{ij} and b_{ij}, $\mathbf{B}^{*}$ is directly determined from $\mathbf{T}^{*}$. However, it is not the case for $\mathbf{R}$ since each region attends multiple triplet variables. In this paper, the region label $\mathbf{R}^{*}$ is determined by majority voting scheme. Although it is heuristic, we observed that this voting scheme is empirically effective for region labeling without much performance degradation.

The triplet-based algorithm discussed so far has two main advantages over the previous one: *first*, instead of explicitly defining hard-constraints between region and edge labels, which is very difficult, our approach enforces their consistency in a soft manner, which is much easier; during training multi-class triplet classifier, the cases of inconsistently-labeled regions and edges are penalized since they do not occur in the training dataset. *second*, feature sharing effect; the region feature affects the edge label, while the edge feature affects the region label.

3. MAX-MARGIN TRAINING VIA STRUCTURED-SVM

In this section, a large-margin training based on structured-SVM is described. To estimate the parameter vector $\mathbf{W}$ of the linear discriminant function $F(\mathbf{X},\mathbf{T};\mathbf{W})$, the following constrained-optimization problem referred to as margin scaling [10, 22, 3] is solved as follows:

$$\begin{aligned}&\min_{\mathbf{w},\boldsymbol{\xi}} \; \frac{1}{2}||\mathbf{W}||^{2} + \frac{C}{N}\sum_{n=1}^{N}\xi_n,\\ &\text{s.t. } d(\mathbf{X}_n,\mathbf{T};\mathbf{W}) \geq \Delta(\mathbf{T}_n,\mathbf{T}) - \xi_n,\;\; \mathbf{T}\in\mathcal{T}\setminus\mathbf{T}_n,\;\forall n,\\ &\xi_n \geq 0,\;\forall n.\end{aligned} \tag{8}$$

Here, $d(\mathbf{X}_n,\mathbf{T};\mathbf{W})$ is the difference of the discriminant function values between the ground-truth label $\mathbf{T}_n$ and the predicted label $\mathbf{T}$ such as

$$d(\mathbf{X}_n,\mathbf{T};\mathbf{W}) = F(\mathbf{X}_n,\mathbf{T}_n;\mathbf{W}) - F(\mathbf{X}_n,\mathbf{T};\mathbf{W}). \tag{9}$$

Here, ξ_n is a slack variable to allow training error for $\mathbf{X}_n$, and C is the balance coefficient to trade-off between the training error minimization and the margin maximization. The loss function $\Delta(\mathbf{T}_n,\mathbf{T})$ is an error measurement of predicting a label $\mathbf{T}$ given the correct label $\mathbf{T}_n$. To overcome the unbalance problem, a modified Hamming loss [13] is used.

The optimization problem of Eq. (8) has exponential number of constraints with respect to the dimensionality of $\mathbf{T}$. Thus, the cutting-plane algorithm [19, 11] is used to reduce the number of constraints. In the algorithm, the most violated label for n_{th} training data is inferred as

$$\bar{\mathbf{T}}_n = \arg\max_{\mathbf{T}\in\mathcal{T}/\mathbf{T}_n} \left[\Delta(\mathbf{T}_n,\mathbf{T}) - d(\mathbf{X}_n,\mathbf{T};\mathbf{W})\right], \tag{10}$$

then added to the constraint set. Note that the considered loss function is decomposable over the test edges for efficient inference of $\bar{\mathbf{T}}_n$ in Eq. (10). Given the constraint set, the optimization problem can be solved using quadratic programming(QP).

4. EXPERIMENTS

Dataset : To compare previous algorithm, proposed algorithm is evaluated on two dataset : Geometric Context Dataset [9] and D-order dataset [12]. We used a half of images as training set and the others as test set and evaluate our algorithm.

Features : For foreground region labeling features $\psi_i(x_i)$, a 52-dimensional low-level features [4] such as color, texture, location, and shape features are extracted from each region. In addition, we use 150 dimensional visual word features [1] that representing the posterior probability of each region belonging. For depth ordering features $\phi_{ij}(x_i,x_j)$, we design geometric cues (4 dim), Convexity cues (2 dim), Position/location cues (2 dim) and saliency cues (27 dim).

Results : To evaluate the effectiveness of the proposed algorithm, we compare the following three methods for two tasks : depth ordering and foreground region labeling.

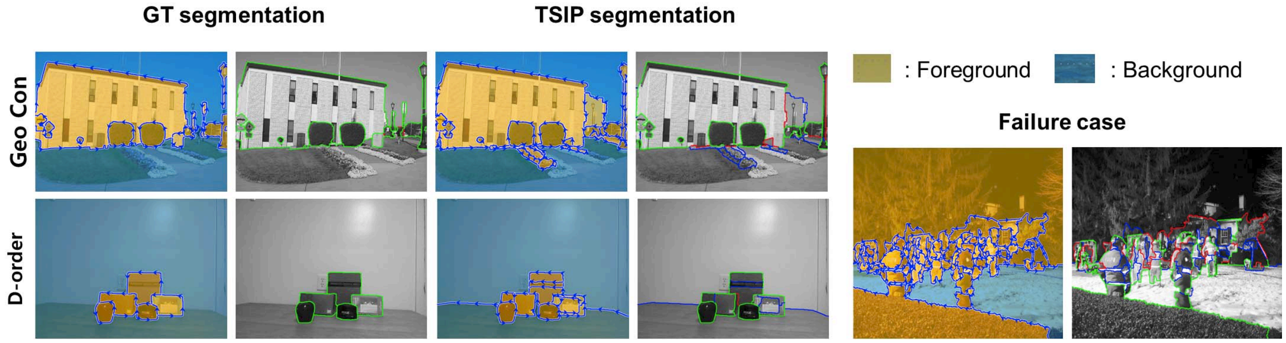

Fig. 2. Qualitative results on benchmark dataset. In gray image, the green boundary denotes correct, red denotes incorrect, and blue denotes unmatched boundaries.

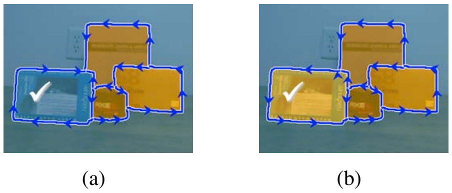

Fig. 3. Global consistency on depth ordering and region labeling. The region on the left side of the arrow is thought to be in front. The marked miss-classified segment in (a) is modified to be correct in (b) under constraints on triplet variable.

F/G only(baseline1) : Depth ordering w/o region labeling.
F/B only(baseline2) : Foreground/background region labeling via SVM.
F/G + F/B(proposed) : The multi-class triplet prediction.

Method	Geometric Context		D-order	
	F/G Acc	F/B Acc	F/G Acc	F/B Acc
F/G only	83.2(76.6)	-	95.2(73.2)	-
F/B only	-	86.0(81.8)	-	100(91.9)
F/G+F/B	**88.3(82.0)**	**95.9(84.8)**	**97.0(84.5)**	**100(92.4)**

Table 2. Quantitative results on benchmark datasets. The non-parenthesized/parenthesized figures report the performance using ground-truth segmentation and TSIP segmentation, respectively.

Method	Geometric Context		D-order
	Seg-ho	GT seg	GT seg
Hoiem *et al* [9]	79.9	-	-
Jia *et al* [12]	-	73.3	91.7
Proposed	**81.5**	**80.8**	**93.2**

Table 3. Average figure/ground accuracies(%) on Benchmark dataset in same condition with others. ”Seg-ho” and ”GT seg” denote using segmentation generated from [9] and ground truth segmentation, respectively.

Table 2 presents the quantitative results of figure/ground labeling and foreground region labeling. The non-parenthesized or parenthesized figures report the performance using ground-truth (GT) segmentation and automatically generated segmentation, respectively. Here, the figure/ground accuracy is calculated by fraction of correctly predicted boundary pixels over total boundaries. In case of using automatically generated segmentation, the performance is measured only on the boundaries that intersect with the ones in the GT segmentation. On Geometric Context and D-order datasets, the proposed algorithm outperforms both individual tasks when GT segmentation is given. Although the baseline performances are high since the given GT segmentation is perfect, the proposed algorithm further improves them by inferring labels of the two tasks jointly. In case of using automatically generated segmentation using [13], the overall performances are degraded. However, the proposed algorithm still shows much better performance than the baselines, especially on figure/ground labeling task. Compare to the other algorithms, proposed joint learning method outperforms both of the dataset. Table 3 shows the figure/ground(depth ordering) accuracies on each dataset. We tested on both dataset under same segmentation with same measurement for fair comparison. Moreover, as shown in Fig.3, the proposed joint inference using triplet variable successfully conserves the global consistency between region and edge labeling. Several qualitative results are shown in Fig.2. Some failure cases occur when the automatically generated segmentation contain very irregular and not object-based segments.

5. CONCLUSION

In this paper, we propose an algorithm that simultaneously estimates consistent depth ordering and foreground region labeling. Instead of inferring region and edge labels separately, we infer them jointly as multi-class triplet variable based on the proposed discriminant function. To efficiently solve the optimization problem, LP relaxation is used. In our experiment, the proposed algorithm outperformed two previously proposed state-of-the-art algorithms on two benchmark datasets.

6. REFERENCES

[1] D. Batra, R. Sukthankar, and T. Chen. Learning class-specific affinities for image labelling. In *CVPR*, 2008.

[2] M. J. Black and D. J. Fleet. Probabilistic detection and tracking of motion discontinuities. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 1, pages 551–558. IEEE, 1999.

[3] B.Taskar, C.Guestrin, and D.Koller. Max-margin Markov networks. In *NIPS*, 2003.

[4] D.Hoiem, A.A.Efros, and M.Hebert. Recovering surface layout from an image. *IJCV*, 75(1):151–172, 2007.

[5] C. C. Fowlkes, D. R. Martin, and J. Malik. Local figure–ground cues are valid for natural images. *Journal of Vision*, 7(8), 2007.

[6] A. Gupta, A. A. Efros, and M. Hebert. Blocks world revisited: Image understanding using qualitative geometry and mechanics. In *Computer Vision–ECCV 2010*, pages 482–496. Springer, 2010.

[7] F. Heitger, R. von der Heydt, and O. Kubler. A computational model of neural contour processing: Figure-ground segregation and illusory contours. In *From Perception to Action Conference, 1994., Proceedings*, pages 181–192. IEEE, 1994.

[8] D. Hoiem, A. A. Efros, and M. Hebert. Putting objects in perspective. *International Journal of Computer Vision*, 80(1):3–15, 2008.

[9] D. Hoiem, A. A. Efros, and M. Hebert. Recovering occlusion boundaries from an image. *International Journal of Computer Vision*, 91(3):328–346, 2011.

[10] I.Tsochantaridis, T.Joachims, and T.Hofmann. Large margin methods for structured and interdependent output variables. *JMLR*, 6:1453–1484, 2005.

[11] J.E.Kelley. The cutting-plane method for solving convex programs. *Journal of the Society for Industrial Applied Mathematics*, 8:703–712, 1960.

[12] Z. Jia, A. Gallagher, Y.-J. Chang, and T. Chen. A learning-based framework for depth ordering. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 294–301. IEEE, 2012.

[13] S. Kim, S. Nowozin, P. Kohli, and C. Yoo. Task-specific image partitioning. 2012.

[14] I. Leichter and M. Lindenbaum. Boundary ownership by lifting to 2.1 d. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 9–16. IEEE, 2009.

[15] Y. Lu, W. Zhang, H. Lu, and X. Xue. Salient object detection using concavity context. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 233–240. IEEE, 2011.

[16] H.-K. Pao, D. Geiger, and N. Rubin. Measuring convexity for figure/ground separation. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 948–955. IEEE, 1999.

[17] A. Stein, D. Hoiem, and M. Hebert. Learning to find object boundaries using motion cues. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[18] X. Y. Stella, T. S. Lee, and T. Kanade. A hierarchical markov random field model for figure-ground segregation. In *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 118–133. Springer, 2001.

[19] T.Joachims, T.Finley, and C.N.J.Yu. Cutting-plane training of structural svms. *Machine Learning*, 77(1):27–59, 2009.

[20] I. Tsochantaridis, T. Joachims, T. Hofmann, Y. Altun, and Y. Singer. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6(2):1453, 2006.

[21] R. Vaillant and O. D. Faugeras. Using extremal boundaries for 3-d object modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):157–173, 1992.

[22] V.Vapnik. Statistical learning theory. *Wiley and Sons Inc., New York*, 1998.