

# Multimodal Video Indexing and Retrieval Using Directed Information

Xu Chen, Alfred Hero and Silvio Savarese

Department of Electrical Engineering and Computer Science  
University of Michigan at Ann Arbor, Ann Arbor, MI, USA  
{xhen, hero}@umich.edu, silvio@eecs.umich.edu

**Abstract**—We propose a novel framework for multimodal video indexing and retrieval using shrinkage optimized directed information assessment (SODA) as similarity measure. The directed information (DI) is a variant of the classical mutual information which attempts to capture the direction of information flow that videos naturally possess. It is applied directly to the empirical probability distributions of both audio-visual features over successive frames. We utilize RASTA-PLP features for audio feature representation and SIFT features for visual feature representation. We compute the joint probability density functions of audio and visual features in order to fuse features from different modalities. With SODA, we further estimate the DI in a manner that is suitable for high dimensional features  $p$  and small sample size  $n$  (large  $p$  small  $n$ ) between pairs of video-audio modalities. We demonstrate the superiority of the SODA approach in video indexing, retrieval and activity recognition as compared to the state-of-the-art methods such as Hidden Markov Models (HMM), Support Vector Machine (SVM), Cross-Media Indexing Space (CMIS) and other non-causal divergence measures such as mutual information (MI). We also demonstrate the success of SODA in audio and video localization and indexing/retrieval of data with misaligned modalities.

**Index Terms**—Multimedia content retrieval, audio-video pattern recognition, shrinkage optimization, overfitting prevention, non-linear information flow, multimodal feature fusion.

## I. INTRODUCTION

IN large-scale video analysis, mutual dependency between pairs of video documents is usually directed and asymmetric: past events influence future events but not conversely. This is mainly because purposeful human behavior generates some of the most highly complex non-linear patterns of directed dependency. Moreover, the content of a video is intrinsically multimodal including visual, auditory and textual channels, which provides different types of channels to convey the meaning of multimedia information to users [31]. For example, it would be difficult to reliably distinguish action movies from detective movies if only the visual information is considered. Combining evidence from multiple modalities for video indexing and retrieval has been shown to improve the accuracy in several applications, including combining overlay text, motion, and audio [14] [7]. To cater to these diverse challenges and applications, model-free information theoretic approaches have been previously proposed to discriminate complex human activity patterns but have only had limited success. What is needed is a different measure of information that is more sensitive to strongly directed non-linear dependencies in human activity events with different modalities. This paper proposes

such a measure, directed information (DI), and introduces a DI estimation approach, shrinkage optimized directed information assessment (SODA), that is well suited to the high dimensional setting of recognition, indexing and retrieval of human activity by fusing the information from different modalities in a video document. Since a single modality does not provide sufficient information for accurate indexing, the DI estimator is adapted to fusion of features from the multiple modalities. The DI is conceptually straightforward, is of low implementation complexity, and is optimal in the mean-square sense over the class of regularized DI estimators. The DI reduces to the log of Granger’s pairwise causality measure under the assumptions that the multivariate video features are stationary and Gaussian. Furthermore, our experiments demonstrate that the performance of the fusion algorithm based on DI on indexing/retrieval tasks and activity recognition tasks is superior to previously proposed methods based on hidden Markov models, (symmetric) mutual information, Cross-Media Indexing Space and SIFT-bag Kernels.

The proposed SODA approach is a natural evolution of previous information theoretic approaches to video event analysis. Zhou et al [38] proposed the Kullback-Leibler divergence as a similarity measure between SIFT features for video event analysis. The work [19] by Liu and Shah applied Shannon’s mutual information (MI) to human action recognition in videos. The work [7] by Fisher and Darrell utilize mutual information between pairs of audio and video signals for cross-modal audio and video localization. Sun and Hoogs [33] utilized compound disjoint information as a metric for image comparison. However, the similarity measures used by these methods do not exploit the transactional nature of human behavior: people’s current behavior is affected by what they have observed in the past [8]. The proposed SODA approach is specifically designed to exploit this directionality in information flow under a minimum of model assumptions.

SODA fuses audio-visual signals by estimation of the joint probability distribution of audio and visual features. Thus, our SODA estimator is completely data-driven: different from event and activity recognition approaches based on key regions detection [15], Markov chains [13], graphical model-based learning [22] or fusion algorithms based on semantic features [12], it relies solely on a non-parametric regularized estimate of the joint probability distribution. Like other non-parametric approaches to indexing/retrieval and event recognition [38], [19], [37], [34], [25], it differs from other model-based meth-

ods for multimodal integration such as hidden Markov models (HMM) [26] [36] [14]. Using TRECVID 2010 human activity video databases, our experiments show that SODA performs indexing and retrieval significantly better than SVM [18] and MI [19] approaches. We also show that SODA outperforms HMM models for activity recognition.

As an analog of Shannon’s MI, the DI was initially introduced by Massey in 1990 [21] as a variant of mutual information that can account for feedback in communication channels. The DI has been applied to the analysis of gene influence networks [28]. *As far as we know this paper represents the first application of DI to multimodal video indexing and retrieval.* Due to the intrinsic complexity of audio and visual features and high dimensionality of the joint feature distribution, the implementation of the DI for fusion of audio and visual features is a challenging problem. In particular, as explained below, a standard empirical implementation of DI estimator suffers from severe overfitting errors. We minimize these overfitting errors with a novel estimator regularization technique.

Similar to MI, DI is a function of the time-aggregated feature densities extracted from a pair of sequences shown in Fig.1. We use the popular Relative Spectra Transform-Perceptual Linear Prediction (RASTA-PLP) for speech feature representation [10] [11] due to their superiority in smoothing over short-term noise variations. We utilize SIFT features for visual feature representation [20], due to their invariance to image scale, rotation and other effects, and the bag of visual words (BOW) model [24] for representing image content in each frame. Implementing DI requires estimates of the joint distribution of the merged RASTA-PLP and bag of words based on SIFT features. Fig.2 illustrates the details of the feature fusion. To estimate these high dimensional feature distributions we apply James-Stein shrinkage regularization methods. Shrinkage estimators reduce mean-squared error (MSE) by shrinking the histogram towards a target, e.g. a uniform distribution. Such a shrinkage approach was adopted by Hauser and Strimmer [9] for entropy estimation. We extend this approach to DI, obtaining an asymptotic expression for the MSE and use this expression to compute an optimal shrinkage coefficient. The extension is non-trivial since it requires an approximation to the bias and variance of the more complicated directed information function.

It is helpful to note that our proposed SODA has advantages over the classical Granger measures of causal influence between two random processes [16] [2] [27]. Different from SODA, Granger causality [16] tends to capture causal influence by computing the residual prediction errors of two linear predictors: one utilizes the previous samples of both processes and another utilizes only the previous samples of one of the processes. The original Granger causality measure [16] was limited to stationary Gaussian time series. These assumptions are slackened in later versions. However, due to non-stationarity and non-linearity of the dependency structure of interesting human activities, classical Granger measures are suboptimal. Our SODA approach can be viewed as an optimized non-parametric and non-linear extension of parametric and linear Granger measures of causality. SODA accounts

for non-linear dependencies while reducing to the classical Granger measure in the case that the processes are jointly Gaussian.

We show experimental results on the TRECVID 2010 video databases that demonstrate the capabilities of SODA for activity recognition, indexing and retrieval, and video-audio temporal and spatial localization. Specifically we show: (1) Use of SODA as a video indexing/retrieval similarity measure results in at least 7% improvement in precision-recall performance as compared to unregularized DI, PCA regularized DI, MI, SVM and cross-media indexing as measured by the area under the curve (AUC) of the precision-recall curve. (2) By plotting the evolution of the DI over time we can accurately localize the emergence of strongly causal interactions between activities in a pair of videos. The DI’s activity recognition performance is as good as or better than HMM-based fusing algorithms for audio-visual features whose emission probabilities are implemented with Kernel Density estimates (KDE) or Gaussian Mixture Models (GMM). (3) SODA improves in terms of average precision by more than 8% compared to MI when used for spatial temporal similarities in localizing audio and video signals.

## II. RELATED WORK

Extensive research efforts have been invested in multimodal video indexing and retrieval problems. Early work on multimodal video indexing used SVM and HMM approaches to multimodal video indexing [14] [18]. The authors in [14] propose different methods for integrating audio and visual information for video classification of TV programs based on HMM. In [18], text features from closed-captions and visual features from images are combined to classify broadcast news videos using meta-classification via SVM. Recently, Snoek and Worring [32] proposed the time interval multimedia event (TIME) framework as a robust approach for classification of semantic events in multimodal video documents. The representation used in TIME extends the Allen temporal interval relations [1] and allows for proper inclusion of context and synchronization of the heterogeneous information sources involved in multimodal video analysis. More recently, the authors in [35] [39] used semantic correlations among multimedia objects of different modalities for cross-media indexing. In cross-media indexing and retrieval, the query examples and retrieval results need not to be of the same media type. For example, users can query images by submitting either an audio example or an image example in cross media retrieval systems. In [39] a correlation graph is built for the media objects of different modalities and a scoring technique is utilized for retrieval. In [35], for each query, the optimal dimension of cross-media indexing space (CMIS) is automatically determined from training data and the cross-media retrieval is performed on a per-query basis. In [29], Rasiwasia et al. resolved the problem of jointly modeling the text and image components of multimedia documents. Correlations between the two components are learned using canonical correlation analysis and abstraction is achieved by representing text and images at a more general, semantic level.

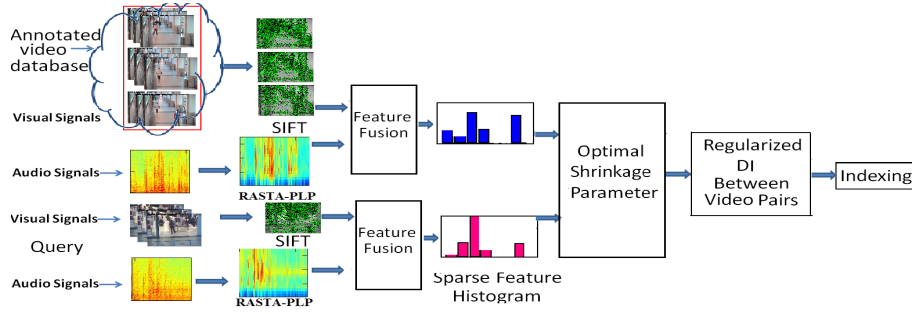


Fig. 1. Block diagram of shrinkage optimized directed information (SODA) for fusion of audio and visual features for video indexing.

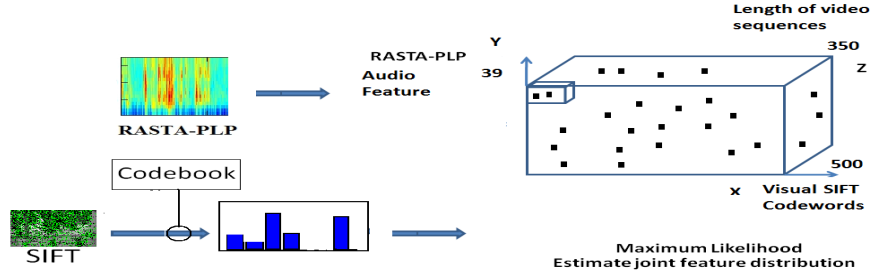


Fig. 2. Visual illustration of the process of fusing audio and visual features where the visual features are obtained from a visual codebook using bag of words (BOW) based on SIFT features. The joint probability density functions which define DI are estimated from multidimensional histograms computed from these cubes obtained from audio features and visual features by counting the number of instances (black square in the figure) falling into each subcube.

It is shown in [29] that accounting for both crossmodal correlations and semantic abstraction improve retrieval accuracy. Unlike the above papers, this paper uses a generalized measure of correlation, the directed information, between multimodal (audio and video) data streams to achieve better classification and retrieval performance.

### III. PROBLEM FORMULATION

Here we propose a DI estimator that is specifically adapted to video and audio sources. Given discrete features  $X$  and  $Y$  we use the multidimensional histogram for the fusion of SIFT and RASTA-PLP features. Continuous features are discretized by quantization over a codebook. The dimension of the joint feature distribution must be sufficiently large to adequately represent inter-frame object interactions as well as capture the variability of appearance and audio across videos within the same class [23]. This high dimension would lead to high variance DI estimates unless adequate countermeasures are taken. We propose using an optimal regularized DI estimation strategy to control estimator variance.

The feature fusion is implemented for bag of words (BOW) based on SIFT and RASTA-PLP features in each video frame as shown in Fig. 2. For a single frame the codebook has an alphabet of  $p$  symbols  $\mathcal{X} = \{x_i\}_{i=1}^p$  corresponding to  $p$  quantization cells (classes)  $\mathcal{C} = \{C_i\}_{i=1}^p$ . The codebook produces the  $i$ -th symbol  $x_i$  when the feature lies in quantization cell  $C_i$ ,  $i = 1, \dots, p$ . For a video sequence  $X^{(m)} = \{X_1, \dots, X_m\}$ , the codebook for the joint feature distribution has  $p^m$  output levels in  $\mathcal{X} \times \dots \times \mathcal{X} \subset \mathbb{R}^m$  and quantization cells  $\mathcal{C} \times \dots \times \mathcal{C} \subset \mathbb{R}^m$ . For a particular frame sequence  $X^{(m)}$  let there be  $n$  i.i.d. feature realizations and let  $Z = [z_1, \dots, z_{p^m}]$  denote the

histogram of these realizations over the respective quantization cells. Then  $Z$  is multinomial distributed with probability mass function

$$P_\theta(z_1 = n_1, \dots, z_{p^m} = n_{p^m}) = \frac{n!}{\prod_{k=1}^{p^m} n_k!} \prod_{k=1}^{p^m} \theta_k^{n_k},$$

where  $\theta = E[Z]/n = [\theta_1, \dots, \theta_{p^m}]$  is a vector of class probabilities and  $\sum_{k=1}^{p^m} n_k = n$ ,  $\sum_{k=1}^{p^m} \theta_k = 1$ .

We consider two multimodal video sequences  $V_x$  and  $V_y$  with  $M_x$  and  $M_y$  frames, respectively. Denote by  $X_m = \{X_{m,a}, X_{m,v}\}$  and  $Y_m = \{Y_{m,a}, Y_{m,v}\}$  the audio and visual feature variables extracted from the  $m$ -th frames of  $V_x$  and  $V_y$ , respectively, where the audio-visual feature is obtained by estimating the joint distribution of the audio and visual features. Define  $X^{(m,a)} = \{X_{k,a}\}_{k=1}^m$  and  $Y^{(m,a)} = \{Y_{k,a}\}_{k=1}^m$  for audio features.  $X^{(m,v)} = \{X_{k,v}\}_{k=1}^m$  and  $Y^{(m,v)} = \{Y_{k,v}\}_{k=1}^m$  for visual features. Further define  $X^{(m)} = \{X_k\}_{k=1}^m$  and  $Y^{(m)} = \{Y_k\}_{k=1}^m$  for fused features. The mutual information (MI) between  $V_x$  and  $V_y$  is

$$\text{MI}(V_x; V_y) = E \left[ \ln \frac{f(X^{(M_x)}, Y^{(M_y)})}{f(X^{(M_x)})f(Y^{(M_y)})} \right],$$

where

$$f(X^{(M_x)}, Y^{(M_y)}) = f(X^{(M,a)}, X^{(M,v)}, Y^{(M,a)}, Y^{(M,v)})$$

is the joint distribution for fusion of the audio and video features for both the sequences  $V_x$  and  $V_y$ , and  $f(X^{(M_x)}) = f(X^{(M,a)}, X^{(M,v)})$  and  $f(Y^{(M_y)}) = f(Y^{(M,a)}, Y^{(M,v)})$  are joint distributions of audio-visual features for each sequence. The time-aligned directed information (DI) from  $V_x$  to  $V_y$  is

a non-symmetric generalization of the MI defined as [21]

$$DI(V_x \rightarrow V_y) = \sum_{m=1}^M I(X^{(m)}; Y_m | Y^{(m-1)}) \quad (1)$$

where  $M = \min\{M_x, M_y\}$ ,  $I(X^{(m)}; Y_m | Y^{(m-1)})$  is the conditional MI between  $X^{(m)}$  and  $Y_m$  given the past  $Y^{(m-1)}$

$$I(X^{(m)}; Y_m | Y^{(m-1)}) = E \left[ \ln \frac{f(X^{(m)}, Y_m | Y^{(m-1)})}{f(X^{(m)} | Y^{(m-1)}) f(Y_m | Y^{(m-1)})} \right]$$

and  $f(W|Z)$  denotes the conditional distribution of random variable  $W$  given random variable  $Z$ . An equivalent representation of DI (1) is in terms of conditional entropies

$$DI(V_x \rightarrow V_y) = \sum_{m=1}^M [H(Y_m | Y^{(m-1)}) - H(Y_m | Y^{(m-1)}, X^{(m)})],$$

which implies that the DI is the cumulative reduction in uncertainty of frame  $Y_m$  when the past frames  $Y^{(m-1)}$  of  $V_y$  are supplemented by information about the past and present frames  $X^{(m)}$  of  $V_x$ . Using the equivalent representation of DI (1) in terms of unconditional entropy

$$\begin{aligned} DI_\theta(V_x \rightarrow V_y) = & \\ & \sum_{m=1}^M \left( H_\theta(X^{(m)}, Y^{(m-1)}) - H_\theta(Y^{(m-1)}) \right) \\ & - \sum_{m=1}^M \left( H_\theta(X^{(m)}, Y^{(m)}) - H_\theta(Y^{(m)}) \right), \quad (3) \end{aligned}$$

the DI can be computed explicitly from the entropy expression for a multinomial random variable  $W$  over  $P$  classes with class probabilities  $\theta = \{\theta_k\}_{k=1}^P$

$$H_\theta(W) = -n \sum_{k=1}^P \theta_k \ln \theta_k,$$

with  $W$  representing one of the four vectors  $[X^{(m)}, Y^{(m-1)}]$ ,  $[Y^{(m)}, X^{(m)}]$ ,  $Y^{(m)}$ , or  $Y^{(m-1)}$ . To estimate the DI in (3), the vector of multinomial parameters  $\theta$  must be empirically estimated from the audio and video sequences. However, due to the large size of the codebook, the multidimensional joint feature histograms are high dimensional and the number of unknown parameters  $p^m$  exceeds the number of feature instances  $n$ . A plug-in maximum likelihood (ML) estimator for  $\theta$  in the expression (3), will therefore suffer severely from high variance due to this high dimensional DI. Specifically, given  $n$  realizations  $\{W_i\}_{i=1}^n$  of the audio-visual feature vector  $W = [X^{(M_x)}, Y^{(M_y)}]$  the ML estimator of the  $k$ -th class probability  $\theta_k$  is  $\hat{\theta}_k = n^{-1} \sum_{i=1}^n I(W_i \in C_k)$ ,  $k = 1, \dots, p^{M_x+M_y}$ . Since  $n \ll p^{M_x+M_y}$ , most  $\hat{\theta}_k$ 's will be equal to zero, leading to overfitting error.

To mitigate high variance, we apply a James-Stein shrinkage approach. A related approach was adopted in [9] for entropy and MI estimation, which is based on shrinking the ML estimator of  $\theta$  towards a target distribution  $t = [t_1, \dots, t_{p^{M_x+M_y}}]$  as,

$$\hat{\theta}_k^\lambda = \lambda t_k + (1 - \lambda) \hat{\theta}_k^{ML}, \quad (4)$$

where  $\lambda \in [0, 1]$  is a shrinkage coefficient. The James-Stein plug-in entropy estimator is defined as:

$$\hat{H}_{\hat{\theta}^\lambda}(X) = -n \sum_{k=1}^P \hat{\theta}_k^\lambda \log(\hat{\theta}_k^\lambda). \quad (5)$$

The corresponding plug-in estimator for DI is simply  $\widehat{DI}^\lambda = \widehat{DI}_{\hat{\theta}^\lambda}(V_x \rightarrow V_y)$  where  $\lambda$  is selected to optimize DI performance. The oracle value of  $\lambda$  minimizes estimator MSE:

$$\lambda^\circ = \arg \min_{\lambda} E(\widehat{DI}^\lambda - DI)^2. \quad (6)$$

The oracle SODA estimator is  $\widehat{DI}^{\lambda^\circ}(X^M \rightarrow Y^M)$ . The MSE in (6) can be decomposed as  $MSE = Bias^2 + Variance$ . The theoretical expressions for bias and variance, given Propositions 1 and 2 in the appendix, will be used to determine the relationship between MSE and the shrinkage coefficient  $\lambda$ . The oracle  $\lambda^\circ$  can then be calculated by minimizing  $MSE = C_1^2 + (2C_1C_2 + T_2\Sigma_2T_2')\frac{1}{n} + O(\frac{1}{n^2})$  over  $\lambda$ , where expressions for  $C_1, C_2, T_2, \Sigma_2$  are given in Propositions 1 and 2. The oracle shrinkage parameter  $\lambda^\circ$  is determined by applying a gradient descent algorithm to numerically minimize the MSE. It can be shown that the oracle shrinkage parameter  $\lambda^\circ$  in equation (6) converges to 0 with increasing numbers of samples  $n$ . As is customary in James-Stein approaches, an empirical estimate of the oracle  $\lambda^\circ$  is obtained by replacing each of the terms  $C_1, C_2, T_2, \Sigma_2$  with their empirical maximum likelihood estimates. We call this empirical estimator of  $\lambda^\circ$  the optimal shrinkage parameter.

#### IV. IMPLEMENTATION OF SODA INDEXING/RETRIEVAL AND RECOGNITION ALGORITHM

A simple flow chart of our implementation of SODA for indexing and retrieval is shown in Fig. 1. For both indexing, retrieval and recognition we estimate the DI by James Stein plug-in estimation as follows. The pairwise DI, defined in (3), is estimated using the shrinkage estimator (4) of the multinomial probabilities, where the optimal shrinkage parameter (6) is selected to minimize the asymptotic expression for the MSE, represented as the sum of the square of the asymptotic bias and the asymptotic variance given in Proposition 2 in the Appendix. The nearest neighbor algorithm is applied to a symmetricized version of the DI similarity measure to index the video database. Indexing refers to organization of the video corpus according to the nearest neighbor graph over videos using the DI as a pairwise video distance. For retrieval, reverse nearest neighbors are used to find and rank the closest matches to a query. Precision is the fraction of retrieved instances that are relevant, while recall is the fraction of relevant instances that are retrieved. Once the DI optimal shrinkage parameter has been determined, the local DI is defined similarly to the DI except that, for a pair of videos  $X$  and  $Y$ , the videos are time shifted and windowed prior to computing the DI via (3). Specifically, let  $\tau_x \in [0, M_x - T]$ ,  $\tau_y \in [0, M_y - T]$  be the respective time shift parameters, where  $T \ll \min\{M_x, M_y\}$  is the sliding window width, and denoted by  $X_{\tau_x}^{M_x}, Y_{\tau_y}^{M_y}$  the time shifted videos. Then the local DI,  $DI(X_{\tau_x}^{M_x} \rightarrow Y_{\tau_y}^{M_y})$ , defines a surface over  $\tau_x, \tau_y$  and the summation indices in

(3) range over smaller sets of  $T$  time samples. We use the peaks of the local DI surface to detect and localize common activity in the pair of videos. As a quantitative measure, we will assign a p-value to the MI and DI. The p-value is defined as the critical threshold that would lead to the rejection of the null hypothesis [4]. The test statistic is computed as

$$T^{a,v} = DI(Y^v, X^a) = \max_{i,j} DI(Y_i^v, X_j^a), (i, j \in \mathbb{Z}^+) \quad (7)$$

where  $i, j$  is the time index in the video sequence. In this work, we utilize both of central limit theorem relying on Proposition 2 and bootstrap resampling to calculate p-values, where the Proposition 2 is presented in the appendix and the overall bootstrap based test procedure is:

- 1) Repeat the following procedure  $B(= 1000)$  times (with index  $b = 1, \dots, B$ ):
  - Generate resampled (with replacement) versions of the times series  $X^a, Y^v$ , denoted by  $X_b^a, Y_b^v$  respectively.
  - Compute the statistic  $t_b^{a,v} = DI(Y_b^v, X_b^a) = \max_{i,j} DI(Y_{i,b}^v, X_{j,b}^a), (i, j \in \mathbb{R})$
- 2) Construct an empirical CDF (cumulative distribution function) from these bootstrapped sample statistics, as  $F_T(t) = P(T \leq t) = \frac{1}{B} \sum_{b=1}^B I_{x>0}(x = t - t_b)$ , where  $I$  is an indicator random variable on its argument  $x$ .
- 3) Compute the true detection statistic (on the original time series)  $t_0 = DI(Y^v, X^a)$  and its corresponding p-value ( $p_0 = 1 - F_T(t_0)$ ) under the empirical null distribution  $F_T(t)$ .

This can be applied to each peak in Fig.4 to specify the p-value.

## V. EXPERIMENTAL RESULTS

In this section we provide results illustrating the potential of SODA for indexing/retrieval, activity recognition, and audio and video localization using public-domain human activity video databases. We first illustrate the DI's capability to detect and localize common activity in pairs of videos (Figs. 6, 5), pairs of audio and video sequences (Fig. 4, Table I) and quantify its activity recognition performance relative to HMM activity recognition methods (Table II). We then give quantitative results demonstrating that the proposed SODA indexing and retrieval method has improved precision/recall performance as compared to other methods including indexing/retrieval algorithms implemented with MI, Granger causality, Cross Media Indexing Space [35], SIFT-bag kernels [38] and SVM (Fig. 7, Table III).

**TRECVID Database used in experiments:** To illustrate and compare these methods we use the TRECVID 2010 corpus for our experiment. The activity-annotated video dataset contains video clips of human activities including: people walking; meeting with others; talking; entering and exiting shops; playing ballgames. A total of 6320 video sequences from 85 different events were used in the following experiments. Each video sequence contained 350 video frames on average. Whenever we report performance comparisons in the following experiments, half of the videos were randomly

selected for training and cross-validation and the remainder were used for testing.

**Feature Fusion:** For audio features, Perceptual Linear Prediction (PLP) is a technique of warping spectra to minimize the differences between speakers while preserving the important speech information [10]. RASTA is a separate technique that applies a band-pass filter to each frequency subband so as to smooth over short-term noise variations and to mitigate effects of static spectral coloration in the speech channel [11]. The output of RASTA-PLP audio feature extraction is a 39 by  $N$  feature matrix where  $N$  is determined by the length of audio signals and is selected to be 350 in our experiment. The visual features are obtained from a visual codebook using bag of words (BOW). The visual codebook is constructed using the k-means algorithm [24], which is used to quantize the SIFT features into codewords (with  $k$  ranging from 300 to 800 clusters). The codebook is estimated using a training set of videos in the database. In the implementation, we have 500 codewords for SIFT features due to its best recognition performance. Thus, for  $N$  frames, we have a cube for joint feature representation with size  $39 \times 500 \times N$ , where here  $N$  is 350. The joint probability density functions which define DI and local DI are estimated from multidimensional histograms computed by counting the number of observed instances in the frames occurring in each cube.

**Investigation of competing algorithms:** We compare the activity recognition performance of DI with that of a HMM proposed for video classification with integration of multi-modal features in [14]. A discrete HMM is characterized by  $\Lambda = (A, B, \Pi)$ , where  $A$  is the state transition probability matrix,  $B$  is the observation symbol probability matrix and  $\Pi$  is the initial state distribution. We first train  $\Lambda_i, i = 1, 2, \dots, C$ , where  $C$  is the number of classes and here  $C = 85$ . For each observation sequence  $O$ , we compute  $P(O|\Lambda_i)$  and the classification is based on the maximum likelihood of  $P(O|\Lambda)$ . In [14], by assuming that features are independent of each other, they train an HMM for the audio and visual modalities separately. The observed sequences of different features are applied into the corresponding HMM. The final observation probability is computed as

$$P(O|\Lambda_i) = P(O^a|\Lambda_i^a)P(O^v|\Lambda_i^v), \quad (8)$$

where  $\Lambda^a = (A^a, B^a, \Pi^a), \Lambda^v = (A^v, B^v, \Pi^v)$ .  $A^a$  is the state transition probability matrix for audio features and  $A^v$  is for visual features. Similar notations are used for  $\Lambda, B, \Pi$ . Specifically, for the GMM given 1039 training video sequences, we implement the HMM by estimating the emission probability of the distribution of audio or visual features with Gaussian mixture models (GMM). We then implement the Baum-Welch algorithm with 50 iterations to estimate the parameters of the GMM model governing frames in each activity class. For a test video, activity is detected and classified using maximum likelihood. In the more recent work of [26] non-parametric kernel density estimation (KDE) is used to estimate emission probability and the authors demonstrate improvement over parametric Gaussian mixture models for action recognition. We therefore also compare with HMM using KDE estimates of emission probability.

The indexing/retrieval performance of the DI will be compared to that of our implementations of three state-of-the-art approaches [18] [35] [38]. In [18] they investigate a meta-classification combination strategy using Support Vector Machine. Compared with a probability-based combination strategy like our work, the meta-classifiers learn the weights for different classifiers. Our SVM implementation is based on libsvm and we use C-SVM with a radial basis function kernel [5]. In [35] the semantic correlations among multimedia objects of different modalities are learned. Then the heterogeneous multimedia objects are analyzed in the form of multimedia document (MMD) and indexing is performed in the cross-media indexing space. In [38] the Kullback-Leibler divergence was used as a similarity measure between SIFT features for video event analysis. We also compare the DI measure to the standard Granger causality measure, implemented with Ledoit-Wolf covariance shrinkage [17] to control excessive MSE. Finally, to show the advantage of shrinkage estimation for stably estimating the DI, we compare to a version of DI that uses PCA instead of shrinkage. PCA can be interpreted as a form of regularization that uses hard thresholding instead of shrinkage.

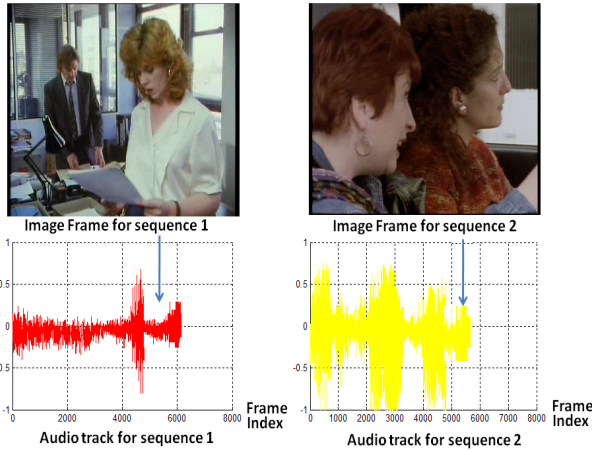


Fig. 3. Visual illustration of audio and video temporal localization, where SODA is able to localize the time of two people talking in two video sequences.

#### A. Multimodal activity recognition and localization

**Audio and video localization:** In multimodal video activity recognition, we need to first solve the correspondences between audio and video data. We demonstrate the application of SODA for audio and video localization. Namely, given the dataset with different speech signals and video signals, SODA is capable of determining the spatial and temporal correspondence between the speech signals and video signals by calculating the directed information between the pairs of speech signal and video signals. In the work by Fisher and Darrell [7] they proposed an approach based on maximum mutual information for cross-modal correspondence detection. They utilize the mutual information and regularization terms as follows:

$$J_1 = \hat{I}(Y^v, X^a) - \alpha^v (h^v)^T h^v - \alpha^a (h^a)^T h^a - \beta (h^v)^T \bar{R}_V^{-1} h^v, \quad (9)$$

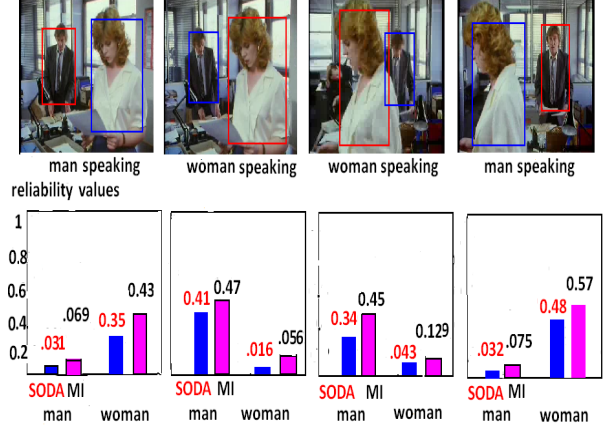


Fig. 4. Top row presents four frames from a video sequence with two speakers in TRECVID dataset. In the first and the fourth frames the man is speaking, while in the second and third frames the woman is speaking. The consistency measure using SODA shown in the bottom row for each frame correctly detects who is speaking and demonstrates the superiority over the MI-based method by Fisher et al [7], where the vertical axis represents the p-values. The corresponding p-values are annotated at the top of the histograms.

where the last term derives from the output energy constraint and  $\bar{R}_V^{-1}$  is the average autocorrelation function (taken over all images in the sequences),  $h^a$  and  $h^v$  are projection functions mapping the audio and video signals into low dimensional spaces,  $\alpha^a$ ,  $\alpha^v$  and  $\beta$  are scalar weighting terms. Different from [7], we define our localization criterion with SODA as:

$$J_2 = \widehat{DI}^\lambda(Y^v, X^a). \quad (10)$$

We evaluate the audio and video localization with 570 speech signals and the corresponding video signals for people talking. We compare the performance with mutual information described in [7] and show the results as a confusion matrix in Table I, where the left value in the elements of confusion matrix represents the accuracy of DI-based localization and the right represents the accuracy of MI-based localization. As shown in Table I, the temporal localization accuracy with DI consistently outperforms the MI-based localization, which demonstrates the competitive performance of SODA for temporal localization. We achieve more than 8% average precision compared to maximum mutual information as shown in Table I. To implement spatial localization, we first localize objects in the video frames using the method of object detection and mode learning described in [6]. The detection method uses strong low-level features based on histograms of oriented gradients (HOG) and efficient matching algorithms for deformable part-based models (pictorial structures). Here the localized objects are people. Using SODA, we calculated the directed information between the visual features in the bounding boxes and audio features. As shown in Fig. 4, the top row presents four frames from a video sequence with two speakers in the TRECVID dataset. In the first and the fourth frames the man is speaking, while in the second and third frames the woman is speaking. The measure using the p-value for SODA shown in the bottom row for each frame correctly detects who is speaking and demonstrates the superiority over



TABLE I

CONFUSION MATRIX FOR AUDIO-VIDEO LOCALIZATION FOR TRECVID 2010 DATASET WITH DI AND MI, WHERE THE COLUMNS INDICATE WHICH AUDIO SEQUENCE WAS USED WHILE THE ROWS INDICATE WHICH VIDEO SEQUENCE WAS USED. CLASSIFICATION IS PERFORMED USING A NEAREST NEIGHBOR CLASSIFIER.

SODA/MI	a1	a2	a3	a4	a5	a6	a7
v1	<b>0.76</b> /0.68	0.02/0.04	0.07/0.08	0.04/0.06	0.02/0.02	0.03/0.02	0.06/0.10
v2	0.05/0.07	<b>0.82</b> /0.73	0.03/0.06	0.02/0.05	0.07/0.04	0/0.02	0.01/0.04
v3	0.03/0.08	0.05/0.06	<b>0.78</b> /0.65	0.02/0.03	0.06/0.07	0.02/0.05	0.04/0.06
v4	0.07/0.09	0.02/0.03	0.04/0.05	<b>0.83</b> /0.71	0.02/0.05	0/0.04	0.02/0.03
v5	0.03/0.06	0.02/0.03	0.04/0.06	0.05/0.02	<b>0.77</b> /0.68	0.03/0.07	0.06/0.08
v6	0.03/0.05	0/0.02	0/0.03	0.01/0.02	0.03/0.04	<b>0.90</b> /0.79	0.03/0.05
v7	0.05/0.08	0.01/0.03	0.03/0.02	0/0.06	0.03/0.04	0.05/0.03	<b>0.83</b> /0.74

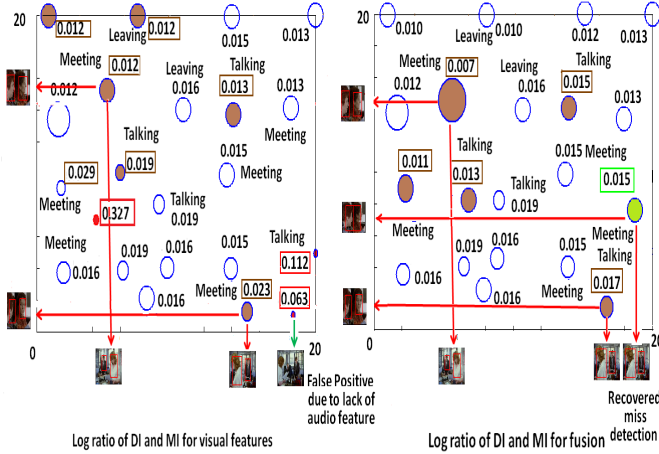


Fig. 5. Bubble graph of log ratio of peak values for local DI with only visual features (left) and with fusion (right) in  $\widehat{DI}(X_{\tau_x}^{M_x} \rightarrow Y_{\tau_y}^{M_y})$  between videos  $X$  and  $Y$ . Here the axes range over  $\tau_x$  and  $\tau_y$ , which represent time shift parameters of the respective video frames, and the sliding window width is  $T = 5$  frames. The size of the bubble is proportional to the log ratio of peak values of DI and MI. Each of the bubbles is annotated by a particular activity and its p-value. The improvement of p-values with fusion is shown by gray bounding boxes. The removal of false positives is highlighted by red bounding boxes on the left panel. The improvement of miss detections is highlighted by the green bounding box on the right panel.

the MI-based method by Fisher et al [7].

**Activity recognition and localization:** In Table II we compare the activity recognition performance of DI to that of the HMM implemented with GMM (first row of table) and KDE emission probability estimates. For purposes of comparison we evaluated performance on the same set of videos as in the TRECVID 2010 that were used in the experiments of [14] [26]. Video is digitized at 10 frames per second and at 240 by 180 pixels per frame and audio is sampled at 22.05 KHz and 16 bits per sample. The table indicates DI outperforms HMM in terms of activity recognition. This improvement might be attributed to the presence of model mismatch and bias in the HMM model as contrasted to the more robust behavior of the proposed model-free shrinkage DI approach.

We next show an anecdotal result suggesting that local DI is capable of identifying common activities in a pair of videos. Typically, the local DI with fusion of visual and audio features further improves the true positives and reduces the false alarm compared to the DI approach using only visual features. We selected two videos from the TRECVID 2010

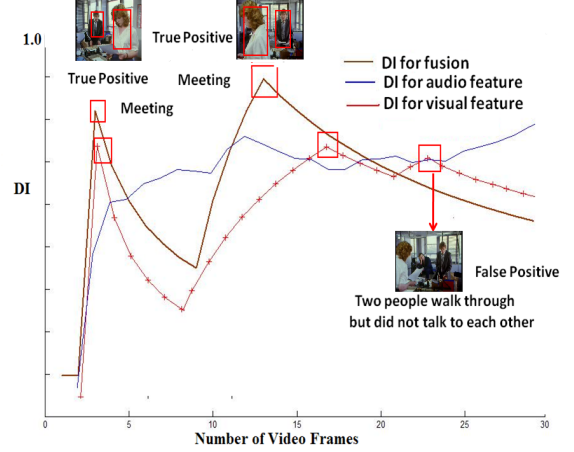


Fig. 6. Comparison of temporal trajectories and peak values of local directed information (DI) by fusing audio and visual features and local DI based on only audio and visual features versus time for two videos  $X, Y$ . The true positives for DI with fusion and false positives for DI with only visual features are highlighted. The fusion of DI provides better accuracy to detect and localize frames in  $Y$  with strong human interactions. Interactions between different people and trajectories corresponding to peak values in DI in the events are indicated in video by bounding boxes.

TABLE II

COMPARISONS OF AVERAGE PRECISION (AP) FOR SODA AND HIDDEN MARKOV MODEL (HMM) WITH GAUSSIAN MIXTURE MODEL (GMM) ( $n$  IS THE NUMBER OF COMPONENTS) AND KERNEL DENSITY ESTIMATION (KDE) FOR VIDEO RETRIEVAL IN TRECVID 2010 DATABASE.

	HMM(n=3)	HMM(n=6)	HMM(n=9)
AP	0.704	0.737	0.718
	KDE/HMM	MI	SODA
AP	0.769	0.693	<b>0.856</b>

dataset: "Two people enter, meet and talk to each other" in different locations, denoted as  $X$  and  $Y$ . The local DI from  $X$  to  $Y$  was rendered as a surface over  $\tau_x, \tau_y$ , as explained above, and the peaks on this surface were used to detect and localize common activities, i.e., activities in  $X$  that were predictive of activities in  $Y$ . The local MI is defined similarly to the local DI. The bubbles (dots) in Fig. 5 occur at the peaks of the log ratio of pairwise DI and MI and the size of each bubble is proportional to the magnitude of the log-ratio of the associated peak. The figure shows that the DI peaks occur at frames containing strong common activities and are higher than the MI at those locations. Moreover, as

shown in the Figure, by fusion we remove three false positives by incorporating the audio signals (red bounding boxes on the left panel). We strengthen most of the true positives by providing lower p-values with fusion (gray bounding boxes). In addition, with fusion, we recover one of the miss detections (green bounding boxes on the right panel). For instance, the peak labeled with reliability value 0.068 in the left figure disappears in the right figure by adding audio features, it can be mainly attributed to the fact that audio features have fewer false alarms and is very helpful for removing false positives. In the video and audio source, it corresponds to the event that two people walk through but they did not greet and talk to each other. Only using visual features is insufficient to discriminate between two people simple walking past each other vs exchanging a greeting. By adding audio signals, the false alarm is significantly reduced. The peak labeled with p-value 0.031 in the left figure is significantly reduced to 0.012 by the addition of audio features in the right figure.

As shown in Fig. 5, the DI detects that the human activity with strongest interactions is "Meeting", corresponding to the highest log ratio (largest bubbles). Lower peaks occurred at other times of common activity such as "Leaving," "Walking". The indicated p-values of DI peaks, computed using the central limit theorem for shrinkage DI, Prop. 2., suggest a high level of statistical significance of these peaks. Using corrected BH procedure with central limit theorem approximation to p-values [3] applied to pairs of video sequences shown in Fig. 5 for DI when  $\alpha$  is equal to 0.05 and 0.1, 8 and 15 peaks are detected. We increase the number of detections with bootstrap resampling [30] BH procedure with 1000 samples to 11 and 23. While for MI, 5 and 12 peaks are detected using corrected BH procedure with central limit theorem when  $\alpha$  is equal to 0.05 and 0.1. The number of peaks detected increased to 9 and 19 with bootstrap resampling BH procedure.

For further illustration, in Fig. 6 we plot the local DI with fusion of visual and audio features and local DI using only visual features as temporal trajectories. These trajectories can be interpreted as scan statistics for localizing common activity in the two videos. Specifically, the curves in Fig. 6 show slices of the local DI surfaces evaluated along the diagonal  $\tau_x = \tau_y$  (no relative time shift between the videos) for another pair of videos in the "people meet and talk" corpus. Fig. 6 shows that by fusion of two modalities we obtain a sharper DI curve (gray curve) as compared to the curve for local DI using only visual features (red) or only audio features (blue). Note that at the local peak value of DI annotated with the visual feature two people walk through but did not talk to each other the audio signal is flat while at the two other peak locations annotated with the feature "Meeting" it is varying. Therefore, the fusion of audio and video signal is capable of identifying the false alarm which cannot be resolved when only visual features are used.

Table III compares the average precision of the proposed SODA method and the SVM method for the TRECVID 2010 dataset. When there are events with low mutual interaction like "people marching," and a large number of associated features, the average precisions of the DI and SVM for retrieval are similar. However on average the proposed DI method results

in at least 10% better average accuracy. With fusion of audio-visual features, we obtain further improvement in recognition of events like "lecture" or "greeting", where the audio features provide important cues in discriminating between them. We also compare the average precision for activity recognition using SODA versus the number of codewords for SIFT features in Table IV. As shown in Table IV, the best recognition performance is achieved when the number of codewords used to construct SIFT features is 500. When the number of codewords is larger than 500, the performance deteriorates slightly which may be due to overfitting.

## B. Video retrieval

**Indexing and retrieval of video with misaligned modalities:** Next we turn to the application of SODA for indexing and retrieval of data with misaligned modalities. The implementation is as follows: (1) Compute marginal DI for the audio and video signals and detect peaks. (2) Segment the audios and videos according to peak locations to capture the beginning and ending points of interactive activity. (3) Compute pairwise DI on the aligned audio and video segments. (4) Repeat for all peak locations/segments. Fig.7 compares precision and recall performance of SODA to other indexing and retrieval methods. The experiments were implemented over the entire database of 6320 videos. As shown in Fig.7, the proposed DI method has the best overall performance exhibiting a significantly better area-under-the-curve (AUC) metric than the competing methods where AUC is computed by a non-parametric method based on constructing trapezoids under the curve as an approximation of area. Compared to the second best method using cross-media indexing [35], SODA provides more than 7% improvement measured using the area under the curve (AUC) of precision and recall curves. Among these methods only the Granger method provides directional measures of information flow. However, unlike DI the Granger causality measure is based on a strong Gaussian model assumption, which may account for its inferior performance. Fig.7 also shows that shrinkage regularized DI is better than PCA regularized DI. We also demonstrate the average running time for different algorithms for processing one video sequence using Matlab on a 3GHz PC in Table V, where SODA method takes about 6-7 seconds for processing one video sequence on average.

## VI. CONCLUSION

We proposed a novel framework for multimodal video indexing/retrieval and recognition based on SODA. The proposed approach estimates the joint PDFs of SIFT and RASTA-PLP and uses James-Stein shrinkage estimation strategies to control high variance. Since DI captures the directional information that videos and audios naturally possess, it demonstrates better performance as compared to other symmetric non-directional methods. We also demonstrate that the proposed SODA approach improves audio and video temporal and spatial localization and can be used to effectively index data with misaligned modalities.



TABLE III

COMPARISON OF AVERAGE PRECISION WITH SODA FOR FUSING AUDIO-VISUAL FEATURES FOR ACTIVITY RECOGNITION FROM TRECVID 2010 DATASET WHERE AVERAGE PRECISION IS MEASURED BY CORRECT RECOGNITION RATE COMPARED TO THE GROUND TRUTH.

Event Name	talking	lecture	greeting	fighting	greeting	people marching
SODA (visual)	<b>0.81</b>	<b>0.68</b>	<b>0.83</b>	<b>0.73</b>	<b>0.77</b>	<b>0.85</b>
SODA (fusion)	<b>0.86</b>	<b>0.75</b>	<b>0.89</b>	<b>0.75</b>	<b>0.86</b>	<b>0.88</b>
SVM (fusion)	0.74	0.73	0.67	0.62	0.71	0.79

TABLE IV

COMPARISON OF AVERAGE PRECISION WITH SODA FOR FUSING AUDIO-VISUAL FEATURES FOR ACTIVITY RECOGNITION FROM TRECVID 2010 DATASET VERSUS THE NUMBER FOR SIFT FEATURE CODEWORDS WHERE AVERAGE PRECISION IS AVERAGED OVER ALL THE ACTIVITIES.

Number of Codewords	300	400	500	600	700	800
SODA (fusion)	0.75	0.78	<b>0.82</b>	0.80	0.78	0.77

TABLE V

COMPARISON OF THE AVERAGE RUNNING TIME FOR DIFFERENT ALGORITHMS FOR PROCESSING ONE VIDEO SEQUENCE FROM TRECVID 2010 DATASET.

Algorithm	SVM	SIFT-bag Kernels	Granger Causality	Cross-Media Indexing	MI	SODA
Running Time (sec)	6.2	7.5	5.3	8.6	5.5	6.7

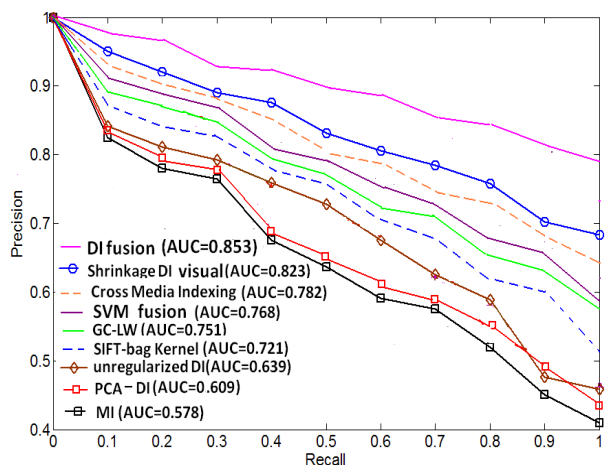


Fig. 7. Comparison of precision and recall curves for indexing using SODA with fusion and only with visual features, SVM with fusion, cross media indexing [35], mutual information (MI), Granger causality measure with LW shrinkage (GC-LW) [17], SIFT-bag Kernel [38], unregularized DI, DI with PCA regularization (PCA-DI) where PCA is implemented with a 20% residual energy threshold. Precision is defined as the fraction of relevant videos among those retrieved, while recall is the fraction of relevant videos retrieved among all relevant videos in the database.

#### ACKNOWLEDGEMENTS

This work was partially supported by a grant from the US Army Research Office, grant W911NF-09-1-0310. The authors would like to thank Dr Joseph P. Campbell at MIT Lincoln Research Laboratory for his suggestions on audio features.

#### REFERENCES

- [1] J. Allen. Maintaining knowledge about temporal intervals. In *Communications of the ACM*, 1983.
- [2] P. Amblard and O. Michel. On directed information theory and granger causality graphs. In *Journal of Computational Neuroscience*, volume 30, 2011.
- [3] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. In *Ann Stat*, volume 29, 2001.
- [4] P. Bickel and K. Doksum. *Mathematical statistics: Basic ideas and selected topics*. volume I, 2005.
- [5] C. Chang and C. Lin. Libsvm: A library for support vector machines. 2001.
- [6] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 32, 2010.
- [7] J. Fisher and T. Darrell. Speaker association with signal-level audiovisual fusion. In *IEEE Transactions on Multimedia*, 2004.
- [8] J. Germana. A transactional analysis of biobehavioral systems. *Integrative Physiological and Behavior Sciences*, 31, 1996.
- [9] J. Hausser and K. Strimmer. Entropy inference and the james-stein estimator, with application to nonlinear gene association networks. *Journal of Machine Learning Research*, 2009.
- [10] H. Hermansky. Perceptual linear predictive (plp) analysis of speech. In *J. Acoust. Soc. Am*, 1990.
- [11] H. Hermansky and N. Morgan. Rasta processing of speech. In *IEEE Trans. on Speech and Audio Proc*, 1994.
- [12] B. Hormler, D. Arsic, B. Schuller, and G. Rigoll. Boosting multimodal camera selection with semantic features. In *IEEE international conference on Multimedia and Expo*, 2009.
- [13] T. Hospedales, S. Gong, and T. Xiang. A Markov clustering topic model for mining behaviour in video. In *ICCV*, 2009.
- [14] J. Huang, Z. Liu, Y. Wang, Y. Chen, and E. K. Wong. Integration of multimodal features for video scene classification based on hmm. In *IEEE Signal Processing Society 1998 Workshop on Multimedia Signal Processing*, 1999.
- [15] Y. Ke, R. Sukthankar, and M. Hebert. Event Detection in Crowded Videos. In *International Conference on Computer Vision (ICCV)*. IEEE, 2007.
- [16] M. Krumin and S. Shoham. Multivariate autoregress modeling and granger causality analysis of multiple spike trains. *Computational Intelligence and Neuroscience*, 2010.
- [17] O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 2004.
- [18] W. Lin and A. Hauptmann. News video classification using svm-based multimodal classifiers and combination strategies. In *ACM Multimedia Conference*, 2002.
- [19] J. Liu and M. Shah. Learning human actions via information maximization. *IEEE Conference Computer Vision and Pattern Recognition*, 2008.
- [20] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal Computer Vision*, 2004.
- [21] J. Massey. Causality, feedback and directed information. *Symp Inf Theory and Its Applications (ISITA)*, 1990.
- [22] R. Messing, C. Pal, and H. Kautz. Activity recognition using the velocity histories of tracked keypoints. In *International Conference on Computer Vision (ICCV)*. IEEE, 2009.
- [23] R. Morris and D. Hogg. *Statistical Models of Object Interaction*. In *International Journal of Computer Vision*. Springer, 2000.
- [24] J. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *International Journal on Computer Vision (IJCV)*, 2008.
- [25] J. C. Niebles, C. W. Chen, and L. Fei-Fei. Modeling Temporal Structure of Decomposable Motion Segments for Activity Classification. In

- European Conference on Computer Vision (ECCV)*. IEEE, 2010.
- [26] M. Piccardi and O. Perez. Hidden markov models with kernel density estimation of emission probabilities and their use in activity recognition. *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.
- [27] C. Quinn, T. Coleman, N. Kiyavash, and N. Hatsopoulos. Estimating the directed information to infer causal relationships in ensemble neural spike train recordings. In *Journal of Computational Neuroscience*, volume 30, 2011.
- [28] A. Rao, A. Hero, D. J. States, and J. D. Engel. Motif discovery in tissue-specific regulatory sequences using directed information. *EURASIP Journal on Bioinformatics and Systems Biology*, 2007.
- [29] N. Rasiwasia, J. Pereira, E. Coviello, G. Doyle, G. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *ACM Proceedings of the 18th international conference on Multimedia*, 2010.
- [30] J. Romano, A. Shaikh, and M. Wolf. Control of the false discovery rate under dependence using the bootstrap and subsampling. In *TEST*, 2008.
- [31] C. Snoek and M. Worring. A state-of-the-art review on multimodal video indexing. 2002.
- [32] C. Snoek and M. Worring. Multimedia event-based video indexing using time intervals. In *IEEE Transactions on Multimedia*, volume 7, 2005.
- [33] Z. Sun and A. Hoogs. Image comparison by compound disjoint information. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.
- [34] X. Wang, K. Tieu, and E. Grimson. Learning semantic scene models by trajectory analysis. In *ECCV*, 2006.
- [35] Y. Yang, F. Wu, D. Xu, Y. Zhuang, and L. Chia. Cross-media retrieval using query dependent search methods. In *Pattern Recognition*, volume 43, 2010.
- [36] J. H. Z. Liu and Y. Wang. Classification of tv programs based on audio information using hidden markov models. In *IEEE Signal Processing Society 1998 Workshop on Multimedia Signal Processing*, 1998.
- [37] H. Zhong, J. Shi, and M. Visontai. Detecting unusual activity in video. In *CVPR*, 2004.
- [38] X. Zhou, X. Zhuang, S. Yan, S. Chang, M. Johnson, and T. Huang. Sift-bag kernel for video event analysis. *ACM International Conference on Multimedia*, 2008.
- [39] Y. Zhuang, Y. Yang, and F. Wu. Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval. In *IEEE Transactions on Multimedia*, volume 10, 2008.

## APPENDIX I THE BIAS AND VARIANCE

**Proposition 1:** The bias of the directed information estimator with James-Stein plug-in estimator can be represented as:

$$\text{Bias}(\widehat{DI}_\theta^\lambda) = C_1 + C_2 \frac{1}{n} + O\left(\frac{1}{n^2}\right), \quad (11)$$

where  $C_1 = C_{b1} - C_{b2} + C_{b3}$ ,  $C_2 = C_{b4} - C_{b5} + C_{b6}$ .

$$\begin{aligned} C_{b1} &= \sum_{m=1}^M \left[ -\frac{\theta_{x,y}(k,l)}{\sum_{k=1}^{p^m} \sum_{l=1}^{p^m} \theta_{x,y}(k,l)} \right. \\ &\log_2 \left( \frac{\theta_{x,y}(k,l)}{\sum_{k=1}^{p^m} \sum_{l=1}^{p^m} \theta_{x,y}(k,l)} \right) S_{k,l} + \\ &\frac{1}{\log 2} (-1 + S_{k,l}) \ln(1 - S_{k,l}) \\ &\left. \frac{\theta_{x,y}(k,l)}{\sum_{k=1}^{p^m} \sum_{l=1}^{p^m} \theta_{x,y}(k,l)} \right], \quad (12) \\ S_{k,l} &= \lambda \left( 1 - \frac{\sum_{k=1}^{p^m} \sum_{l=1}^{p^m} \theta_{x,y}(k,l)}{p^m \theta_{x,y}(k,l)} \right), \end{aligned}$$

$$\begin{aligned} C_{b2} &= \sum_{m=1}^M \left[ -\frac{\theta_{x,y}(k,l)}{\sum_{k=1}^{p^m} \sum_{l=1}^{p^{(m-1)}} \theta_{x,y}(k,l)} \right. \\ &\log_2 \left( \frac{\theta_{x,y}(k,l)}{\sum_{k=1}^{p^m} \sum_{l=1}^{p^{(m-1)}} \theta_{x,y}(k,l)} \right) V_{k,l} + \\ &\left. \frac{1}{\log 2} (-1 + V_{k,l}) \ln(1 - V_{k,l}) \frac{\theta_{x,y}(k,l)}{\sum_{k=1}^{p^m} \sum_{l=1}^{p^{(m-1)}} \theta_{x,y}(k,l)} \right], \end{aligned}$$

$$V_{k,l} = \lambda \left( 1 - \frac{\sum_{k=1}^{p^m} \sum_{l=1}^{p^{(m-1)}} \theta_{x,y}(k,l)}{p^{(2m-1)} \theta_{x,y}(k,l)} \right),$$

$$\begin{aligned} C_{b3} &= \left[ -\frac{\theta_y(l)}{\sum_{l=1}^p \theta_y(l)} \log_2 \left( \frac{\theta_y(l)}{\sum_{l=1}^p \theta_y(l)} \right) W_{k,l} + \right. \\ &\left. \frac{1}{\log 2} (-1 + W_{k,l}) \ln(1 - W_{k,l}) \frac{\theta_y(l)}{\sum_{l=1}^p \theta_y(l)} \right], \quad (13) \end{aligned}$$

$$W_{k,l} = \lambda \left( 1 - \frac{\sum_{k=1}^p \theta_y(l)}{p \theta_y(l)} \right),$$

$$C_{b4} = \sum_{m=1}^M \frac{1}{2 \log 2} \frac{1}{(1 - S_{k,l})} \left( \frac{\theta_{x,y}(k,l)}{\sum_{k=1}^{p^m} \sum_{l=1}^{p^m} \theta_{x,y}(k,l)} - 1 \right),$$

$$C_{b5} = \sum_{m=1}^M \frac{1}{2 \log 2} \frac{1}{(1 - V_{k,l})} \left( \frac{\theta_{x,y}(k,l)}{\sum_{k=1}^{p^m} \sum_{l=1}^{p^{(m-1)}} \theta_{x,y}(k,l)} - 1 \right),$$

$$C_{b6} = \frac{1}{2 \log 2} \frac{1}{(1 - W_{k,l})} \left( \frac{\theta_y(l)}{\sum_{l=1}^p \theta_y(l)} - 1 \right).$$

Remark: In the above equations,  $p^{(m-1)}$  comes from the dimension of the PDF for  $Y^{(m-1)}$  and  $p^{(2m-1)}$  comes from the dimension of the joint PDF for  $X^{(m)}$  and  $Y^{(m-1)}$ .

**Proposition 2:** The directed information (DI) with plug-in JS shrinkage estimator is asymptotically Gaussian, where the asymptotical mean  $\mu_0 = \sum_{m=1}^M (A \log A - B \log B) + C \log C$ , where  $A = \frac{\lambda}{p^{2m}} + (1 - \lambda) \frac{\sum_{k=1}^{p^m} \theta_{x,y}(k,l)}{\sum_{k=1}^{p^m} \sum_{l=1}^{p^m} \theta_{x,y}(k,l)}$ ,  $B = \frac{\lambda}{p^{(2m-1)}} + (1 - \lambda) \frac{\sum_{k=1}^{p^m} \theta_{x,y}(k,l)}{\sum_{k=1}^{p^m} \sum_{l=1}^{p^{(m-1)}} \theta_{x,y}(k,l)}$ ,  $C = \lambda/p + (1 - \lambda) \frac{\theta_y(l)}{\sum_{l=1}^p \theta_y(l)}$ . The asymptotic variance is given by  $T_2 \Sigma_2 T_2' \frac{1}{n}$ , where the first  $p^{2M}/2$  diagonal elements in  $\Sigma_2$  are denoted by  $\theta_x(k)(1 - \theta_x(k))$ , the last  $p^{2M}/2$  diagonal elements in  $\Sigma_2$  are denoted by  $\theta_y(l)(1 - \theta_y(l))$ . The non-diagonal elements in  $(k, l)$  in the first  $p^{2M}/2$  rows and the first  $p^{2M}/2$  columns in  $\Sigma_2$  are denoted by  $-\theta_x(k)\theta_x(l)$ . The non-diagonal element in the last  $p^{2M}/2$  rows and the last  $p^{2M}/2$  columns is denoted by  $-\theta_y(l)\theta_y(k)$  and the rest of them are denoted by  $-\theta_x(k)\theta_y(l)$ .  $T_2 = (\frac{\partial \widehat{DI}_\theta^\lambda}{\partial \theta_x(k)}, \frac{\partial \widehat{DI}_\theta^\lambda}{\partial \theta_y(l)})$  is a  $1 \times p$  vector. Therefore,

$$\begin{aligned} \frac{\partial \widehat{DI}_\theta^\lambda}{\partial \theta_x(k)} &= \sum_{m=1}^M \left[ (\log A + 1) \frac{\partial A}{\partial \theta_x(k)} + (\log B + 1) \frac{\partial B}{\partial \theta_x(k)} \right], \\ \frac{\partial \widehat{DI}_\theta^\lambda}{\partial \theta_y(l)} &= \sum_{m=1}^M \left[ (\log A + 1) \frac{\partial A}{\partial \theta_y(l)} + (\log B + 1) \frac{\partial B}{\partial \theta_y(l)} \right] + \\ &(\log C + 1) \frac{\partial C}{\partial \theta_y(l)} \quad (14) \end{aligned}$$

where if  $k = k_0$  or  $l = l_0$ ,

$$\left( \frac{\partial A}{\partial \theta_x(k_0)}, \frac{\partial A}{\partial \theta_y(l_0)} \right) = (1 - \lambda) \frac{\sum_{k=1}^{p^m} \sum_{l=1}^{p^m} \theta_{x,y}(k, l) - 1}{\left( \sum_{k=1}^{p^m} \sum_{l=1}^{p^m} \theta_{x,y}(k, l) \right)^2} \left( \frac{\partial \theta_{x,y}(k_0, l)}{\partial \theta_x(k_0)}, \frac{\partial \theta_{x,y}(k, l_0)}{\partial \theta_y(l_0)} \right), \quad (15)$$

$$\left( \frac{\partial B}{\partial \theta_x(k_0)}, \frac{\partial B}{\partial \theta_y(l_0)} \right) = (1 - \lambda) \frac{\sum_{k=1}^{p^m} \sum_{l=1}^{p^{(m-1)}} \theta_{x,y}(k, l) - 1}{\left( \sum_{k=1}^{p^m} \sum_{l=1}^{p^{(m-1)}} \theta_{x,y}(k, l) \right)^2} \left( \frac{\partial \theta_{x,y}(k_0, l)}{\partial \theta_x(k_0)}, \frac{\partial \theta_{x,y}(k, l_0)}{\partial \theta_y(l_0)} \right), \quad (16)$$

$$\frac{\partial C}{\partial \theta_y(l_0)} = (1 - \lambda) \frac{\sum_{l=1}^p \theta_y(l) - 1}{\left( \sum_{l=1}^p \theta_y(l) \right)^2} \quad (17)$$

otherwise

$$\left( \frac{\partial A}{\partial \theta_x(k_0)}, \frac{\partial A}{\partial \theta_y(l_0)} \right) = (1 - \lambda) \frac{-1}{\left( \sum_{k=1}^{p^m} \sum_{l=1}^{p^m} \theta_{x,y}(k, l) \right)^2} \left( \frac{\partial \theta_{x,y}(k_0, l)}{\partial \theta_x(k_0)}, \frac{\partial \theta_{x,y}(k, l_0)}{\partial \theta_y(l_0)} \right). \left( \frac{\partial B}{\partial \theta_x(k_0)}, \frac{\partial B}{\partial \theta_y(l_0)} \right) = (1 - \lambda) \frac{-1}{\left( \sum_{k=1}^{p^m} \sum_{l=1}^{p^{(m-1)}} \theta_{x,y}(k, l) \right)^2} \left( \frac{\partial \theta_{x,y}(k_0, l)}{\partial \theta_x(k_0)}, \frac{\partial \theta_{x,y}(k, l_0)}{\partial \theta_y(l_0)} \right), \left( \frac{\partial C}{\partial \theta_y(l_0)} \right) = (1 - \lambda) \frac{-1}{\left( \sum_{l=1}^p \theta_y(l) \right)^2} \quad (18)$$

## APPENDIX II

### DERIVATION FOR THE BIAS AND VARIANCE

In order to derive the bias and variance of regularized directed information shown in Proposition 1 and 2, we first compute the bias of shrinkage entropy estimator. The bias of the entropy estimator for features in a single frame with plug-in estimator can be represented as:

$$\text{Bias}(\hat{H}_\theta^\lambda) = \sum_{k=1}^p [-\theta_k \log_2(\theta_k) U_k + \frac{1}{\log 2} (-(1 + U_k) \ln(1 - U_k)) \theta_k] + \sum_{k=1}^p \frac{1}{2 \log 2} \frac{1}{(1 - U_k)} (\theta_k - 1) \frac{1}{n} + O\left(\frac{1}{n^2}\right), \quad (19)$$

where  $U_k = (1 - \frac{1}{p\theta_k})\lambda$ . The entropy estimator is asymptotically Gaussian. The asymptotic mean can be represented as  $(-\sum_{k=1}^p (\lambda/p + (1 - \lambda)\theta_k) \log(\lambda/p + (1 - \lambda)\theta_k))$  and the asymptotic variance of the entropy estimator with plug-in estimator  $\text{Var}(\hat{H}_\theta^\lambda)$  can be represented as:  $(1 - \lambda)^2 T_1 \Sigma T_1^T \frac{1}{n}$ , where  $T_1 = [\log(\lambda/p + (1 - \lambda)\theta_1) + 1, \dots, \log(\lambda/p + (1 - \lambda)\theta_p) + 1]$ . The  $k$ th diagonal element in the  $p \times p$  covariance matrix  $\Sigma$  is  $\theta_k(1 - \theta_k)$  and the  $k$ th row and  $j$ th column non-diagonal elements in  $\Sigma$  is  $-\theta_k \theta_j$ . Since the ML estimator of parameter  $\theta$  in the multinomial distribution converges to multivariate Gaussian distribution for large  $n$ , using delta method, asymptotic expressions for variance can be established. We briefly state the main idea behind delta method here: Let

a consistent asymptotic Gaussian estimator  $B$  converges in probability to its true value  $\beta$ :  $\sqrt{n}(B - \beta) \rightarrow N(0, \Sigma)$ . Then if  $H$  is a differentiable function, the delta method says that  $\sqrt{n}(H(B) - H(\beta)) \rightarrow N(0, \nabla(H(\beta))^T \Sigma \nabla(H(\beta)))$ . Furthermore, in the entropy estimation context, it is easy to show that

$$\nabla H = \left[ \frac{\partial \hat{H}_\theta^\lambda}{\partial \theta_1}, \dots, \frac{\partial \hat{H}_\theta^\lambda}{\partial \theta_p}, \frac{\partial \hat{H}_\theta^\lambda}{\partial \theta_k} \right] = (1 - \lambda) T_1,$$

**Remark:** With increasing  $\lambda$ , the variance decreases for fixed  $n$ . For fixed  $n$ , if the shrinkage coefficient  $\lambda$  is increasing, the square of the bias is increasing and the variance is decreasing. Therefore, the optimal choice of  $\lambda$  provides the optimal trade-off between the bias and variance by minimizing the mean square error which is the sum of the square of the bias and the variance. In the extreme case, when  $\lambda = 1$ , the shrinkage estimator boils down to maximum likelihood estimator. In this case, the bias is 0 and the variance is maximized. We now use the expressions for the bias and variance of entropy estimator to find the bias and the variance of estimated directed information. Based on the formulation of directed information shown in the equation, the directed information can be further simplified as:

$$\widehat{DI}_\theta^\lambda(X^M \rightarrow Y^M) = \sum_{m=1}^M [\hat{H}^\lambda(X_1, \dots, X_m, Y_1, \dots, Y_m) - \hat{H}^\lambda(X_1, \dots, X_m, Y_1, \dots, Y_{m-1})] + \hat{H}^\lambda(Y_M). \quad (20)$$

Let us assume the joint distribution of the two sequences with the length  $M$  (or  $M$  states) of  $X$  and  $Y$  is multinomial distribution  $f(X_1, \dots, X_M, Y_1, \dots, Y_M)$  with the frequency parameters  $\theta_{x,y}(k, l)$ . The marginal distribution  $f(X_1, \dots, X_m, Y_1, \dots, Y_m)$  for a segment of the two sequences with length  $m$  is also multinomial. Therefore, we can apply the similar approach as we show for entropy estimation to compute the bias and variance for the estimator of directed information.

#### A. Proof for Proposition 1

**Proof:** We use the Taylor expansion of the entropy function  $\hat{H}(\theta_1^\lambda, \dots, \theta_p^\lambda)$  around the true value of the entropy for  $\theta_k$ ,  $k = 1, \dots, p$  as follows:

$$\hat{H}(\theta_1^\lambda, \dots, \theta_p^\lambda) = H(\theta_1, \dots, \theta_p) + \sum_{k=1}^p \frac{\partial H(\theta_1, \dots, \theta_p)}{\partial \theta_k} (\theta_k^\lambda - \theta_k) + \sum_{k=1}^p \sum_{j=1}^p \frac{1}{2} \frac{\partial^2 H(\theta_1, \dots, \theta_p)}{\partial \theta_k \partial \theta_j} (\hat{\theta}_k^\lambda - \theta_k)(\hat{\theta}_j^\lambda - \theta_j) + \dots \quad (21)$$

where the coefficients are as follows:

$$\frac{\partial H(\theta_1, \dots, \theta_p)}{\partial \theta_k} = -\log_2 \theta_k - \frac{1}{\log 2} \frac{\partial^2 H(\theta_1, \dots, \theta_p)}{\partial \theta_k \theta_j} = -\frac{1}{\theta_k \log 2} \delta_{j,k}, \dots \frac{\partial^n H(\theta_1, \dots, \theta_p)}{\partial \theta_k \dots \theta_l} = \frac{(-1)^{n-1} (n-2)!}{\theta_k^{n-1} \log 2} \delta_{i, \dots, l}, \quad (22)$$

where  $\delta_{j,k} = 1$  when  $j = k$ , and  $\delta_{j,k} = 0$  when  $j \neq k$ . Therefore, the bias of the entropy can be represented as

$$\begin{aligned} \text{Bias}(\hat{H}^\lambda) &= E(\hat{H}^\lambda) - H = \\ &= \sum_{k=1}^p \left(-\log_2 \theta_k - \frac{1}{\log 2}\right) E[(\theta_k^\lambda - \theta_k)] + \\ &= \sum_{k=1}^p \frac{1}{\theta_k \log 2} E[(\theta_k^\lambda - \theta_k)^2] + \dots + \\ &= \sum_{k=1}^p \frac{(-1)^{n-1}}{(n-1)n\theta_k^{n-1} \log 2} E[(\theta_k^\lambda - \theta_k)^n] \end{aligned} \quad (23)$$

Meanwhile, we have  $\theta_k^\lambda - \theta_k = \lambda(1/p - \theta_k) + (1-\lambda)(\hat{\theta}_k^{ML} - \theta_k)$ . It can be seen that  $E[(\hat{\theta}_k^{ML} - \theta_k)^m]$  satisfies the following recursive formula where  $\mu_n = E[(X_k - \theta_k N)^n] = E[(\hat{\theta}_k^{ML} N - \theta_k N)^n]$ :  $\mu_{n+1} = \theta_k(1-\theta_k)(Nn\mu_{n-1} + \frac{\partial \mu_n}{\partial \theta_k})$ . By substituting the first few terms with  $\mu_0 = 1, \mu_1 = 0, \mu_2 = \theta_k(1-\theta_k)N$  into the recursion formula, the  $n$ th order central moment of  $X_k$  can be seen to be a polynomial in terms of  $N$ , when  $n$  is a even number, the order of the polynomial at most  $n/2$ , namely,  $(O(N^{n/2}))$ , and at most  $(O(N^{(n-1)/2}))$  for even  $n$ . Since  $\hat{\theta}_k^{ML} = X_k/N$ , the  $n$ th order central moment of  $\hat{\theta}_k^{ML}$  is a polynomial in terms of  $1/N$  of the order at most  $n/2$ , namely,  $O(1/N^{n/2})$ , when  $n$  is an even number and at most  $(n+1)/2$ , namely,  $O(1/N^{(n+1)/2})$ , when  $n$  is an odd number. We have the first few terms as follows:

$$\begin{aligned} E[(\hat{\theta}_k^{ML} - \theta_k)^2] &= \frac{1}{N}(\theta_k(1-\theta_k)), \\ E[(\hat{\theta}_k^{ML} - \theta_k)^3] &= \frac{1}{N^2}(\theta_k(1-\theta_k)(1-2\theta_k)), \\ E[(\hat{\theta}_k^{ML} - \theta_k)^4] &= \frac{1}{N^3}(\theta_k(1-\theta_k)(1+3\theta_k(1-\theta_k)(N-2))\}24 \\ E(\hat{\theta}_k^\lambda - \theta_k) &= \lambda(1/p - \theta_k) + (1-\lambda)E(\hat{\theta}_k^{ML} - \theta_k) = \lambda(1/p - \theta_k) \\ E[(\hat{\theta}_k^\lambda - \theta_k)^n] &= (\lambda(1/p - \theta_k))^n + \\ &= n(\lambda(1/p - \theta_k))^{n-1} E(\hat{\theta}_k^{ML} - \theta_k) + \\ &= \sum_{i=2}^n C_n^i (\lambda(1/p - \theta_k))^{n-i} E[(\hat{\theta}_k^{ML} - \theta_k)^i] = \\ &= (\lambda(1/p - \theta_k))^n + \frac{n(n-1)}{2} (\lambda(1/p - \theta_k))^{n-2} \\ &= E[(\hat{\theta}_k^{ML} - \theta_k)^2] + \sum_{i=3}^n C_n^i (\lambda(1/p - \theta_k))^{n-i} E[(\hat{\theta}_k^{ML} - \theta_k)^i] \end{aligned} \quad (25)$$

Since when  $i \geq 3$ ,  $\sum_{i=3}^n C_n^i (\lambda(1/p - \theta_k))^{n-i} E[(\hat{\theta}_k^{ML} - \theta_k)^i]$  are at least  $O(\frac{1}{N^2})$  and can be ignored, only the first two terms are considered. Considering the first term in the above and combining the equation (23), we obtain

$$\begin{aligned} &= \sum_{k=1}^p \frac{(-1)^{n-1}}{(n-1)n\theta_k^{n-1} \log 2} (\lambda(1/p - \theta_k))^n = \\ &= \sum_{k=1}^p \frac{(-1)^{n-1} \theta_k}{(n-1)n \log 2} (\lambda(1/(p\theta_k) - 1))^n \end{aligned} \quad (26)$$

A sufficient condition for the convergence of the right side of the equation (26) is that  $|\lambda(1/(p\theta_k) - 1)| < 1$ , which establishes a sufficient condition for asymptotical unbiasedness.

For computation of the bias, observe that,

$$\begin{aligned} &= \lim_{n \rightarrow \infty} \sum_{n=2}^{\infty} \sum_{k=1}^p \frac{(-1)^{n-1} \theta_k}{(n-1)n \log 2} (\lambda(1/(p\theta_k) - 1))^n = \\ &= \sum_{k=1}^p \frac{1}{\log 2} (-U_k - (1-U_k) \ln(1-U_k)) \theta_k, \end{aligned} \quad (27)$$

where  $U_k = (1 - \frac{1}{p\theta_k})\lambda$ . The equality shown in the equation (27) can be shown as follows:

$$\begin{aligned} &= \sum_{n=2}^{\infty} \sum_{k=1}^p \frac{(-1)^{n-1} \theta_k}{(n-1)n \log 2} (\lambda(1/(p\theta_k) - 1))^n = \\ &= \sum_{n=2}^{\infty} \sum_{k=1}^p \frac{-\theta_k}{(n-1)n \log 2} (\lambda(1 - 1/(p\theta_k)))^n = \\ &= \sum_{n=2}^{\infty} \sum_{k=1}^p \left[ \frac{-\theta_k}{(n-1) \log 2} (\lambda(1 - 1/(p\theta_k)))^n - \right. \\ &= \left. \frac{-\theta_k}{n \log 2} (\lambda(1 - 1/(p\theta_k)))^n \right] \end{aligned} \quad (28)$$

Let  $U_k = \lambda(1 - 1/(p\theta_k))$ , first consider

$$\begin{aligned} &= \sum_{n=2}^{\infty} \frac{U_k^n}{n} = \sum_{n=1}^{\infty} \frac{U_k^n}{n} - U_k = \\ &= \int_0^{U_k} \left( \sum_{n=1}^{\infty} \frac{U_k^n}{n} \right)' dU_k - U_k = \int_0^{U_k} \sum_{n=1}^{\infty} U_k^{n-1} dU_k - U_k \\ &= \int_0^{U_k} \frac{1}{1-U_k} dU_k - U_k = -\ln(1-U_k) - U_k \end{aligned} \quad (29)$$

$$\begin{aligned} &= \sum_{n=2}^{\infty} \frac{U_k^n}{n-1} = U_k \sum_{n=2}^{\infty} \frac{U_k^{n-1}}{n-1} = \\ &= U_k \sum_{n=1}^{\infty} \frac{U_k^n}{n} = -U_k \ln(1-U_k). \end{aligned} \quad (30)$$

Therefore, the equation (27) can be established.

$$\begin{aligned} &= \sum_{n=2}^{\infty} \sum_{k=1}^p \frac{(-1)^{n-1}}{(n-1)n\theta_k^{n-1} \log 2} \\ &= \frac{n(n-1)}{2} (\lambda(1/p - \theta_k))^{n-2} E[(\hat{\theta}_k^{ML} - \theta_k)^2] = \\ &= \sum_{k=1}^p \frac{1}{2 \log 2} \frac{1}{(1-U_k)} (\theta_k - 1) \frac{1}{N} \end{aligned} \quad (31)$$

Recall in the formulation of bias in (23), we have:

$$\sum_{k=1}^p \left(-\log_2 \theta_k - \frac{1}{\log 2}\right) E[(\hat{\theta}_k^\lambda - \theta_k)] = - \sum_{k=1}^p \theta_k \log_2(\theta_k) U_k \quad (32)$$

Combining the equations (27), (31) and (32), the bias shown in Proposition 1 can be established.

Let  $f(U_k) = -U_k - (1-U_k) \ln(1-U_k)$ . Since  $\frac{\partial f(U_k)}{\partial U_k} = \ln(1-U_k)$  and  $U_k < \min\{1, (1 - \frac{1}{p\theta_k})\}$ , when  $0 \leq U_k < 1 - \frac{1}{p\theta_k}$ ,  $f(U_k)$  is monotonically decreasing. Therefore when  $X \in (-1, 1)$ ,  $f(U_k) \in [f(U_k)_{min}, 0]$ , where  $f(U_k)_{min} = \sum_{k=1}^p \frac{1}{\log 2} (-U_{max} - (1-U_{max}) \ln(1-U_{max})) \theta_k$

## B. Proof for Variance of regularized DI

Since the directed information can be represented as:

$$\widehat{DI}_\theta^\lambda(X^M \rightarrow Y^M) = \sum_{m=1}^M [\widehat{H}^\lambda(X_1, \dots, X_m, Y_1, \dots, Y_m) - \widehat{H}^\lambda(X_1, \dots, X_m, Y_1, \dots, Y_{m-1})] + \widehat{H}^\lambda(Y_M). \quad (33)$$

According to the delta method, we only need to compute  $(\frac{\partial \widehat{DI}_\theta^\lambda}{\partial \theta_x(k)}, \frac{\partial \widehat{DI}_\theta^\lambda}{\partial \theta_y(l)})$ . We need to find  $\sum_{m=1}^M \frac{\partial \widehat{H}^\lambda(X_1, \dots, X_m, Y_1, \dots, Y_m)}{\partial \theta_x(k)}$  and  $\sum_{m=1}^M \frac{\partial \widehat{H}^\lambda(X_1, \dots, X_m, Y_1, \dots, Y_{m-1})}{\partial \theta_x(k)} + \frac{\partial \widehat{H}^\lambda(Y_M)}{\partial \theta_x(k)}$ ,  $\sum_{m=1}^M \frac{\partial \widehat{H}^\lambda(X_1, \dots, X_m, Y_1, \dots, Y_m)}{\partial \theta_y(l)}$  and  $\sum_{m=1}^M \frac{\partial \widehat{H}^\lambda(X_1, \dots, X_m, Y_1, \dots, Y_{m-1})}{\partial \theta_y(l)} + \frac{\partial \widehat{H}^\lambda(Y_M)}{\partial \theta_y(l)}$ . Here we provide the derivation for computing  $\frac{\partial \widehat{DI}_\theta^\lambda}{\partial \theta_x(k)}$ , the process of computing  $\frac{\partial \widehat{DI}_\theta^\lambda}{\partial \theta_y(l)}$  can be shown similarly. Considering  $P(X_1, \dots, X_m, Y_1, \dots, Y_m)$  is the multinomial distribution with frequency parameter  $\theta_a = \frac{\theta_{x,y}(k,l)}{\sum_{k=1}^{p^m} \sum_{l=1}^{p^m} \theta_{x,y}(k,l)}$  and the dimension  $p^m$ ,

$$\frac{\partial \widehat{H}^\lambda(X_1, \dots, X_m, Y_1, \dots, Y_m)}{\partial \theta_x(k)} = \frac{\partial \sum_{p=1}^m [\lambda/p^{2m} + (1-\lambda)\theta_a] \log(\lambda/p^{2m} + (1-\lambda)\theta_a)}{\partial \theta_x(k)} \quad (34)$$

where  $k = 1, \dots, m/M$ . According to the chain rule, we obtain  $\frac{\partial \widehat{H}^\lambda(X_1, \dots, X_m, Y_1, \dots, Y_m)}{\partial \theta_x(k)} = (\log A + 1) \frac{\partial A}{\partial \theta_x(k)}$ , where  $A = \frac{\lambda}{p^m} + (1-\lambda) \frac{\theta_{x,y}(k,l)}{\sum_{k=1}^{p^m} \sum_{l=1}^{p^m} \theta_{x,y}(k,l)}$ . Then we only need to compute  $\frac{\partial A}{\partial \theta_x(k)}$ . It has been noted that if  $k \neq k_0$ ,  $\frac{\partial \theta_{x,y}(k,l)}{\partial \theta_x(k_0)} = 0$ . Therefore, according to the chain rule, we can compute: for  $k = k_0$ ,

$$\frac{\partial A}{\partial \theta_x(k_0)} = (1-\lambda) \frac{\sum_{k=1}^{p^m} \sum_{l=1}^{p^m} \theta_{x,y}(k,l) - 1}{(\sum_{k=1}^{p^m} \sum_{l=1}^{p^m} \theta_{x,y}(k,l))^2} \frac{\partial \theta_{x,y}(k_0,l)}{\partial \theta_x(k_0)}, \quad (35)$$

For  $k \neq k_0$ ,

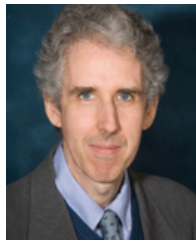
$$\frac{\partial A}{\partial \theta_x(k_0)} = (1-\lambda) \frac{-1}{(\sum_{k=1}^{p^m} \sum_{l=1}^{p^m} \theta_{x,y}(k,l))^2} \frac{\partial \theta_{x,y}(k_0,l)}{\partial \theta_x(k_0)}. \quad (36)$$

The other terms can be derived similarly.



**Xu Chen** Xu Chen received the B.S. from Shanghai Jiao Tong University (SJTU) in, Shanghai, China (2006) and the PhD from University of Illinois (2010) both in Electrical Engineering. He has also been research intern with Ecole Polytechnique Federale de Lausanne (EPFL) in Lausanne, Switzerland and Kodak Research Lab in Eastman Kodak Company in Rochester, New York, USA in 2008 and 2009 respectively. Since 2010 he has been research fellow with University of Michigan, Ann Arbor in Department of Electrical Engineering and

Computer Science. He coauthored the book chapter "Motion trajectory-based video retrieval, classification, and summarization," Video Search and Mining, Studies in Computational Intelligence Series, Springer-Verlag in 2010. Xu Chen's main research interests are in image and video processing, machine learning, computer vision and statistical signal processing.



**Alfred Hero** Alfred O. Hero III received the B.S. (summa cum laude) from Boston University (1980) and the Ph.D from Princeton University (1984), both in Electrical Engineering. Since 1984 he has been with the University of Michigan, Ann Arbor, where he is the R. Jamison and Betty Williams Professor of Engineering. His primary appointment is in the Department of Electrical Engineering and Computer Science and he also has appointments, by courtesy, in the Department of Biomedical Engineering and the Department of Statistics. In 2008 he was awarded

the the Digiteo Chaire d'Excellence, sponsored by Digiteo Research Park in Paris, located at the Ecole Supérieure d'Electricité, Gif-sur-Yvette, France. He has held other visiting positions at LIDS Massachusetts Institute of Technology (2006), Boston University (2006), I3S University of Nice, Sophia-Antipolis, France (2001), Ecole Normale Supérieure de Lyon (1999), Ecole Nationale Supérieure des Télécommunications, Paris (1999), Lucent Bell Laboratories (1999), Scientific Research Labs of the Ford Motor Company, Dearborn, Michigan (1993), Ecole Nationale Supérieure des Techniques Avancées (ENSTA), Ecole Supérieure d'Electricité, Paris (1990), and M.I.T. Lincoln Laboratory (1987 - 1989). Alfred Hero is a Fellow of the Institute of Electrical and Electronics Engineers (IEEE). He has been plenary and keynote speaker at major workshops and conferences. He has received several best paper awards including: a IEEE Signal Processing Society Best Paper Award (1998), the Best Original Paper Award from the Journal of Flow Cytometry (2008), and the Best Magazine Paper Award from the IEEE Signal Processing Society (2010). He received a IEEE Signal Processing Society Meritorious Service Award (1998), a IEEE Third Millennium Medal (2000) and a IEEE Signal Processing Society Distinguished Lectureship (2002). He was President of the IEEE Signal Processing Society (2006-2007). He sits on the Board of Directors of IEEE (2009-2011) where he is Director Division IX (Signals and Applications).

Alfred Hero's recent research interests have been in detection, classification, pattern analysis, and adaptive sampling for spatio-temporal data. Of particular interest are applications to network security, multi-modal sensing and tracking, biomedical imaging, and genomic signal processing.



**Silvio Savarese** Silvio Savarese received the B.S./M.S. degree (Summa Cum Laude) from the University of Napoli Federico II (Italy) in 1999 and a PhD in Electrical Engineering from the California Institute of Technology in 2005. He joined the University of Illinois at Urbana-Champaign from 2005 to 2008 as a Beckman Institute Fellow. Since 2008 he has been an Assistant Professor of Electrical Engineering at the University of Michigan, Ann Arbor. He is recipient of an NSF Career Award in 2011 and Google Research Award in 2010. In 2002 he was

awarded the Walker von Brimer Award for outstanding research initiative. He served as workshops chair and area chair in CVPR 2010, and as area chair in ICCV 2011. Silvio Savarese has been active in promoting research in the field of object recognition and scene representation. He co-chaired and co-organized the 1st, 2nd and 3rd edition of the IEEE workshop on 3D Representation for Recognition (3dRR-07, 3dRR-09, 3dRR-11) in conjunction with the ICCV. He was editor of the Elsevier Journal in Computer Vision and Image Understanding, special issue on 3D Representation for Recognition in 2009. He authored a book chapter on "Studies in Computational Intelligence-Computer Vision", edited by Springer in 2010 and co-authored a book on 3D object and scene representation published by Morgan and Claypool in 2011. His work has received several best paper awards including the CETI Award at the 2010 FIATECHs Technology Conference. His research interests include computer vision, object recognition and scene understanding, shape representation and reconstruction, human activity recognition and visual psychophysics.