PORNOGRAPHY DETECTION USING BOSSANOVA VIDEO DESCRIPTOR

Carlos Caetano[*], *Sandra Avila*[*†], *Silvio Guimarães*[‡], *Arnaldo de A. Araújo*[*]

[*] Federal University of Minas Gerais, NPDI Lab — DCC/UFMG, Minas Gerais, Brazil
[†] University of Campinas, RECOD Lab — DCA/UNICAMP, Campinas, Brazil
[‡] Pontifical Catholic University of Minas Gerais, VIPLAB — ICEI/PUC Minas, Minas Gerais, Brazil
carlos.caetano@dcc.ufmg.br, sandra@dca.fee.unicamp.br, sjamil@pucminas.br, arnaldo@dcc.ufmg.br

ABSTRACT

In certain environments or for certain publics, pornographic content may be considered inappropriate, generating the need to be detected and filtered. Most works regarding pornography detection are based on the detection of human skin. However, a shortcoming of these kind of approaches is related to the high false positive rate in contexts like beach shots or sports. Considering the development of low-level local features and the emergence of mid-level representations, we introduce a new video descriptor, which employs local binary descriptors in conjunction with BossaNova, a recent mid-level representation. Our proposed method outperforms the state-of-the-art on the Pornography dataset.

***Index Terms*—** Pornography detection, binary descriptors, BossaNova representation, visual recognition.

1. INTRODUCTION

The immediate challenge that comes with the exponential growth of multimedia content is the development of new tools for content deduplication, multimedia retrieval and understanding [1]. Regarding content deduplication, we are interested in removing all near-duplicated content in order to facilitate the multimedia retrieval. Multimedia understanding is related to visual recognition in which we are interested in identifying, e.g. pornographic content and action recognition.

Usually, the approaches to cope with those problems consider (i) extraction of local image descriptors; (ii) encoding of the local features in a mid-level representation; and (iii) classification and/or search of the image descriptor. Here, instead of using the local image descriptors for multimedia classification, we propose studying the behavior of a video descriptor computed from local image descriptors for pornography detection in video content. Considering that the pornography content is forbidden in several environments (e.g., schools, workplaces), this work can be useful for detecting and filtering out this content to avoid undesirable visualizations.

As can be seen in [2], most of the works regarding the detection of pornographic material has been done for the image domain. Moreover, the majority of those works is based on the detection of human skin [3–6]. However, the main drawback of those kind of approaches is related to the high false positive rate in contexts like beach or sports. Other methods have been explored, like Bag-of-Words (BoW) models [7–9] and BossaNova representation [10, 11]. In fact, Deselaers et al. [7] were the first to propose a BoW model to filter pornographic images. On a similar way, Lopes et al. [8, 9] developed a BoW-based approach, which used the HueSIFT, a SIFT variant including color information, to classify images and videos. In [10], the authors applied the BossaNova representation and HueSIFT descriptors to classify pornographic videos using majority voting. Recently, in our previous work [11], we achieved similar results on [10]'s dataset by using BossaNova and binary descriptors.

Those previous works have explored bags of static features. Very few works have applied spatiotemporal features [12] or other motion information for detection of pornography [13, 14]. Moreover, other approaches have been employed audio analysis as an additional feature [15].

In the face of our good previous results [11], obtained by combining binary descriptors and BossaNova mid-level image representation (a recent extension of the BoW approach), we propose a new video descriptor by applying the median function on BossaNova vectors. Our proposal has as its advantage the fact that it does not depend on any skin detector or shape models to classify pornography; besides, it outperforms the state-of-the-art results on the Pornography dataset [10].

The remainder of this paper is organized as follows. In Section 2, we give a brief explanation of the most recent binary descriptors and the BossaNova mid-level representation. In Section 3, we introduce our video descriptor. In Section 4, we analyze our empirical results, validating the good performance of the BinBoost binary descriptor and our proposed video descriptor. Finally, in Section 5, we present some conclusions and future works.

2. THEORETICAL BACKGROUND

Recognizing categories of objects and scenes is a fundamental human ability and an important, yet elusive, goal for com-

puter vision research. Images consist of pixels that have no semantic information by themselves, making the task very challenging. Thanks to the development of very discriminant low-level local features (such as SIFT descriptors [16]), and the emergence of mid-level aggregate representations, based on the quantization of those features (such as Bag-of-Words[1] model [17]), progress in visual recognition tasks has been made successfully. This section reviews those key concepts that lay the foundation of this work, in particular, the binary descriptors and the BossaNova representation.

2.1. Low-level Features: Binary Descriptors

Binary descriptors have recently emerged as low-complexity alternatives to state-of-the-art descriptors. They have received considerable attention having a similar or better recognition performance when compared to state-of-the-art non-binary descriptors. Also, it benefits from using the Hamming distance instead of the usual Euclidean distance.

The basic idea is that one can encode most of the information of a patch as a binary string using only simple binary tests comparing pixel intensities. BRIEF descriptor [18] does this binary tests comparison on a randomly way according to a Gaussian distribution with respect to the keypoint of the patch. As an extension of BRIEF, ORB [19] makes the selection of points to be compared by a k-nearest neighborhood strategy based on error-prone. The random sampling has been replaced to a sampling scheme that uses machine learning for decorrelating BRIEF features under rotational invariance. BRISK descriptor [20] uses a limited number of points in a concentric pattern. Each point contributes to various pairs used to build the binary descriptor. FREAK descriptor [21] uses a cascade of binary strings by comparing image intensities over a retinal sampling pattern.

Recently, Trzcinski et al. [22] proposed the BinBoost descriptor, which is a extremely compact binary descriptor robust to illumination and viewpoint changes. Each bit of the descriptor is computed with a boosted binary hash function. That function relies on weak learners that consider the orientations of intensity gradients over image patches.

2.2. Mid-level Representation: BossaNova

This section provides a brief introduction to the BossaNova mid-level image representation, which offers more information-preserving pooling operation based on a distance-to-codeword distribution. More details can be found in [10, 23].

Let $\mathcal{X}$ be an unordered set of binary descriptors extracted from an image. $\mathcal{X} = \{\mathbf{x}_j\}$, $j \in [1, N]$, where $\mathbf{x}_j \in \mathbb{R}^D$ is a binary descriptor vector and N is the number of binary descriptors in the image. Let $\mathcal{C}$ be a visual codebook[2] obtained by the k-medians algorithm. $\mathcal{C} = \{\mathbf{c}_m\}$, $m \in [1, M]$, where $\mathbf{c}_m \in \mathbb{R}^D$ is a codeword and M is the number of visual codewords. $\mathbf{z}$ is the final vectorial BossaNova representation of the image used for classification.

The BossaNova approach follows the BoW formalism, but proposes an image representation which keeps more information than BoW during the pooling step. Thus, the BossaNova pooling function g estimates the probability density function of α_m: $g(\alpha_m) = \mathrm{pdf}(\alpha_\mathbf{m})$, by computing the following histogram of distances $z_{m,b}$:

$$\begin{aligned} g : \mathbb{R}^N &\longrightarrow \mathbb{R}^B, \\ \alpha_\mathbf{m} &\longrightarrow g(\alpha_m) = z_m, \\ z_{m,b} &= \mathrm{card}\left(\mathbf{x}_j \mid \alpha_{m,j} \in \left[\frac{b}{B}; \frac{b+1}{B}\right]\right), \\ &\quad \frac{b}{B} \geq \alpha_m^{min} \text{ and } \frac{b+1}{B} \leq \alpha_m^{max}, \end{aligned} \quad (1)$$

where B denotes the number of bins of each histogram z_m, and $[\alpha_m^{min}; \alpha_m^{max}]$ limits the range of distances for the descriptors considered in the histogram computation.

After computing the local histograms z_m for all the $\mathbf{c}_m$ centers, the BossaNova vector $\mathbf{z}$ [10] can be written as:

$$\mathbf{z} = \left[\left[z_{m,b}\right], st_m\right]^{\mathrm{T}}, \quad (m, b) \in \{1, ..., M\} \times \{1, ..., B\}, \quad (2)$$

where $\mathbf{z}$ is a vector of size $M \times (B + 1)$, s is a nonnegative constant and t_m is a scalar value for each codeword, counting the number of binary descriptors $\mathbf{x}_j$ close to that codeword.

In brief, the BossaNova vector is defined by three parameters: the number of codewords M, the number of bins B in each histogram, and the range of distances $[\alpha_m^{min}, \alpha_m^{max}]$. As in [10], we set up the bounds as $\alpha_m^{min} = \lambda_{min} \cdot \sigma_m$ and $\alpha_m^{max} = \lambda_{max} \cdot \sigma_m$, where σ_m is the standard deviation of each cluster $\mathbf{c}_m$ obtained by k-medians clustering algorithm.

Avila et al. [10] applied their representation to the context of visual recognition. In comparison to the BoW, BossaNova significantly outperforms it. Besides, the BossaNova approach was ranked at the second place, considering only visual-based approaches, in the ImageCLEF 2012 challenge [24]. Furthermore, by using a simple histogram of distances to capture the relevant information, the method remains very flexible and keeps the representation compact. For those reasons, we choose the BossaNova approach for mid-level features.

3. A NEW VIDEO DESCRIPTOR BASED ON BOSSANOVA REPRESENTATION

As can be seen in [11], it is possible to represent local binary descriptors with BossaNova for visual recognition. Despite its promising results for pornography detection, the final

[1] Bag-of-Words models have blurred somewhat the distinction between local and global descriptors, because they propose a single (global) feature vector based upon several (local) features.

[2] The codebook is usually built by clustering a set of local descriptors. It can be defined by the set of codewords (or visual words) corresponding to the centroids of clusters.

video classification is obtained by a majority voting over the images. A shortcoming is that the number of pornographic images must be greater than the number of non-pornographic ones, which cannot be always true. Furthermore, this approach consider only static images ignoring the probability of correct and wrong classifications.

In order to address those problems, instead of computing the majority voting, we propose a new video descriptor which improves the BossaNova video classification performance by combining the image signatures as follows.

Let $\mathcal{V}$ be a video sequence. $\mathcal{V} = \{f^i\}$, $i \in [1, N]$, where f^i is the keyframe[3] of the shot i and N is the number of keyframes. Let $\mathcal{Z}$ be a set of BossaNova vectors computed for the video $\mathcal{V}$. $\mathcal{Z} = \{\mathbf{z}^i\}$, $i \in [1, N]$, where $\mathbf{z}^i$ is a BossaNova vector extracted for the keyframe f^i. The BossaNova video descriptor $\mathcal{Z}$ can be modeled by a function h as follows:

$$\begin{aligned} h : \mathbb{R}^Z &\longrightarrow \mathbb{R}^Z, \\ \mathcal{Z} &\longrightarrow h(\{\mathbf{z}^i\}) = [[o_{m,b}], p_m]^{\mathrm{T}}, \\ o_{m,b} &= \mathrm{median}(z^i_{m,b}), \\ p_m &= \mathrm{median}(t^i_m), \end{aligned} \tag{3}$$

where $Z \subset \{1, ..., M\} \times \{1, ..., B\}$, and $\mathbf{z}^i = \left[\left[z^i_{m,b}\right], t^i_m\right]^{\mathrm{T}}$.

Intuitively, this new video descriptor represents the median distance for each visual word to the codeword, since each BossaNova representation contains information about the distance-to-codeword distribution. Moreover, outliers are eliminated by the median function.

Also, it is important to note that the computational complexity of the proposed video descriptor is equivalent to the previous approach [11], since the computation and the number of BossaNova vectors remain the same. Furthermore, in the classification step, a single BossaNova vector is used, thus saving processing time.

In Figure 1, we illustrate the overview of our approach for pornography detection. It is important to note that this methodology is adapted from [10, 11], in which the binary image descriptors are computed followed by the computation of BossaNova representation for each keyframe. Our proposed video descriptor is computed according to Equation 3 for representing each video to be classified.

Our experiments show the advantage of our BossaNova video descriptor when compared to the previous BossaNova, including the majority voting proposed in [11].

4. EXPERIMENTS

To perform our experiments, we first use the PASCAL VOC 2007 dataset to validate our choice of using the BinBoost descriptor. Then, we assess our video descriptor proposal in the challenging real-world application of pornography detection.

[3]A keyframe is a frame that represents the content of a logical video unit, like a shot or scene, for example.

In the interest of a fair comparison, we apply the same feature extraction process. The five most recent and promising binary descriptors (BRIEF, ORB, BRISK, FREAK and BinBoost) are densely extracted at every 6 pixels. For all binary descriptors, except for BinBoost[4], we obtained the code from OpenCV's repository [25]. Also, we used the BossaNova code made available[5].

To learn the codebooks, we apply a k-medians clustering algorithm with Hamming distance over one million randomly sampled descriptors. That setup is also employed in our previous work [11]. Classification is performed by Support Vector Machines (SVM) classifier using a nonlinear kernel.

4.1. Results on PASCAL VOC 2007 dataset

The PASCAL VOC 2007 dataset [26] consists of 9,963 images of 20 object categories collected from Flickr. Those images are split into three subsets: training (2,501 images), validation (2,510 images) and test (4,952 images). Our experimental results are obtained on training+validation/test sets. The classification performance is measured by the mean Average Precision (mAP) across all classes.

Our goal for this experiment is to validate the BinBoost binary descriptor in conjunction with BossaNova representation. For that reason, we kept the BossaNova parameter values the same as in [11] ($B = 2$, $\lambda_{min} = 0.4$ and $\lambda_{max} = 2.0$, $s = 10^{-3}$, $M = 1024$). In addition, we assess the BinBoost descriptor varying among 8, 16 and 32 dimensions.

Table 1 shows the results of our experiments over PASCAL VOC 2007 dataset. We can notice that BinBoost descriptor with BossaNova outperforms the previous methods. Also, it should be noted that the 16-dimensional BinBoost gives the best result (mAP = 44.6%), surpassing even the 32-dimensional ones. In order to verify the statistical significance of those results, a statistical test for the differences between the means was performed using a Student t-test [27], paired over the dataset classes. The test consists of determining a confidence interval for the differences and simply checking if the interval includes zero, i.e, if the confidence interval does not include zero, the difference is significant at that confidence level. Thus, at 95% confidence level, the difference is significant for BinBoost16 and all others descriptors, except for BinBoost32. Thereafter, we decided to employ as low-level descriptor the 16-dimensional BinBoost.

4.2. Results on Pornography dataset

The Pornography dataset [10] contains nearly 80 hours of 400 pornographic and 400 non-pornographic videos. It comes already separated into 16,727 video keyframes. To make a fair comparison, we carefully followed the experimental protocol

[4]BinBoost is available at `http://www.cvlab.epfl.ch/research/detect/binboost`

[5]`http://www.npdi.dcc.ufmg.br/bossanova`

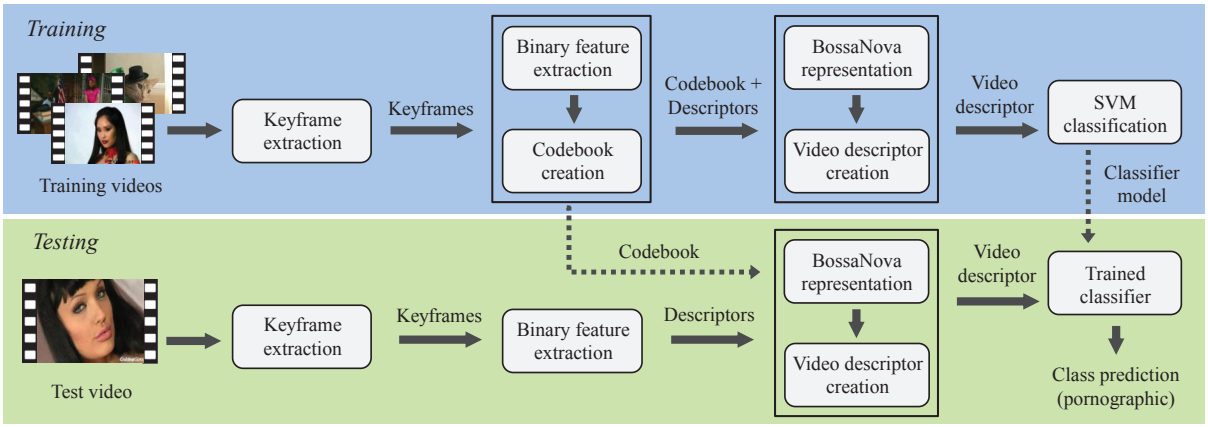

Fig. 1. Overview of our methodology for pornography detection with the proposed BossaNova Video Descriptor.

	Approach	mAP (%)
Published results	BossaNova (BRIEF) [11]	36.2
	BossaNova (ORB) [11]	37.1
	BossaNova (BRISK) [11]	38.0
	BossaNova (FREAK) [11]	33.3
Our results	BossaNova (BinBoost8)	43.2
	BossaNova (BinBoost16)	**44.6**
	BossaNova (BinBoost32)	44.2

Table 1. Image classification mAP (%) results of BinBoost + BossaNova and some published results that used binary descriptors on PASCAL VOC 2007 dataset [26].

	Approach	Acc. (%)
Published results	BossaNova (HueSIFT) [10]	89.5 ± 1
	BossaNova (BRIEF) [11]	86.3 ± 3
	BossaNova (ORB) [11]	86.5 ± 3
	BossaNova (BRISK) [11]	88.6 ± 2
	BossaNova (FREAK) [11]	86.9 ± 3
Our results	BossaNova VD (BRIEF)	89.0 ± 1
	BossaNova VD (ORB)	89.0 ± 1
	BossaNova VD (BRISK)	89.3 ± 1
	BossaNova VD (FREAK)	89.7 ± 2
	BossaNova VD (BinBoost16)	**90.9** ± 1

Table 2. Video classification (%) results (and standard deviations) of our approach and published results on Pornography dataset [10].

of the dataset, i.e., a 5-fold cross-validation (640 videos for training and 160 for testing on each fold). The video classification performance is reported by accuracy rate.

Here, our goal is to evaluate our BossaNova Video Descriptor (VD) on pornography detection and to compare it with previous published methods that also employed the BossaNova representation. Thus, we kept the BossaNova parameter values the same as in [10, 11]: $M = 256$, $B = 10$, $\lambda_{min} = 0$, $\lambda_{max} = 3$ and $s = 10^{-3}$.

Table 2 shows our results and the ones reported on the literature over the Pornography dataset. We can notice the considerable improvement obtained with our BossaNova VD, reaching 90.9% of accuracy with BinBoost16. Also, it is important to observe that our best result outperforms the best published one, as far as we know, which used HueSIFT descriptors [10]. *To the best of our knowledge, ours is the best result reported to date on Pornography dataset*.

The comparison with our previous work [11] is particularly relevant, because we employed the same binary descriptors (BRIEF, ORB, BRISK and FREAK with default parameters). We note an absolute improvement of (by up to) 2.8% from BossaNova to our BossaNova VD. That confirms the advantages introduced by our video descriptor.

Furthermore, we investigated the cases where our method failed. The misclassified non-pornographic videos correspond to very challenging cases, such as breastfeeding sequences, sequences of children being bathed, and beach scenes. Also, the analysis of the misclassified pornographic videos revealed that the method presented difficulties with poor quality videos or when the clip is borderline pornographic, with few explicit elements. The same difficulty was also reported by [10].

5. CONCLUSION

In this paper, we proposed a new video descriptor which employs local binary image descriptors in conjunction with the recent mid-level image representation, namely BossaNova.

We also validated the performances of BinBoost descriptor in conjunction with BossaNova on PASCAL VOC 2007 dataset, a benchmark in visual object category recognition. With respect to the video descriptor, we experimentally compared the performances of our approach to the published results on a real-world application of pornography detection.

Also, the use of BinBoost descriptor outperformed the previous binary descriptors-based-methods (up to nearly 6.6%) on PASCAL VOC 2007. For pornography classification, our approach also outperformed the state-of-the-art results.

In order to provide more comprehensive analysis of our video descriptor, we propose evaluating its behavior on other video classification problems. Furthermore, we propose exploring the key aspects of the parametric space of BossaNova in the visual recognition task.

6. ACKNOWLEDGMENTS

The authors are thankful to CNPq, CAPES and FAPEMIG, Brazilian research and development agencies, and the INCT InWeb, for the support to this work.

REFERENCES

[1] M. Cord and P. Cunningham, *Machine Learning Techniques for Multimedia: Case Studies on Organization and Retrieval*, Cognitive Technologies, Springer, 2008.

[2] C. Ries and R. Lienhart, "A survey on visual adult image recognition," *MTA*, pp. 1–28, 2012.

[3] M. Fleck, D. Forsyth, and C. Bregler, "Finding naked people," in *ECCV*, pp. 593–602. 1996.

[4] M. Jones and J. Rehg, "Statistical color models with application to skin detection," *IJCV*, pp. 81–96, 2002.

[5] J.-S. Lee, Y.-M. Kuo, P.-C. Chung, and E.-L. Chen, "Naked image detection based on adaptive and extensible skin color model," *Pattern Recognition*, vol. 40, no. 8, pp. 2261–2270, 2007.

[6] H. Zuo, W. Hu, and O. Wu, "Patch-based skin color detection and its application to pornography image filtering," in *WWW*, 2010, pp. 1227–1228.

[7] T. Deselaers, L. Pimenidis, and H. Ney, "Bag-of-visual-words models for adult image classification and filtering," in *ICPR*, 2008, pp. 1–4.

[8] A. Lopes, S. Avila, A. Peixoto, R. Oliveira, and A. de A. Araújo, "A bag-of-features approach based on Hue-SIFT descriptor for nude detection," in *EUSIPCO*, 2009, pp. 1552–1556.

[9] A. Lopes, S. Avila, A. Peixoto, R. Oliveira, M. Coelho, and A. de A. Araújo, "Nude detection in video using bag-of-visual-features," in *SIBGRAPI*, 2009, pp. 224–236, 10.1109/SIBGRAPI.2009.32.

[10] S. Avila, N. Thome, M. Cord, E. Valle, and A. de A. Araújo, "Pooling in image representation: the visual codeword point of view," *CVIU*, vol. 117, no. 5, pp. 453–465, 2013.

[11] C. Caetano, S. Avila, S. Guimarães, and A. de A. Araújo, "Representing local binary descriptors with BossaNova for visual recognition," in *ACM SAC*, 2014, pp. 49–54.

[12] E. Valle, S. Avila, A. da Luz, F. Souza, M. Coelho, and A. de A. Araújo, "Content-based filtering for video sharing social networks," *CoRR*, 2011, arXiv:1101.2427.

[13] T. Endeshaw, J. Garcia, and A. Jakobsson, "Classification of indecent videos by low complexity repetitive motion detection," in *AIPR*, 2008, pp. 1–7.

[14] C. Jansohn, A. Ulges, and T. Breuel, "Detecting pornographic video content by combining image features with motion information," in *ACM MM*, 2009, pp. 601–604.

[15] Y. Liu, X. Wang, Y. Zhang, and S. Tang, "Fusing audio-words with visual features for pornographic video detection," in *TrustCom*, 2011, pp. 1488–1493.

[16] D. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, pp. 91–110, 2004.

[17] J. Sivic and A. Zisserman, "Video Google: a text retrieval approach to object matching in videos," in *ICCV*, 2003, pp. 1–8.

[18] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "BRIEF: binary robust independent elementary features," in *ECCV*, 2010, pp. 778–792.

[19] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *ICCV*, 2011, pp. 2564–2571.

[20] S. Leutenegger, M. Chli, and R. Siegwart, "BRISK: Binary robust invariant scalable keypoints," in *ICCV*, 2011, pp. 2548–2555.

[21] R. Ortiz, "FREAK: Fast retina keypoint," in *CVPR*, 2012, pp. 510–517.

[22] V. Lepetit T. Trzcinski, M. Christoudias and P. Fua, "Boosting Binary Keypoint Descriptors," in *CVPR*, 2013, pp. 2874–2881.

[23] S. Avila, N. Thome, M. Cord, E. Valle, and A. de A. Araújo, "BOSSA: Extended BoW formalism for image classification," in *ICIP*, 2011, pp. 2909–2912.

[24] S. Avila, N. Thome, M. Cord, E. Valle, and A. de A. Araújo, "BossaNova at ImageCLEF 2012 flickr photo annotation task," in *Working Notes of the CLEF*, 2012, pp. 1–6.

[25] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[26] M. Everingham, L.-V. Gool, C. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results," 2007.

[27] R. Jain, *The art of computer systems performance analysis*, John Wiley & Sons, Inc., 1991.