FINGERPRINT INFORMATION MAXIMIZATION FOR CONTENT IDENTIFICATION [1]

Rohit Naini, Pierre Moulin

University of Illinois
Beckman Inst., Coord. Sci. Lab., & ECE Dept.
405 N. Mathews Ave., Urbana, IL 61801, USA.

ABSTRACT

This paper presents a novel design of content fingerprints based on maximization of the mutual information across the distortion channel. We use the information bottleneck method to optimize the filters and quantizers that generate these fingerprints. A greedy optimization scheme is used to select filters from a dictionary and allocate fingerprint bits. We test the performance of this method for audio fingerprinting and show substantial improvements over existing learning based fingerprints.

Index Terms— Audio fingerprinting, Information bottleneck, Information maximization, Content Identification

1. INTRODUCTION

Organization, matching and retrieval of multimedia content from internet scale databases pose significant challenges. With increasing popularity of content sharing and social networking websites, there is a data deluge where new multimedia content is created and existing contents are duplicated by users. It is desirable to have a scalable identification engine for efficient data organization and management. Content Identification (ID) has other well known applications such as copyright enforcement for sharing sites such as Youtube, content based searching such as Google Image search, and popular smartphone applications such as Shazam and Soundhound.

Fingerprint based methods for Content ID is an active area of research. Good fingerprints should have be compact to keep the database search complexity manageable while being highly distinguishable and robust to distortions introduced during operations such as transcoding and processing.

A variety of algorithms for content fingerprinting are readily available in literature. Examples of audio fingerprinting based on signal processing primitives include the Philips audio fingerprinting scheme [1] and Google's Waveprint [2]. Another body of research relies on machine learning methods to identify perceptually relevant features from a data dictionary [3]. Examples include, Symmetric Pairwise Boosting [4], and learning hash codes [5].

[1]Research supported by NSF under grant CCF 12-19145.

A formal approach to fingerprinting is presented in [6] where an information-theoretic relationship is derived between database size, hash length, and robustness that holds for any reliable, fingerprint-based, content ID system, under some structural assumptions on the fingerprinting code and a statistical assumption on the signals of interest. Our previous work in this area focuses on exploiting the underlying statistical models for fingerprints [7], in [8] we established a link between the information and learning theoretic aspects of Content ID. The problem of designing efficient quantizers for Content ID is also examined in [9, Ch. 2]. There fingerprint mutual information is used as the design criterion for quantizer design on Gaussian data with 4 quantizer levels. In contrast, our proposed scheme may be thought of as a practical application of the concepts of [6]. We use an *information bottleneck* scheme [13] for quantizer design which optimizes thresholds at multiple bit levels using 2-D joint kernel density estimates of actual filter outputs. We rely on estimates of mutual information of the fingerprint bits across common content distortion channels to identify optimal features and for the corresponding bit-allocations to generate the fingerprint.

The paper is organized as follows. Section 2 formalizes the content ID problem. Section 3 develops the information maximization framework and simplifies it into a tractable form. Section 4 presents the information bottleneck procedure and a greedy optimization for feature selection and bit-allocation. Section 5 employs this technique to audio fingerprinting and presents results for fingerprint matching performance. We conclude with a brief discussion in Section 6.

2. STATEMENT OF CONTENT ID

A *content database* is defined as a collection of M elements (content items) $\mathbf{x}(m) \in \mathcal{X}^N, \quad m = 1, 2, \cdots, M$, each of which is a sequence of N *frames* $\{x_1(m), x_2(m), \cdots, x_N(m)\}$. Here $\mathcal{X}$ is the set of possible values of a frame. A frame could be a short video segment, a short sequence of image blocks, or a short audio segment. Content frames may overlap spatially, temporally, or both. In audio fingerprinting, overlapping time windows that are 2 sec long and start every 185 ms; with overlap of 15/16 are used as frames in [4] . A

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

3-minute second song is represented by $N = 1000$ frames. It is desired that the audio be identifiable from a short segment, say 5 sec long, corresponding to $L = 16$ frames. This is called the granularity of the audio ID system [4]. Typically $L \ll N$.

The problem is to determine whether a given *probe* consisting of L frames, $\mathbf{y} \in \mathcal{X}^L$, is related to some element of the database, and if so, identify which one. To this end, an algorithm $\psi(\cdot)$ must be designed, returning the decision $\psi(\mathbf{y}) \in \{0, 1, 2, \cdots, M\}$ where $\psi(\mathbf{y}) = 0$ indicates that $\mathbf{y}$ is unrelated to any of the database elements.

In this paper, we consider a general class of fingerprints for content ID generated by frame-wise processing. The codes of [1, 4, 11], among others, fall in this category. Each of the M database elements $\mathbf{x} \in \mathcal{X}^N$ is fingerprinted using a mapping $\Phi : \mathcal{X} \to \mathcal{F}$ such that the fingerprint $\tilde{\mathbf{x}} \in \mathcal{F}^N$ with frame-wise components $\tilde{\mathbf{x}}^{(i)} = \Phi(x_i)$, $1 \le i \le N$; a probe $\mathbf{y}$ is encoded into a probe fingerprint $\tilde{\mathbf{y}} \in \mathcal{F}^L$ with components $\tilde{\mathbf{y}}^{(i)} = \Phi(y_i)$, $1 \le i \le L$; the decoder returns $\hat{m} = \psi(\tilde{\mathbf{y}})$ using a decoding function $\psi : \mathcal{F}^L \to \{0, 1, \cdots, M\}$.

Additional structure is imposed on the fingerprinting function Φ. The mapping $\Phi : \mathcal{X} \to \mathcal{F}$ is obtained by applying a set of K filters (optimized from a dictionary) to each frame and quantizing the kth filter output to 2^{b_k} levels such that $\sum_k b_k = B$. Hence $\mathcal{F}$ takes the form $\mathcal{F} = \prod_{k=1}^{K} \tilde{\mathcal{X}}_k$ with $\tilde{\mathcal{X}}_k = \{q_1, q_2, ..., q_{2^{b_k}}\}$. Thus the sub-fingerprint for the ith frame is given by the vector $\tilde{\mathbf{x}}^{(i)} = (\tilde{x}_1^{(i)}, \tilde{x}_2^{(i)}, ..., \tilde{x}_K^{(i)})$. The entire fingerprint takes the form of an array $\tilde{\mathbf{x}} = \{\tilde{x}_k^{(i)},\ 1 \le i \le N,\ 1 \le k \le K\}$. The decoding function is based on a sliding window estimation of the minimum Hamming distance between the arrays $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{x}}$.

3. FINGERPRINT INFORMATION MAXIMIZATION

For a pair $(\mathbf{x}, \mathbf{y})$ of related content elements, i.e. the probe $\mathbf{y}$ is obtained by passing $\mathbf{x}$ through a degradation channel. For the sake of simplicity, consider that the probe and database elements are perfectly aligned frame by frame. Assume each frame X is a random vector drawn from a distribution P_X over $\mathcal{X}$. Similarly, Y denotes the degraded frame drawn from a conditional distribution $P_{Y|X}$ over $\mathcal{X}$. Assume that the pairs of frames $(X^{(i)}, Y^{(i)})$ are iid.

Let $\tilde{\mathbf{X}} = (\tilde{X}_1, \tilde{X}_2, ..., \tilde{X}_K)$ denote a vector of random variables generated by applying to X a set of K filters Φ drawn from a dictionary $\mathbf{D}$, and quantizing the output of each filter ϕ_k to 2^{b_k} levels. The bit allocation satisfies $\sum_k b_k = B$, where B is the total fingerprint bit budget. Let $\tilde{\mathbf{Y}}$ denote the corresponding random vector obtained by applying the same set of filters to content Y. We seek to maximize the mutual information $I(\tilde{\mathbf{X}}; \tilde{\mathbf{Y}})$ between these fingerprint vectors. The problem can be stated as

$$(\Phi^*, \vec{b}^*) = \arg \max_{\Phi \subset \mathbf{D},\ \sum b_k = B} I(\tilde{\mathbf{X}}; \tilde{\mathbf{Y}}) \tag{1}$$

The solution to this problem yields the set of filters from the dictionary to be used and the optimal bit allocation. Unfortunately, this problem is combinatorial in nature and an approximate solution is needed. We simplify the problem by making Markov type approximations resulting in simplified conditional entropies

$$H(\tilde{X}_1|\tilde{Y}_1\tilde{Y}_2) \approx H(\tilde{X}_1|\tilde{Y}_1) \tag{2}$$
$$H(\tilde{X}_2|\tilde{Y}_2\tilde{Y}_1\tilde{X}_1) \approx H(\tilde{X}_2|\tilde{Y}_2) \tag{3}$$

This reflects the typical property that filter outputs of original and degraded frames from the same filter are more closely related than across filters. This approximation is validated by estimation of these conditional entropies from the training data. We plot the histogram of ratios $\frac{H(\tilde{X}_1|\tilde{Y}_1\tilde{Y}_2)}{H(\tilde{X}_1|\tilde{Y}_1)}$ and $\frac{H(\tilde{X}_2|\tilde{Y}_2\tilde{Y}_1\tilde{X}_1)}{H(\tilde{X}_2|\tilde{Y}_2)}$ for pairs of filters selected by the algorithm. It is seen from the histograms in Fig. 1 that with high probability, the approximation is accurate to 5% of the actual. The following analysis by considers $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$ for just two filters and approximates the mutual information as follows.

$$\begin{aligned}
I(\tilde{\mathbf{X}}; \tilde{\mathbf{Y}}) &= I(\tilde{X}_1\tilde{X}_2; \tilde{Y}_1\tilde{Y}_2) \\
&= H(\tilde{X}_1\tilde{X}_2) - H(\tilde{X}_1\tilde{X}_2|\tilde{Y}_1\tilde{Y}_2) \\
&= H(\tilde{X}_1) + H(\tilde{X}_2) - I(\tilde{X}_1; \tilde{X}_2) \\
&\quad -H(\tilde{X}_1|\tilde{Y}_1\tilde{Y}_2) - H(\tilde{X}_2|\tilde{Y}_2\tilde{Y}_1\tilde{X}_1) \\
&\approx H(\tilde{X}_1) + H(\tilde{X}_2) - I(\tilde{X}_1; \tilde{X}_2) \\
&\quad -H(\tilde{X}_1|\tilde{Y}_1) - H(\tilde{X}_2|\tilde{Y}_2) \\
&= I(\tilde{X}_1; \tilde{Y}_1) + I(\tilde{X}_2; \tilde{Y}_2) - I(\tilde{X}_1; \tilde{X}_2)
\end{aligned}$$

This approximation is shown in Fig. 2. The shaded regions of intersection are approximated to zero.

Extending the approximation to K filters, and characterizing higher order common information [12] terms like $I(\tilde{X}_1; \tilde{X}_2; \tilde{X}_3)$ using a tuning parameter β we have the following approximation to the optimization in (1):

$$\begin{aligned}
(\Phi^*, \vec{b}^*) = \arg \max_{\Phi \subset \mathbf{D},\ \sum b_k = B} \Bigg[&\sum_{k:\phi_k \in \Phi} I(\tilde{X}_k; \tilde{Y}_k) \\
&-\beta \sum_{k \ne l:\phi_{k,l} \in \Phi} I(\tilde{X}_k; \tilde{X}_l) \Bigg].
\end{aligned} \tag{4}$$

Typically, $\beta \in [0, 1)$ depending on the actual values of the higher order common information terms. When $\beta = 0$, the algorithm neglects the dependencies between responses across filters.

4. ITERATIVE OPTIMIZATION METHOD

In order to solve the optimization problem in (4), we need to estimate the mutual information terms $I(\tilde{X}_k; \tilde{Y}_k)$ and $I(\tilde{X}_k; \tilde{X}_l)$ for the filters in the dictionary $\mathbf{D}$. The goal is to

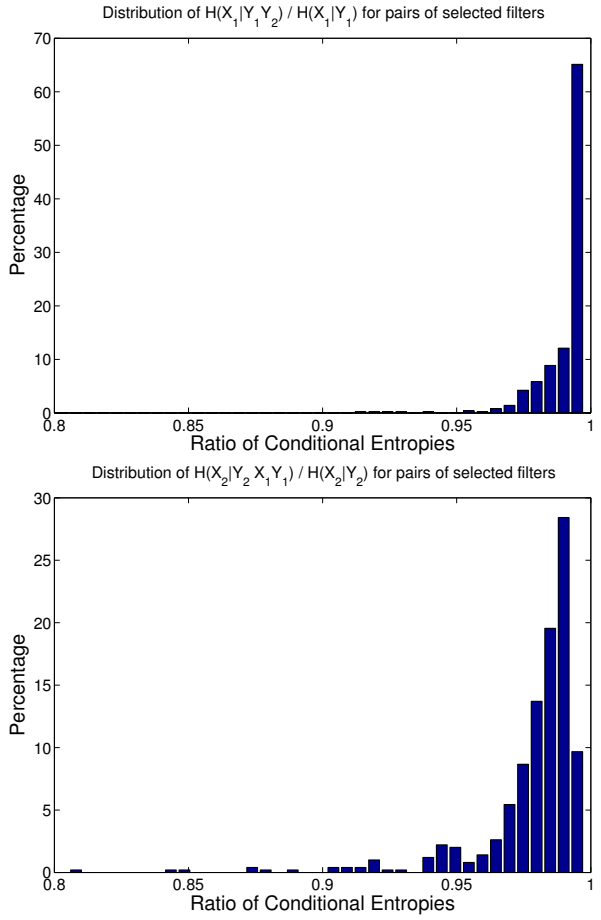

Fig. 1: Histograms showing the accuracy of conditional entropy approximations (2)and (3).

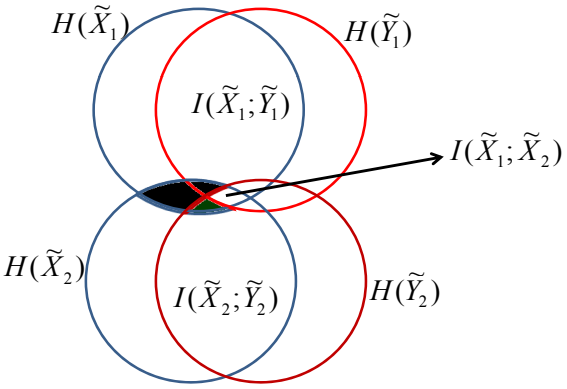

Fig. 2: Fingerprint Vectors Mutual Information Venn Diagram

identify the filters that retain high individual mutual information as represented by the $I(\tilde{X}_k;\tilde{Y}_k)$ terms while simultaneously accounting for the mutual information loss due to $I(\tilde{X}_k;\tilde{X}_l)$. The training dataset is used to estimate the two dimensional joint statistics of the filter outputs. Bandwidth tuned 2 dimensional kernel density estimation is performed on a finely quantized grid using real valued filter outputs.

The optimization problem is solved in three steps. Step 1, solves the problem of optimal quantization of filter outputs given an arbitrary number of quantization levels. This is done using a variation to the information bottleneck method [13]. Following this, Step 2 performs a greedy filter selection process based on mutual information estimates obtained in Step 1. Step 3 allocates fingerprint bits to the filters selected in Step 2 using estimates of $I(\tilde{X}_k,\tilde{Y}_k)$ and $I(\tilde{X}_k,\tilde{Y}_l)$ and quantization thresholds obtained in Step 1. Thus, sub-fingerprints are extracted for each content frame using the selected filters and their quantization thresholds.

1) The Information Bottleneck Method: The information bottleneck method [13] provides an scheme to quantize a scalar random variable X into $\tilde{X}$ such that its relevance to a random variable Y, given by $I(\tilde{X};Y)$, is maximized. We have a slight variation of this formulation as both X_k and Y_k, need to be quantized using an identical quantization scheme such that the mutual information $I(\tilde{X}_k;\tilde{Y}_k)$ is maximized. We use the agglomerative clustering method to generate a hard quantization scheme as shown in [14]. The only input to this clustering algorithm is the estimate of the 2-dimensional joint probability density $p(x_k,y_k)$, outputs of ϕ_k on a finely quantized grid. The method repeatedly cycles through pairs of adjacent quantization bins and collapses two levels that lead to least information loss. The output of this algorithm provides the estimate of $I(\tilde{X}_k,\tilde{Y}_k)$ for every possible number of quantization levels. The algorithm generates optimal quantizer thresholds under the hard quantization assumption [14].

2) Filter support set selection: The filters are selected greedily from the dictionary assuming a fixed number of quantization bits per filter $b_k = 3, \forall k$ in this step. After each iteration, the entire dictionary of candidate filters is examined for two terms, the mutual information $I(\tilde{X}_k;\tilde{Y}_k)$ at 3 bits quantization and the interaction of ϕ_k with previously selected filters within the support set. At iteration T, this second term is given by $\sum_{t=1}^{T-1} I(\tilde{X}_k,\tilde{X}_{\phi_t^*})$ calculated using the joint statistics generated from the training dataset and quantizer thresholds obtained in Step 1. The selection of filter k at iteration T is given by

$$\phi_T^* = \arg\max_{\phi_k \in \mathbf{D}} \left[I(\tilde{X}_k;\tilde{Y}_k) - \beta \sum_{t=1}^{T-1} I(\tilde{X}_k,\tilde{X}_{\phi_t^*}) \right]. \quad (5)$$

The process is repeated until K filters are selected as support set for fingerprint generation.

3) Bit allocation: Bits are allocated to each of the K filters selected in Step 2. Initially one bit per filter is pre-allocated.

Then at each iteration, an additional bit is allocated to the filter that most increases the cost function of (1). Quantization thresholds generated using the *information bottleneck* scheme in Step 1 are reused to estimate mutual information terms in the cost as a function of number of bits. This process iterates until the entire bit budget B is allocated. In general, we allow the maximum number of bits available per filter to exceed the average bit budget in order to leverage filters with superior performance.

5. EXPERIMENTAL RESULTS

To validate the performance of the proposed algorithm, we build a fingerprinting scheme for audio content and compare the matching characteristics on a large database of audio content. Our audio database consists of $M = 1000$ music files, each duplicated under 5 different perceptual distortion channels: Bandpass filtering, 20 percent echo addition, Equalization, Flanging, and Pitch changes. For the training and testing purposes, the database of total 7 million frames of 376 ms each is split into 2 sets of 10^5 disjoint 10 frame, 2.5 second snippets. All the statistical modelling to estimate the joint densities are performed on the training set.

Audio spectrograms provide an adequate representation for these audio snippets. A short time Fourier transform (STFT) is performed on each snippet with a temporal window of 376 ms corresponding to the frame width. For added robustness, we capture the Spectral Sub-band Centroids (SSC) using 16 bins on the Bark scale as described in [4]. Therefore, each audio snippet is transformed to a 16×10 real valued SSC image.

The audio fingerprints are built from a dictionary of Viola-Jones type filters [3]. These filters capture perceptual aspects of audio such as spatio-temporal shifts in energy. The dictionary consists of 3808 filters constructed across 5 types of filters in various sizes and shifts of application to the SSC.

We apply our fingerprinting algorithm with $K = 32$ filters and a total bit budget of $B = 64$ bits. For the estimation of joint densities to feed to the agglomerative clustering, we use 2 dimensional kernel density estimation [15] on training data to obtain joint pdf values on a 128×128 grid spanning the range of outputs for each filter. The optimization using the information bottleneck method takes about 90 minutes on a 2.2 Ghz Core i7 processor using an unoptimized MATLAB implementation. We build the fingerprints using different values of the tuning parameter $\beta = 0, 0.1, 0.5, 1.0$.

For matching fingerprints, we use a simple Hamming distance based thresholding to declare a pair of snippets as a match or a non-match using an exhaustive search over the database. Alternatively, we can use Euclidean distance with marginal performance improvement but prefer Hamming distance due its reduced storage and ease of matching binary fingerprints using bitwise operations. The matching performance is evaluated on the testing database using 10^5 song snippets to plot the ROC curve by varying the matching threshold. We compare our results to a a fingerprint scheme built Symmetric Pairwise Boosting (SPB) in [4] and 2-bit information maximizing quantization (2b-IM) in [9]. For fairness of comparison, both SPB and 2b-IM use 32 boosted filters with 2 bits allocated per filter. Qualitatively, the filters selected by both SPB and our algorithm are ones with a long temporal support and short support in the frequency direction. The improved performance from our method is due to better quantization and selection of filters which are nearly independent and don't share significant time-frequency support with each other. The bit allocation in Step 3 results in 3,8,7,14 filters with 4,3,2 and 1 bit respectively when $\beta = 0.1$. We show the efficacy of our bit allocation step by comparing performance with a uniform 2-bit allocation per filter. The results are shown in Fig. 3. We observe that ROC performance

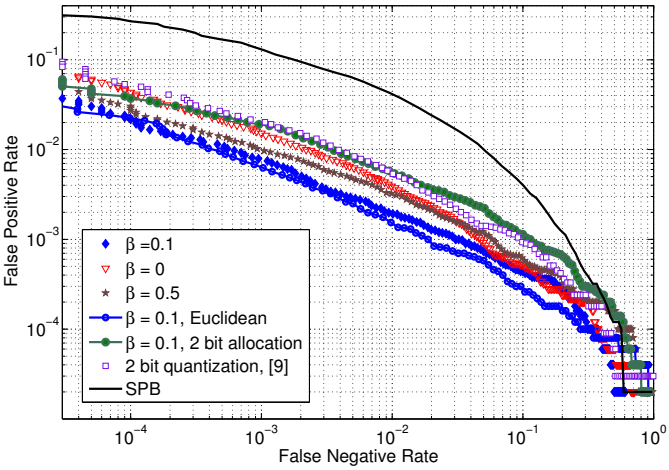

Fig. 3: ROC curves for matching audio fingerprints across a distortion channel

varies depends on β, since β trades off the two different aspects of the selected filters. It is seen that when $\beta = 0.5$, a heavy penalty is levied on fingerprint correlation which leads to a suboptimal choice of filters and to inferior performance. From our experiments, it is seen that the best performance is achieved when $\beta = 0.1$.

6. SUMMARY

We have presented a method for designing fingerprints that maximizes a mutual information metric. The problem is high dimensional, so we have proposed a greedy optimization method that relies on the information bottleneck (IB) method. Our algorithm selects filters from a dictionary and optimizes bit allocation to these filters. We have demonstrated the validity of this technique by testing it for audio fingerprinting. Our method outperforms the benchmark learning based method of [4].

7. REFERENCES

[1] J. Haitsma and T. Kalker, "A highly robust audio fingerprinting system.," in *ISMIR*, 2002.

[2] S. Baluja and M. Covell, "Waveprint: Efficient wavelet-based audio fingerprinting," *Pattern recognition*, vol. 41, no. 11, pp. 3467–3480, 2008.

[3] Y. Ke, D. Hoiem, and R. Sukthankar, "Computer vision for music identification," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. IEEE, 2005, vol. 1, pp. 597–604.

[4] D. Jang, C. D. Yoo, S. Lee, S. Kim, and T. Kalker, "Pairwise boosted audio fingerprint," *Information Forensics and Security, IEEE Transactions on*, vol. 4, no. 4, pp. 995–1004, 2009.

[5] S. Baluja and M. Covell, "Learning to hash: forgiving hash functions and applications," *Data Mining and Knowledge Discovery*, vol. 17, no. 3, pp. 402–430, 2008.

[6] P. Moulin, "Statistical modeling and analysis of content identification," in *Information Theory and Applications Workshop (ITA), 2010*. IEEE, 2010, pp. 1–5.

[7] R. Naini and P. Moulin, "Model-based decoding metrics for content identification," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 1829–1832.

[8] R. Naini and P. Moulin, "Real adaboost for content identification," in *Statistical Signal Processing Workshop (SSP), 2012 IEEE*. IEEE, 2012, pp. 756–759.

[9] A.L. Varna, *Multimedia Protection using Content and Embedded Fingerprints*, Ph.D. thesis, University of Maryland, July 2011.

[10] N. Slonim and N. Tishby, "Document clustering using word clusters via the information bottleneck method," in *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2000, pp. 208–215.

[11] S. Lee, C. D. Yoo, and T. Kalker, "Robust video fingerprinting based on symmetric pairwise boosting," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 19, no. 9, pp. 1379–1388, 2009.

[12] W. J. McGill, "Multivariate information transmission," *Psychometrika*, vol. 19, no. 2, pp. 97–116, 1954.

[13] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," *arXiv preprint physics/0004057*, 2000.

[14] N. Slonim, N. Friedman, and N. Tishby, "Agglomerative multivariate information bottleneck," in *Advances in neural information processing systems*, 2001, pp. 929–936.

[15] Z.I. Botev, J.F. Grotowski, and D.P. Kroese, "Kernel density estimation via diffusion," *The Annals of Statistics*, vol. 38, no. 5, pp. 2916–2957, 2010.