MULTI-VIEW FACE DETECTION IN VIDEOS WITH ONLINE ADAPTATION

Yao-Chuan Chang[1], *Yen-Yu Lin*[1,2], *Hong-Yuan Mark Liao*[1,2]

[1]Research Center for Information Technology Innovation, Academia Sinica, Taiwan
[2]Institute of Information Science, Academia Sinica, Taiwan

ABSTRACT

Most learning-based approaches to face detection suffer from the problem of performance degradation on faces that are not covered by training data. However, including all variations of faces in training is practically infeasible due to the scalability restriction of machine learning algorithms and expensive manual labeling. In this work, we focus on face detection in videos, and alleviate this problem by exploiting strong correlation among video frames. We augment a pre-trained multi-view face detection with an incrementally derived Gaussian process regressor. The regressor can extract and propagate visual knowledge across frames, and adapts the detector to handle unseen faces. Testing on two datasets, the promising results manifest the effectiveness of the proposed approach.

Index Terms— Face detection, video analysis, transfer learning, boosting, Gaussian process regression

1. INTRODUCTION

Face detection is an important and challenging problem in the fields of computer vision and pattern recognition. As a key component for image and video analysis, face detection is an inherent part in plenty applications, such as surveillance, human computer interface, image and video retrieval [1, 2, 3]. In literature, e.g., [4, 5, 6, 7, 8], face detection is often cast as a classification task through supervised learning. Although the design of face detectors has gained significant progress, owing to the nature of supervised learning, most face detection algorithms suffer from a common problem that the performance degrades when testing faces are not covered by training data. The appearances of faces are determined by miscellaneous factors, such as illuminations, poses, and their combinations. It is practically infeasible to have all possible variations of faces in training due to expensive labeling cost and scalability restrictions in machine learning.

In this work, we aim to alleviate this problem in detecting faces in videos. Motivated by the strong correlations among video frames, we propagate visual knowledge regarding detection across frames, and leverage the additional information to dynamically adapt the pre-trained detector. Besides, we propose an efficient way to carry out online detector adaptation along streamed video frames.

1.1. Related work

The literature of face detection is quite comprehensive [9, 10]. A number of techniques, e.g., [4, 5, 6, 11], have been demonstrated to be effective for detecting frontal faces in a wide range of image databases. Since faces in images often appear with different illuminations, poses, and occlusions, numerous approaches, such as [7, 12, 13, 14, 15], have been proposed for tackling multi-view (class) face detection. However, those approaches rely on a large amount of training samples for each class of faces, say 10^4, to compile a stable detector. Furthermore, although most previous approaches can deal with one kind of face variation effectively, such as rotation, occlusion, and lighting condition, they generally fail to find faces with combination of multiple variation soruces. In realistic applications, both issues induce difficulties for representative and sufficient data collection. It results in differences between the distributions of training and testing data.

Object tracking [16, 17, 18] gives another way for face detection in videos. Despite the fact that several tracking models have met the challenge caused by alternation of face appearances, if target faces are occluded for a period of time and if the faces enter or leave the scene, most trackers fail. Moreover, face trackers require prior knowledge about target faces and the number and type of parameters being tracked, all of which are needed to be specified by users in advance.

Transfer learning [19] aims at delivering abundant knowledge available in the source tasks to facilitate the target task. For face detection, Jain and Learned-Miller [20] have proposed an adaptive face detector via transferring information in an image. They collect the patterns that are predicted by the detector with high confidence, and use these patterns to train a model of *Gaussian process regression* (GPR). The resulting GPR can adapt the detector and improve performance. However, information propagated within an image is typically restricted, since only one or few faces appear in a single image. Besides, learning GPR leads to a computational burden [21, 22], and may hinder face detection from real-time processing. In this work, we leverage correlation among video frames, and carry out spatial-temporal adaptation. Specifically, we treat sub-windows confidently predicted by the detector as a knowledge source, and design an incremental mechanism to efficiently adapt the pre-trained face detector.

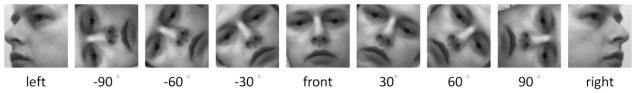


Fig. 1. Faces of nine views used in learning the detector.

1.2. Our Approach

In this work, we address the task of face detection in videos, and aim to resolve the difficulty arising when testing faces are not completely covered by the training data. To this end, we leverage the spatial and temporal correlation among video frames, and propose to augment the pre-trained detector [15] with an incrementally learned GPR models. The GPR is designed to capture the evidences regarding the face and non-face distributions in the video, and transfer the additional visual knowledge across frames. It follows that the resulting detector can be adapted to deal with faces unseen in training. Our approach gives the advantage that once the face detector is trained, the whole procedure is run in an unsupervised manner. Besides, we make use of the streamed frames, and design an block-wise update structure to efficiently derive GPR models. Our approach is evaluated on two datasets. One is the *NRC-IIT* facial video dataset [23]. The other is a set of *variety & talk shows* collected by ourselves.

2. PRELIMINARY

This section describes two key components used in the establishment of the proposed framework, including multi-view face detection and Gaussian process regression.

2.1. Multi-View Face Detection

There exist several off-the-shelf face detectors, upon which our approach can be developed. Among these detectors, we illustrate our approach with the detector derived by *MBH-Boost* [15], and improve its performance of finding faces in videos by online detector adaptation. The reasons why MBHBoost detector is adopted are three-fold. First, it supports real-time, multi-view face detection. These properties increase the generalization of our approach to work on unconstrained video pools, e.g., YouTube, since those videos may be composed of numerous frames and faces of various poses may appear. Second, it implements the principle of *classifier sharing* [24], and iteratively derives weak learners with theoretic merit to ensure the accuracy. Third, it carries out multi-view/class detection, and treats each face class in a unified manner. Namely, we could develop our approach without worrying about the relationships among face classes.

Specifically, we learn the face detector with training data of nine classes, shown in Fig. 1. Then we investigate the correlation among frames in a video, and adapt the detector to handle unseen faces, such as faces with mix of in-plane and out-of-plane rotations. An example is given in Fig. 5.

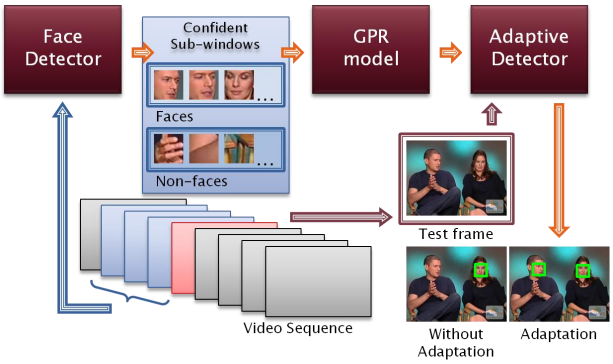


Fig. 2. Online detector adaptation in a video.

2.2. Gaussian Process Regression

Gaussian process regression (GPR) [25] will be used to adapt the detector in this work. It is a stochastic process that adopts Gaussian prior distribution with Bayesian treatment for linearly predicting the target value $g(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} \in \mathbb{R}$ of an instance $\mathbf{x}$. With a set of training data, $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, let $X = [\mathbf{x}_1\ \mathbf{x}_2\ \cdots\ \mathbf{x}_N]$ and $\mathbf{y} = [y_1\ y_2\ \cdots\ y_N]^\top$ denote the data matrix and target vector, respectively. The posterior distribution of weight $\mathbf{w}$ in GPR is modeled as

$$p(\mathbf{w}|X, \mathbf{y}) = \mathcal{N}(\frac{1}{\sigma^2} A^{-1} X^\top \mathbf{y}, A^{-1}), \tag{1}$$

where $A = \frac{1}{\sigma^2} X X^\top + I$, and σ is a positive constant.

Gaussian process, like SVM [26], is a kernel machine. It can be derived that all operations regarding data samples in GPR can be completed in form of *inner product*. It follows that the *kernel trick/fuctnion* [27] can be adopted to efficiently compute the inner product of data that are nonlinearly projected into some predefined high dimensional feature space. After learning a GPR model with kernel, the regression value y^* of a testing sample $\mathbf{x}^*$ is predicted by

$$y^* \equiv g(\mathbf{x}^*) = \mathbf{k}(\mathbf{x}^*)(K + \sigma^2 I)^{-1} \mathbf{y}, \tag{2}$$

$$\text{where } \mathbf{k}(\mathbf{x}^*) = [k(\mathbf{x}^*, \mathbf{x}_1)\ \cdots\ k(\mathbf{x}^*, \mathbf{x}_N)] \in \mathbb{R}^N, \tag{3}$$

$$K = [k(\mathbf{x}_i, \mathbf{x}_j)] \in \mathbb{R}^{N \times N}, \tag{4}$$

and k is the kernel function. In this work, we use RBF kernel, i.e., $k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(\frac{-||\mathbf{x}_i - \mathbf{x}_j||^2}{\gamma}\right)$, for its stability and good performance. In all the experiments, we set the value of parameter γ as the average squared distance among training data, which is suggested in [27], and determine σ by method used in [28].

3. THE PROPOSED FRAMEWORK

Our approach, augmenting the pre-trained detector with an incrementally learned GPR, is described in the section. For a quick start, we outline the proposed approach in Fig. 2.

Algorithm 1: Adapted Detector of A Boosted Cascade

Input: Sub-windows to be classified: X;
The face detector: $\{f_{n,k}\}_{n=1,k=1}^{N,K}$;
The GPR models at time t: $\{g_{n,k}^t\}_{n=1,k=1}^{N,K}$;
Output: Detected faces of class k: X_k, $k = 1 \sim K$;
Initialize: $X_k \leftarrow X$, for $k = 1, 2, ..., K$;
for $n = 1; n \leq N;$ **do**
  for $k = 1; k \leq K;$ **do**
    $X_m \leftarrow \{\mathbf{x} \in X_k | \, |f_{n,k}(\mathbf{x})| < \epsilon\}$;
    $X_o \leftarrow X_k \setminus X_m$;
    $\forall \mathbf{x} \in X_o, \; f'_{n,k}(\mathbf{x}) = f_{n,k}(\mathbf{x})$;
    $\forall \mathbf{x} \in X_m, \; f'_{n,k}(\mathbf{x}) = g_{n,k}^t(\mathbf{x})$;
    $X_k \leftarrow \{\mathbf{x} \in X_k | \, f'_{n,k}(\mathbf{x}) \geq 0\}$;

3.1. Online Face Detector Adaptation

The multi-class face detector derived by MBHBoost [15] is composed of a boosted cascade, i.e., $\{\text{sign}(f_{n,k}(\cdot))\}_{n=1,k=1}^{N,K}$, where N and K are the numbers of stages and classes respectively. In detection, a pattern $\mathbf{x}$, a sub-window of 24×24 pixels here, is predicted stage-by-stage, and $\mathbf{x}$ is considered a face of the kth class if and only if it passes all the classifiers of the kth class, i.e., $f_{n,k}(\mathbf{x}) \geq 0$ for $n = 1, 2, ..., N$.

The adaptation of the MBHBoost detector in this work is developed based on two observations. First, video content is spatially and temporally continuous. The appearances of faces and non-faces vary gradually in a video. Hence, the patterns *reliably predicted* in previous frames provide rich information to facilitate detection in successive frames. Second, the larger the prediction magnitude $|f_{n,k}(\mathbf{x})|$ in a boosted classifier is, the more likely the prediction is correct. This phenomenon has been pointed out in [20]. By integrating the two observations, we aim to borrow the confidently predicted patterns in the past frames to adjust the detector. While the procedure is done frame-by-frame, the adapted detector will successfully locate a face in the video if the training data cover the appearance of the face in some frame, rather than the whole video.

Specifically at time stamp t, we derive the adapted detector $\{f_{n,k}^t\}_{n=1,k=1}^{N,K}$ by augmenting it with GPR models $\{g_{n,k}^t\}_{n=1,k=1}^{N,K}$. Suppose that the confidently predicted sub-windows as well as the associated predictions in the previous frames are collected. The GPR $g_{n,k}^t$ is learned with training data $S_{n,k}^t = \{(\mathbf{x}, f_{n,k}(\mathbf{x})) \mid |f_{n,k}(\mathbf{x})| \geq \epsilon\}$. Here each sub-window $\mathbf{x}$ resides in one of the previous frames, and is represented by the pixel intensities.

For a testing image, the adapted detector will scan the whole sub-windows generated with different scales and locations. However, after predicting these sub-windows by the original detector, sub-windows that can't be confidently predicted, i.e., $|f_{n,k}(\mathbf{x})| < \epsilon$, will be *re-classified* by GPR $g_{n,k}^t$.

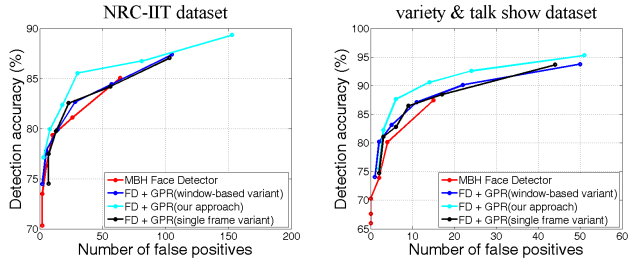


Fig. 3. The ROC curves of different face detection methods.

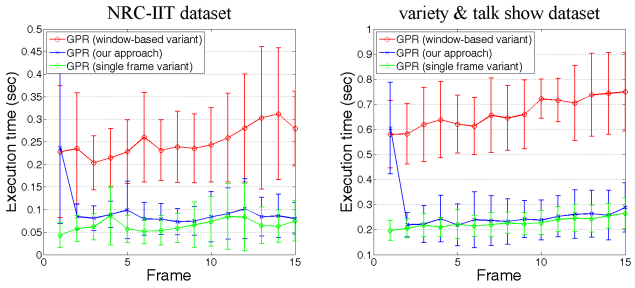


Fig. 4. The average execution time of various adaptation methods at each video frame.

Namely, the adapted detector $\{f_{n,k}^t\}_{n=1,k=1}^{N,K}$ is given by

$$f_{n,k}^t = \begin{cases} f_{n,k}(\mathbf{x}), & \text{if } |f_{n,k}(\mathbf{x})| \geq \epsilon, \\ g_{n,k}^t(\mathbf{x}), & \text{otherwise.} \end{cases} \tag{5}$$

The testing procedure of the adapted detector is summarized in Algorithm 1. Note that the GPR models are learned with training sub-windows collected in all the previous frames. For comparison, we implement two variants of the adapted detector, including 1) *single frame variant*: training data for GPR models contain sub-windows selected in frame t. It can be considered to be the same as the work by Jain and Leaned-Miller [20], except the MBHBoost detector is used; 2) *window-based variant*: training data are collected from the previous w frames, i.e., frames $t - w \sim t - 1$.

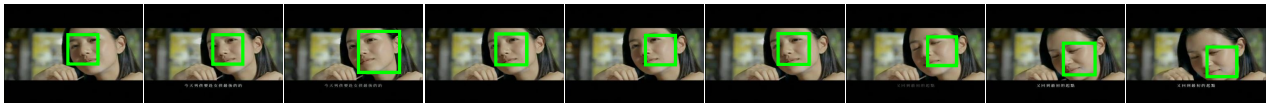
3.2. Incremental Gaussian Process Regression

As shown in (2), the computation bottleneck of learning GPR is to calculate the inverse of the kernel matrix. *Block-wise inversion* can be applied to GPR for incremental update since the new information comes into our system sequentially. Suppose the matrix X_{new} denotes the collection of P incoming training data at frame t, and X_{t-1} represents the collection of all M training data from previous frames. Then X_t can be defined as

$$X_t = [X_{t-1} \; X_{new}] \, . \tag{6}$$

It denotes the collection of the $M + P$ training input vectors since beginning. Then the kernel matrix regarding data X_t

(a) Detection results by the original face detector



(b) Detection results by the adapted face detector

Fig. 5. Detection results by applying the face detectors to one of our *variety & talk show* video sequences.

can be represented by

$$K(X_t, X_t) = \begin{bmatrix} K(X_{t-1}, X_{t-1}) & K(X_{t-1}, X_{new}) \\ K(X_{new}, X_{t-1}) & K(X_{new}, X_{new}) \end{bmatrix} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{D} & \mathbf{E} \end{bmatrix}.$$

As both $K(X_{t-1}, X_{t-1})$ and $K(X_{new}, X_{new})$ are covariance matrices, they share positive semi-definite property, and their relevant inverses exist. Therefore, using the small rank adjustment formula, the inverse of $K(X_t, X_t)$ can be alternatively written as

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{D} & \mathbf{E} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}\Delta^{-1}\mathbf{D}\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}\Delta^{-1} \\ -\Delta^{-1}\mathbf{D}\mathbf{A}^{-1} & \Delta^{-1} \end{bmatrix},$$

where $\Delta = \mathbf{E} - \mathbf{D}\mathbf{A}^{-1}\mathbf{B}$. In our cases where $M >> P$, it can be verified that the computational complexity of calculating matrix inversion can be reduced from $O(M^3)$ to $O(M^2)$.

4. EXPERIMENTAL RESULTS

4.1. Datasets

Two datasets are used to measure the performance of the proposed approach to online face detector adaptation, including *NRC-IIT* facial video database [23] and *variety & talk show* dataset collected by us. In NRC-IIT database, 18 single-face videos with identical resolution (160×120) are used to approximately estimate the execution time. Those videos are composed of faces belong to 10 persons in different poses, scales, or orientations. In addition, 10 video sequences with 20 different faces taken from several variety or talk shows are used to test whether our system can work robustly with multi-face appearances, complex human actions, and miscellaneous backgrounds. These videos were selected from the Internet. These videos sequences are of resolution from 480×360 to 720×480. The appearances of faces vary dramatically along video frames, and are not fully covered in training data, i.e., those shown in Fig. 1. Although this makes the dataset very challenging, the dataset hence serves as a good test bed to evaluate our approach to detector adaptation.

4.2. Experimental Setting

We have the multi-view face detector by MBHBoost [15] as the base detection algorithm. For comparison in both the aspects of accuracy and efficiency, we also implement three types of adaptation methods, including the proposed approach and its single frame and window-based variants, all of which have been mentioned in section 3.1. Performance is measured in form of the *ROC curves* which are generated with different numbers of stages in the boosted cascades (with $17 \sim 22$ stages). The less the stages are, the higher the detection rates and false positive rates are. For evaluation of efficiency, the average computational time of learning GPR and related training data collection of different adaptation methods are illustrated in Fig. 4.

4.3. Results

It can be observed in the ROC curves of Fig. 3 that the proposed approach outperforms all of the other methods in almost the entire range of different numbers of stage usages in the two datasets. Additionally, according to Fig. 4, the proposed approach can effectively reduce the computational cost. The good performances in both accuracy and efficiency validate the effectiveness of our approach. For visualization, Fig. 5 shows the improvement of detection performance by applying the proposed adapted face detector to one of the video sequences in the variety & talk show dataset.

5. CONCLUSIONS

One of the underlying problems in face detection is to collect sufficient and representative training data. Leveraging the strong correlation in a video sequence, we alleviate this problem by proposing a novel approach to face detector adaptation. It uses the GPR to transfer visual knowledge across frames without human labeling. To reduce the computational cost, the block-wise incremental update framework is adopted for fast GPR training. Besides, our system is comprehensively evaluated under different settings with a public video dataset and a collection of variety and talk shows. Promising performances in both accuracy and efficiency are obtained.

Acknowledgments. The work is supported in part by grant NSC 101-2221-E-001-018.

6. REFERENCES

[1] G.L. Foresti, C. Micheloni, L. Snidaro, and C. Marchiol, "Face detection for visual surveillance," in *Proc. Int'l Conf. Pattern Recognition*, 2003, pp. 115–120.

[2] E. Corvee and F. Bremond, "Combining face detection and people tracking in video sequences," in *Proc. Int'l Conf. Crime Detection and Prevention*, 2006, pp. 1–6.

[3] T.-K. Kim, S.-U. Lee, J.-H. Lee, S.-C. Kee, and S.-R. Kim, "Integrated approach of multiple face detection for video surveillance," in *Proc. Int'l Conf. Pattern Recognition*, 2002, pp. 394–397.

[4] E. Osuna, R. Freund, and F. Girosit, "Training support vector machines: an application to face detection," in *Proc. Conf. Computer Vision and Pattern Recognition*, 1997, pp. 130–136.

[5] H.A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 1, pp. 23–28, 1998.

[6] P. Viola and M. Jones, "Rapid object detectionusing a boosted cascade of simple feature," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2001, pp. 511–518.

[7] P. Viola and M. Jones, "Fast multi-view face detection," Tech. Rep., Mitsubishi Electric Research Laboratories, 2003.

[8] Y.-Y. Lin, T.-L. Liu, and C.-S. Fuh, "Fast object detection with occlusions," in *Proc. European Conf. Computer Vision*, 2004, pp. 402–413.

[9] M.-H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 1, pp. 34–58, 2002.

[10] C. Zhang and Z. Zhang, "A survey of recent advances in face detection," Tech. Rep., Microsoft Research, 2010.

[11] R. Lienhart and J. Maydt, "An extended set of haar-like features for rapid object detection," in *Proc. Int'l Conf. Image Processing*, 2002, pp. 900–903.

[12] S. Li, L. Zhu, Z. Zhang, A. Blake, H. Zhang, and H. Shum, "Statistical learning of multi-view face detection," in *Proc. European Conf. Computer Vision*, 2002, pp. 67–81.

[13] K.K. Sung and T. Poggio, "Example based learning for view-based human face detection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 1, pp. 39–51, 1998.

[14] C. Huang, H. Ai, Y. Li, and S. Lao, "Vector boosting for rotation invariant multi-view face detection," in *Proc. Int'l Conf. Computer Vision*, 2005, pp. 446–453.

[15] Y.-Y. Lin and T.-L. Liu, "Robust face detection with multi-class boosting," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2005, pp. 680–687.

[16] S. Birchfield, "Elliptical head tracking using intensity gradients and color histograms," in *Proc. Conf. Computer Vision and Pattern Recognition*, 1998.

[17] T.-L. Liu and H.-T. Chen, "Real-time tracking using trust-region methods," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 26, no. 3, pp. 397–402, 2004.

[18] B. Babenko and M.-H. Yang, "Robust object tracking with online multiple instance learning," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1619–1632, 2011.

[19] S. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[20] Vidit Jain and Erik Learned-Miller, "Online domain adaptation of a pre-trained cascade of classifiers," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2011, pp. 577–584.

[21] D. Nguyen-Tuong, M. Seeger, and J. Peters, "Local gaussian process regression for real time online model learning and control," in *Advances in Neural Information Processing Systems*, 2008, pp. 1193–1200.

[22] D. Nguyen-Tuong, M. Seeger, and J. Peters, *Real-Time Local GP Model Learning. In From Motor Learning to Interaction Learning in Robots*, vol. 264, Springer, 2010.

[23] D. Gorodnichy, "Associative neural networks as means for low-resolution video-based recognition," in *Proc. Int'l Joint Conf. Neural Networks*, 2005.

[24] A. Torralba, K. Murphy, and W. Freeman, "Sharing visual features for multiclass and multiview object detection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 5, pp. 854–869, 2007.

[25] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*, MIT Press, 2006.

[26] V. Vapnik, *Statistical Learning Theory*, Wiley, 1998.

[27] B. Schölkopf and A. Smola, *Learning with Kernels*, MIT Press, 2002.

[28] A. Kapoor, K. Grauman, R. Urtasun, and T. Darrell, "Gaussian processes for object categorization," *Int'l J. Computer Vision*, vol. 88, no. 2, pp. 169–188, 2010.