

CONTEXT RANKING MACHINE AND ITS APPLICATION TO RIGID LOCALIZATION OF DEFORMABLE OBJECTS

B. Tunç[†], S. K. Zhou^{*}, J. H. Park^{*}, M. Gökmen[†]

[†]Informatics Institute, Istanbul Technical University, Turkey

^{*}Integrated Data Systems Department, Siemens Corporate Research, USA

ABSTRACT

In this paper, we exploit the context information embodied in an image to develop a machine learning method called *context ranking machine (CRM)*. Specifically, we leverage two kinds of context information: *identity context* and *metric context*. The identity context of an image patch refers to its origin (e.g., from which image it is cropped), and the metric context refers to its distance to the exact surrounding box of the target object inside the image. We use these context information in two ways. First, for object localization, instead of learning classifiers to separate the whole negative pool from all positives, we separate each positive from its own negatives sharing the same identity context. Second, we rank image patches according to their resemblance to the ground truth by establishing a connection between appearance based features and metric properties of the image. The CRM learns an image-based ranking algorithm via boosting and achieves an improved localization accuracy. We performed tests on echocardiogram images to localize heart chambers, and face images for eye band localization.

Index Terms— Image segmentation, Object detection, Biomedical image processing

1. INTRODUCTION

We study the problem of object localization, which is related to yet different from object detection. Object detection finds the topological location of the object inside the image but object localization determines the exact metric location. For instance, a face detection algorithm may return any surrounding box containing the face; however, for face localization, the tightest box around the face or the exact face boundary must be considered. The candidate image patches in Fig. 1(a) could be both regarded as successful face detection candidates; while the first candidate can easily be selected over the second one.

The rigidly localized objects in the paper are often of deformable nature. Fig. 1(b) shows one such example: left ventricle (LV) of human heart. The LV is present in an echocardiogram that is an ultrasound image of human heart. It is generally impossible to define a generic rigid template for deformable object due to the complexity of shapes. Therefore, after rigid object localization, a separate deformable shape inference module is needed [1], [2]. As we select closer candidate patches to the ground truth box, the performance of the final shape inference increases critically.

While it is relatively easy to roughly localize the object, it is difficult to obtain an accurate location due to the spatial smoothness of an image. Often, candidate patches with sufficient overlap are considered and merged into one window in an *ad hoc* manner. In [3], the average of the corners of all candidate patches is used as the corners of the final bounding box; in [1], [4], the candidate that maximizes

the object presence probability is used as the final output. In object localization, a rigorous, objective error measurement should be defined. In the paper, we use the distance from the detected result to the ground truth.

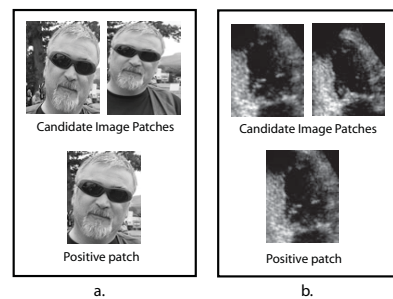


Fig. 1. Different image patches (a) without and (b) with a useful context information

By learning an image-based ranking algorithm via boosting we propose a novel object localization method in Section 3. Section 4 presents the experimental results with improved localization accuracy. Below, we first give a brief overview of related work on rigid object detection/localization and context information.

2. RELATED WORK

The state-of-the-art object detection methods [3, 5, 4] learn a decision boundary to separate all positives from all negatives as shown in Fig. 2(a). However, the identity context lends us an opportunity: we only need to learn the decision boundary that separates the positive from its own negatives as shown in Fig. 2(b).

Using context information for object detection is an emerging topic in computer vision. Torralba and Sinha [6] studied the context information for general object detection: they viewed the context from a holistic perspective and used a mixture of Gaussians to relate the global image content represented by windowed Fourier transform with the object's location and scale. Torralba *et al.* [7] presented a context-based vision system for place and object recognition, aiming to identify familiar locations in new environments and to provide contextual priors for object recognition. Ramstrom and Christensen [8] performed a staged recognition by utilizing context information as a precursor to object detection. Zhou and Comaniciu [2] viewed the context from a local perspective, believing that what is locally observed is a part of a big picture, and develop a regression algorithm to relate the local image content to the target's attributes. The proposed method in the paper is somewhat similar to [2] as both learn the association between appearance and metric distance; but

we learn a ranker instead of regressor, which is arguably easier to learn and hence more accurate.

Image ranking for information retrieval is a well studied subject in the literature [9]. In computer vision, supervised ranking has also been studied. Yan *et al.* [10] proposed to learn a constrained RankBoost algorithm in a shape localization problem in a Bayesian framework. Athitsos *et al.* [11] proposed the so-called BoostMap approach that approximated similarity ranking using AdaBoost to combine simple 1D embeddings, leading to efficient image retrieval with minimal loss in accuracy. Zheng *et al.* [12] used the RankBoost algorithm to perform example-based nonrigid shape segmentation.

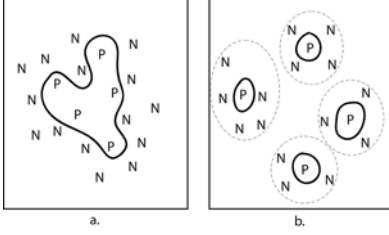


Fig. 2. Decision boundaries (a) in a regular classification and (b) when contextual information is available

3. OBJECT LOCALIZATION USING CONTEXT RANKING MACHINE

The proposed ranking algorithm utilizes a boosting scheme to rank image patches according to their resemblance to ground truth. Based on this ranked list, we keep only the top n candidates. Finally, for complete deformable shape segmentation, those n candidates are used as inputs to a shape inference module.

3.1. Utilizing context information

Any image patch coming from a well-structured image such as LV images presents two kinds of context information: (i) identity context and (ii) metric context. The first one tells us the source image from which this patch was extracted while the latter one is related to the distance of that patch from its ground truth.

To better understand how we can exploit the context information, one can consider a medical image as an example. Since the appearance of the image is totally determined by the anatomical structure, every different image belonging to different patients constitutes its own identity. Inside each identity, there is a hidden metric hierarchy among different patches. For our LV case, each heart image belonging to different people has its own *identity context* since different people's heart images look different due to anatomical features. Inside each image there are lots of patches including ground truth. Each patch has a distance to the ground truth. These distance based hierarchy among patches constitutes *metric context* e.g. any patch being cropped from the upper left part of the image looks like the upper left part of the heart.

We use these context information in two ways. First, for object localization, instead of learning classifiers to separate the whole negative pool from all positives, we separate each positive from its own negatives sharing the same identity context. Second, we rank image patches according to their resemblance to the ground truth by establishing a connection between appearance based features and metric context i.e. using an extra label in training related to the distance from the ground truth.

3.2. The CRM algorithm

Since the CRM algorithm selects top n candidates after ranking, we do not expect a complete ordering of the all image patches. Said differently, it is sufficient for us to push closer image patches into the top n and far ones to the bottom. We call the top n candidates as positives and the rest $N - n$ as negatives.

The CRM algorithm is a modified version of the RankBoost algorithm proposed by Freund *et al.* [13]. The RankBoost algorithm minimizes a cost function of the form

$$\mathcal{R}(H) = \sum_{k=1}^K \sum_{i=1}^I \exp\{-H(x_k) + H(x_i)\} \quad (1)$$

assuming that we have K negatives (e.g. those far away from the ground truth) and I positives (those very close to the ground truth). In CRM algorithm, we utilize the identity context to separate positives and negatives into different groups. Metric context is exploited as an extra weighting function.

Suppose that C is the total number of training identities, x_k^c is the positive and x_i^c the negative data elements for the identity c which have localization errors (distance from ground truth) d_k^c and d_i^c , respectively. Note that the condition $d_k^c < d_i^c$ always holds. We would like to learn a ranking function H that minimizes the following cost function:

$$\mathcal{R}(H) = \sum_{c=1}^C \sum_{k=1}^n \left(\sum_{i=1}^{N-n} (d_i^c - d_k^c)^q \exp\{-H(x_k^c) + H(x_i^c)\} \right)^p \quad (2)$$

where p and q are control parameters.

We build upon the p -norm cost function to further absorb the metric context in a unified treatment [14]. When two data elements are far apart (e.g., $d_i^c - d_k^c$ is big), using the coefficient $(d_i^c - d_k^c)^q$ further pushes them away as it calls a small term $\exp\{-H(x_k^c) + H(x_i^c)\}$ to compensate. If $q = 0$, the cost function in (2) reduces to one used in [14]; further if $p = 1$, it reduces to (1) used in [13].

We invoke the powerful boosting framework to learn the strong ranker $H(x)$, assuming that the strong ranker is a linear combination of weak rankers, i.e., $H(x) = \sum_t \lambda_t h_t(x)$. The weak rankers are selected iteratively. At the t^{th} iteration, one new weak ranker $h_t(x)$ with its blending coefficient λ_t is added to the strong ranker. A simple optimization scheme similar to the one proposed in [13] can be used for derivation of weak ranker $h_t(x)$ and its corresponding coefficient λ_t for each iteration.

Simple mathematical derivation tells that the selection of weak ranker at the t^{th} iteration is done by solving the following task:

$$\max_{h_t} \sum_{c=1}^C \sum_{k=1}^n \left\{ \left(\sum_{i=1}^{N-n} W_{ki}^{ct} \right)^{p-1} \left(\sum_{i=1}^{N-n} W_{ki}^{ct} \Delta_{ki}^{ct} \right) \right\} \quad (3)$$

where,

$$\Delta_{ki}^{ct} := h_t(x_k^c) - h_t(x_i^c)$$

$$W_{ki}^{ct} := \frac{(d_i^c - d_k^c)^q \exp\left(\sum_{j=1}^{t-1} -\lambda_j \Delta_{ki}^{cj}\right)}{\sum_{i,k} (d_i^c - d_k^c)^q \exp\left(\sum_{j=1}^{t-1} -\lambda_j \Delta_{ki}^{cj}\right)} \quad (4)$$

The coefficient λ_t can be found by solving the following equation:

$$0 = \sum_{c=1}^C \sum_{k=1}^n (A_k^c B_k^c) \quad (5)$$

where,

$$A_k^c = \left(\sum_{i=1}^{N-n} W_{ki}^{ct} \exp(-\lambda_t \Delta_{ki}^{ct}) \right)^{p-1}$$

and

$$B_k^c = \left(\sum_{i=1}^{N-n} \Delta_{ki}^{ct} W_{ki}^{ct} \exp(-\lambda_t \Delta_{ki}^{ct}) \right)$$

We associate each weak ranker with an image feature. We use the Haar-like features [3]. Each single feature used in the ranker provides a simple *weak* ranking on the data. By combining these weak rankers, we attain a strong ranking as shown on Figure 3.

Given the identity list $\{1, 2, \dots, C\}$, data for each identity X^c in ascending order according to their d^c , local rectangle feature pool F , the number of rounds T , and the power factors p, q

- $W_{ki}^{c,1} = (d_i^c - d_k^c)^q / Z^1$ where $Z^1 = \sum_{c,k,i} W_{ki}^{c,1}$
 - For $t = 1, \dots, T$
 - For each image feature, train a weak ranker $h(x)$;
 - Select h_t by Eq. (3); */*feature selection*/*
 - Calculate λ_t using Eq. (5);
 - Update the pair weight distribution as in Eq. (4)
 - The final strong ranker is $H(x) = \sum_{t=1}^T \lambda_t h_t(x)$
-

Fig. 3. CRM: Context Ranking Machine Algorithm

4. EXPERIMENTAL EVALUATION

Two kinds of experiments were conducted during testing: (1) left ventricle (LV) localization from echocardiogram, (2) eye band localization inside a detected face. Eye band localization is an already well-studied subject, and we only performed this test to compare our algorithm with the regular AdaBoost algorithm. We used the minimum, average, and maximum error distances of the top n candidates as performance gauges. The resulting boxes are compared to the ground truth box. AdaBoost is used for comparison to show the difference between localization and detection. By this way, we conclude the fact that one can exploit metric context information to transform a detection algorithm to a localization algorithm. To rank the candidate patches coming from AdaBoost, we used detection confidence values.

4.1. Left ventricle localization

For LV localization, the goal is to segment the LV border from an echocardiogram of apical two-chamber view (A2C), in which two heart chambers are visible. Image patches from several identities are labeled with distances to their ground truth boxes for training. For each test with varying training sizes, cross validation is performed and averages are taken. Minimum, mean, and maximum distances of candidate boxes are represented on Figure 4. In figure, the bottom and the top of the vertical lines indicate minimum and maximum errors respectively and the short horizontal lines correspond to average errors. As it is clear from the graphs, CRM is continuously outperforms AdaBoost.

To make the problem harder, we also tried plugging CRM as a final level to a detection cascade. Probabilistic Boosting Network (PBN) [4] cascade is used for this test. The resulting candidate patches are already accumulated around ground truth since it uses a probabilistic network with several levels, and this makes ranking a difficult task. Comparison is made between PBN confidence and

CRM ranking. Table 1 shows final contour errors after a shape inference module [1] is applied. For each identity, we selected the best among the top n candidates using the shape inference score and generated the resulting LV contour. The final segmentation error is calculated as point-to-point maximum distance between the resulting contour and the ground truth. Even only using the top 5 candidates of CRM, we can reach better segmentation accuracy than that achieved by using the top 10 of the detector. This improvement is critical since it brings an important reduction in computational cost during shape inference. Some final inference results can be seen in Figure 5.

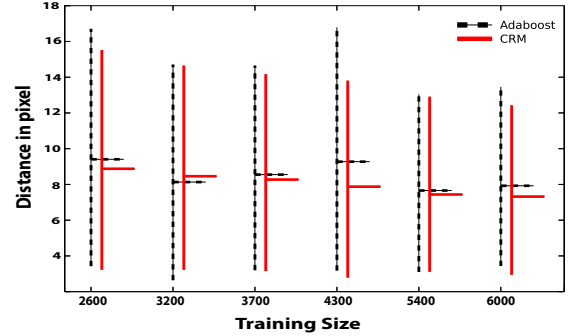


Fig. 4. Minimum, mean, and maximum distances of candidate boxes for Top 3

Table 1. Final deformable segmentation errors after shape inference applied

	Using Top-10	Using Top-5
Detec. + shape inf.[1]	10.19	10.82
CRM + shape inf.	10.12	10.15

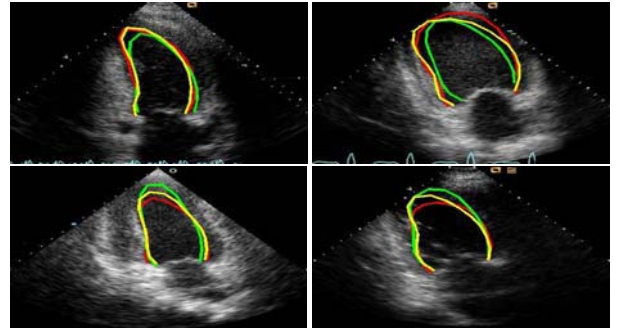


Fig. 5. LV segmentation results. The red line is the ground truth, the yellow is the result by the CRM approach, and the green is the result obtained by the detector approach.

4.2. Eye localization

The goal is to determine the exact location of the eye band inside a face image, assuming that faces are already detected. The whole data set includes 25 different people with pose variations, resulting in 60 identities. Similar to the LV localization, we used cross validation for each training size. The comparison is again performed between

CRM and AdaBoost. We conducted 5 different tests with different training sizes. For each test, 50 identities are used for training and 10 for testing. Figure 6 represents minimum, mean, and maximum distances of candidate boxes detected by CRM and AdaBoost. On Figure 7, several localization results are presented. For both AdaBoost and CRM, the simple mean of the top 3 is taken as the final localization result.

Efficiency of CRM algorithm may be increased by tuning parameters q and p of equation (2). A simple analysis is shown on Figure 8. The limits can be determined by the machine precision since too big q value may result in a zero weight value.

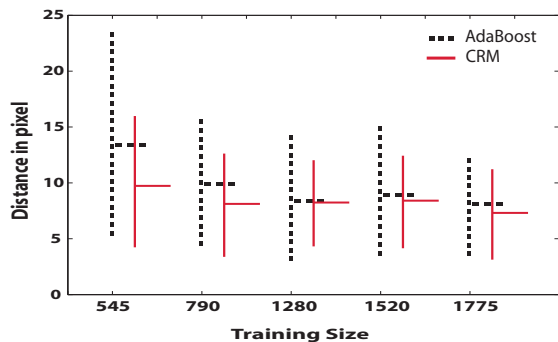


Fig. 6. Minimum, mean, and maximum distances of candidate boxes for Top 3

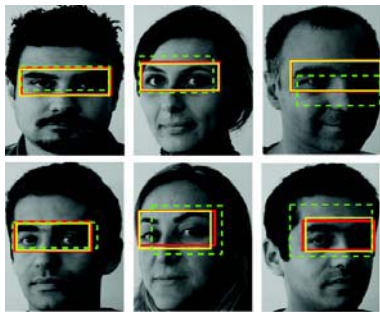


Fig. 7. Eye localization results. The red line is the ground truth, the yellow is the result by the CRM approach, and the green is the result obtained by the detector approach.

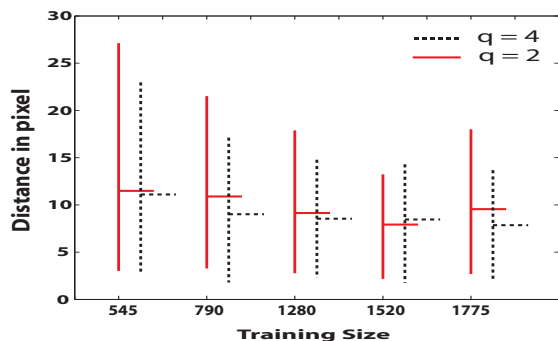


Fig. 8. Sensitivity analysis for parameter q

5. CONCLUSION

We have proposed a novel object localization scheme to achieve improved object localization. The main idea is to incorporate the context information embodied in an image. When there is a well defined metric context inside an image or an aperture around the object, we can order the image patches according to their closeness to the ground truth. The second advancement is the usability of the context information to facilitate the training. Instead of learning a single global hyperplane, we learn lots of individual ones without causing any overfitting. This approach decreases the number of pairs to be ordered in a ranking study and therefore reduces computational complexity in training. Our experimental results on LV segmentation and eye band localization have demonstrated the effectiveness of CRM over regular detector.

6. REFERENCES

- [1] B. Georgescu, X. S. Zhou, D. Comaniciu, and A. Gupta, "Database-guided segmentation of anatomical structures with complex appearance," in *Proc. CVPR*, 2005, vol. 2, pp. 429–436.
- [2] S. Zhou and D. Comaniciu, "Shape regression machine," in *Proc. IPMI*, 2007.
- [3] P. Viola and M. Jones, "Robust real-time object detection," *Technical Report CRL 20001/01, Cambridge Research Laboratory*, 2001.
- [4] J. Zhang, S. K. Zhou, L. McMillan, and D. Comaniciu, "Joint real-time object detection and pose estimation using probabilistic boosting network," *Comp. Vis. and Patt. Recog.*, pp. 1–8, 2007.
- [5] S. Li and Z. Zhang, "FloatBoost learning and statistical face detection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 26, no. 9, pp. 1112 – 1123, 2004.
- [6] A. Torralba and P. Sinha, "Statistical context priming for object detection," in *Proc. ICCV*, 2001.
- [7] A. Torralba, K.P. Murphy, W.T. Freeman, and M.A. Rubin, "Context-based vision system for place and object recognition," in *Proc. ICCV*, 2003.
- [8] O. Ramstrom and H. I. Christensen, "Object detection using background context," in *Proc. ICPR*, 2004, pp. 45–48.
- [9] R. C. Veltkamp and M. Tanase, "Content-based image retrieval systems: a survey," *Technical Report UU-CS-2000-34, Universiteit Utrecht*, 2000.
- [10] S. Yan, M. Li, H. Zhang, and Q. Cheng, "Ranking prior likelihood distributions for bayesian shape localization framework," in *Proc. ICCV*, 2003, vol. 1.
- [11] V. Athitsos, J. Alon, S. Sclaroff, and G. Kollias, "Boostmap: A method for efficient approximate similarity rankings," in *Proc. CVPR*, 2004, vol. 2, pp. 268–275.
- [12] Y. Zheng, X.S. Zhou, B. Georgescu, S. Zhou, and D. Comaniciu, "Example based non-rigid shape detection," in *ECCV*, 2006.
- [13] Y. Freund and et al, "An efficient boosting algorithm for combining preferences," *Procs. of Int. Conf. Machine Learning*, pp. 170–178, 1998.
- [14] C. Rudin, "Ranking with a p-norm push," *Conf. Learning Theory*, pp. 589–604, 2006.