INFORMATIVE SUBSPACES FOR AUDIO-VISUAL PROCESSING: HIGH-LEVEL FUNCTION FROM LOW-LEVEL FUSION

John W. Fisher III [*]

Massachusetts Institute of Technology
Cambridge, MA 02139

Trevor Darrell [†]

Massachusetts Institute of Technology
Cambridge, MA 02139

ABSTRACT

We propose a new probabilistic model of single source multi-modal generation, and show algorithms for maximizing mutual information which find correspondences between signal components. We show a nonparametric method for finding informative subspaces that captures complex statistical relationships between different modalities. We extend a previous subspace method to include new priors on the projection weights, yielding more robust results. Applied to human speakers, our model finds a relationship between audio speech and video of facial motion, and partially segments background events in both channels. We present new results on the problem of audio-visual verification, and show how the audio and video of a speaker can be matched without a prior model of the speaker's voice or appearance.

1. INTRODUCTION

Relating multi-modal signals is a challenging task for automated perception systems. Given signals from multiple modalities, one would like to find correspondences: portions of the signals from the same underlying source. In the domain of audio and video (A/V) of human speakers this ability is useful for a variety of tasks. Namely, speaker localization in a video frame, speaker audio enhancement with respect to noise, and verification as to whether the observed person is actually the person speaking.

We propose an independent cause model to capture the relationship between generated signals in each individual modality. Using principles from information theory we show how an approach for learning maximally informative joint subspaces can find cross-modal correspondences. Nonparametric statistical models have been used to measure the degree of mutual information (MI) in complex A/V phenomena [1]. This method simultaneously learns projections of images and periodograms from the A/V sequence. The projections are computed adaptively such that the image and audio projections have maximum MI. Early results of this method applied to A/V data have been reported [1], but without any derivation from a probabilistic framework. In this paper we ground the MI algorithm in a probabilistic model, and extend the informative subspace algorithm to include a prior bias toward small projection coefficients. We also present new results on the problem of A/V verification without prior models of user speech or appearance, an application not previously addressed in the literature.

In the next section we review related work on audio-visual fusion. We then present our probabilistic model for cross-modal signal generation, and show how audio-visual correspondences can be found by identifying components with maximal MI. We then review techniques for efficient estimation of MI using non-parametric entropy models. Finally, we show a new application to a verification task where we detect whether audio and video come from the same speaker. In an experiment comparing the audio and video of every combination of a group of eight users, our technique was able to perfectly match the corresponding audio and video from a single user. These results are based purely on the instantaneous cross-modal mutual information of the two signals, and do not rely on any prior experience or model of user's speech or appearance.

2. RELATED WORK

There has been substantial progress on feature-level integration of speech and vision for speech recognition (e.g. Meier *et al* [2] and Stork [3]). However, many of these systems assume that no significant motion distractors are present and that the camera was "looking" at the source of the audio signal. Indeed, speech systems are easily confused if there are nearby speakers also making utterances. Our method, which is not specific to A/V, is used to detect whether audio and video signals come from the same or different sources.

Other work, closely related to ours, is that of Hershey and Movellan [4] which examined the per-pixel correlation relative to an audio track, detecting which pixels have related variation. An inherent assumption of this method was that of joint Gaussian statistics. Slaney and Covell [5] consider temporal alignment between audio and video tracks,

[*] J. Fisher is supported in part by MIT Project Oxygen

[†] T. Darrell is supported in part by MIT Project Oxygen

0-7803-7402-9/02/$17.00 ©2002 IEEE

but do not address the problem of detecting whether two signals came from the same person or not. Their technique was more general than [4] in that pixel changes were considered jointly, although there is also an implicit Gaussian assumption and use of training data. We are not aware of any prior work which can perform audio-visual verification at a signal-level without prior speech or appearance models.

3. LINEAR FUSION MODEL

For reasons of brevity, we consider the following linear fusion model (any *differentiable* model is suitable).

$$\begin{bmatrix} y_1^v \cdots y_N^v \\ y_1^a \cdots y_N^a \end{bmatrix} = \begin{bmatrix} h_v^T & 0^T \\ 0^T & h_a^T \end{bmatrix} \begin{bmatrix} x_1^v \cdots x_N^v \\ x_1^a \cdots x_N^a \end{bmatrix} \qquad (1)$$

where $x_i^v \in \Re^{N_v}$ and $x_i^a \in \Re^{N_a}$ are lexicographic samples of images and periodograms, respectively, from an A/V sequence. The linear projection defined by $h_v^T \in \Re^{M_v \times N_v}$ and $h_a^T \in \Re^{M_a \times N_a}$ maps A/V samples to low dimensional features $y_i^v \in \Re^{M_v}$ and $y_i^a \in \Re^{M_a}$.

We address two issues here. First, is the criterion (and algorithm) for designing the features. Specifically, we shall treat x_i's and y_i's as samples of random variables X and Y. We propose (and justify) maximizing the MI, $I(Y^v; Y^a)$, between the derived features defined as [6]

$$I(Y^v; Y^a) = h(Y^v) + h(Y^a) - h(Y^v, Y^a) \qquad (2)$$

as the criterion for choosing h_v and h_a, where

$$h(Y) = -\int p_Y(u) \log(p_Y(u))\, du \qquad (3)$$

is the differential entropy of either the marginal or joint feature densities. The means by which we approximate entropy and infer densities from samples differentiates our approach from other methods. More importantly, they lead to a computationally tractable algorithm with the capacity to model complex joint A/V properties. The second issue we address is that of solving the system of equations described by 1. The number of samples, N, is much less than the dimension of the A/V measurements, $N_v + N_a$, consequently the system of equations is under-determined. We propose constraints on the projections which improve performance.

4. MI AS A FUSION CRITERION

Many multimodal scenes can be modeled with one common A/V source and distinct interfering sources for each modality. Each observation combines information from the joint and interfering sources for that channel. Figure 1 represents the model graphically. High-dimensional observations of each modality, $\{X^v, X^a\}$, are independent *conditioned* on the causes $\{A, B, C\}$ (B is the only common cause). The joint statistical model consistent with figure 1a is

$$P(A,B,C,X^a,X^v) = P(A)P(B)P(C) \\ P(X^a|A,B)P(X^v|B,C)\,.$$

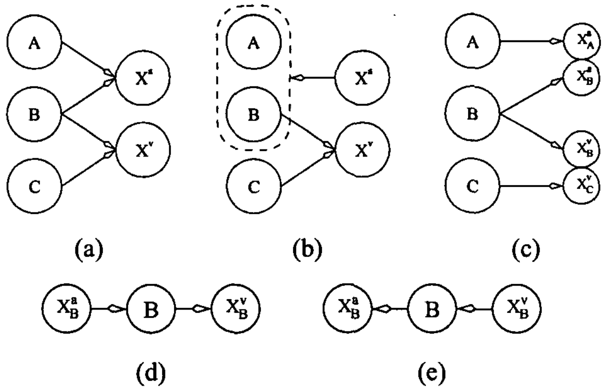

Fig. 1. Graphs of pertinent statistical models: (a) the independent cause model - $\{X^a, X^v\}$ are independent conditioned on $\{A, B, C\}$, (b) information about X^a contained in X^v is conveyed through *joint* statistics of A and B, (c) the graph implied by the existence of a separating function, and (d) two equivalent Markov chains which can be extracted from the graphs.

Bayes' rule yields the model (graph of figure 1b)

$$P(A,B,C,X^a,X^v) = P(X^a)P(C)P(A,B|X^a)P(X^v|B,C)$$

in which information about X^a contained in X^v is conveyed through the *joint* statistics of A and B. Consequently, we cannot disambiguate the influences of A and B on the measurements. A similar model results when conditioning on X^v. However, if decompositions $X^a = [X_A^a, X_B^a]$ and $X^v = [X_B^v, X_C^v]$ *exist* (e.g. a segmented image or filtered audio) such that the joint density

$$P(A,B,C,X^a,X^v) = P(A)P(B)P(C) \\ P(X_A^a|A)P(X_B^a|B)P(X_B^v|B)P(X_C^v|C)$$

can be written (graph of figure 1c) we can extract a Markov chain containing elements related only to B. Figure 1d shows equivalent Markov chain graphs.

Given the decomposition, it can be shown via the data processing inequality [6] that the following inequalities hold

$$I(X_B^a; X_B^v) \le I(X_B^a; B) \quad \text{and} \quad I(X_B^a; X_B^v) \le I(X_B^v; B)$$

The inequalities hold for functions of X_B^a and X_B^v as well (e.g. Y^a and Y^v), so maximizing $I(Y_B^a; Y_B^v)$ increases $I(Y_B^a; B)$ and $I(Y_B^v; B)$. The implication is that such fusion discovers the underlying common cause of the observations.

5. FINDING INFORMATIVE SUBSPACES

Features with high mutual information are desirable as they are predictive of each other. However, we wish to avoid strong assumptions about the features' joint statistics (e.g. joint gaussianity). The method of [1] uses a nonparametric density estimate for which there is an *analytic* gradient of the mutual information with respect to projection parameters. As in [7], we replace *each* entropy term in equation 2 with a second-order Taylor series expanded about the uniform density, $p_u(x)$,

$$\begin{aligned} \hat{I}\left(Y^v;Y^a\right) \propto &- \int \left(\hat{p}_{Y^a}(u) - p_u(u)\right)^2 du \\ &- \int \left(\hat{p}_{Y^v}(u) - p_u(u)\right)^2 du \\ &+ \int \left(\hat{p}_{Y^v,Y^a}(u,z) - p_u(u,z)\right)^2 du dz \,. \end{aligned} \quad (4)$$

Note, $\hat{p}(x)$ is a Parzen density estimate [8] defined as

$$\hat{p}\left(y\right) = \frac{1}{N}\sum_i \kappa\left(y - y_i, \Sigma\right) \quad (5)$$

where we use a Gaussian product kernel for $k(\ )$ (i.e. $\Sigma = \sigma^2 I$). The resulting approximation is the integrated squared error between the density of the projections to the uniform density. As stated, this particular choice of entropy approximation and density estimate lead to a closed form gradient of MI with respect to the projection coefficients which can be computed by evaluating a *finite* number of functions at a *finite* number of points in the output space. The update term for the *individual* entropy terms in 4 (note the opposite sign on the third term) of sample y_i at iteration k as a function of y_i's at iteration $k-1$ is (where y_i denotes a sample of either Y^a or Y^v or their concatenation depending on which term of 4 is being computed)

$$\Delta {y_i}^{(k)} = b_r({y_i}^{(k-1)}) - \frac{1}{N}\sum_{j \neq i} \kappa_a\left({y_i}^{(k-1)} - {y_j}^{(k-1)}, \Sigma\right) \quad (6)$$

$$b_r(y_i)_j \approx \frac{1}{d}\left(\kappa\left(y_i + \frac{d}{2}, \Sigma\right)_j - \kappa\left(y_i - \frac{d}{2}, \Sigma\right)_j\right) \quad (7)$$

$$\begin{aligned} \kappa_a\left(y,\Sigma\right) &= \kappa(y,\Sigma) * \kappa'(y,\Sigma) \\ &= -\left(2^{M+1}\pi^{M/2}\sigma^{M+2}\right)^{-1} \exp\left(-\frac{y^T y}{4\sigma^2}\right) y \end{aligned} \quad (8)$$

where $M = M_a$, M_v, or $M_a + M_v$ again depending on which entropy term. Both $b_r(y_i)$ and $\kappa_a(y_i, \sigma)$ are vector-valued functions (M-dimensional) and d is the support of the output (i.e. a hyper-cube with volume d^M). The notation $b_r(y_i)_j$ indicates the jth element of $b_r(y_i)$. Adaptation consists of the update rule above followed by a modified least squares solution for h_v and h_a until a local maximum is reached. In the experiments that follow $M_v = M_a = 1$ with 150 to 300 iterations.

Early results [1] demonstrated video localization of a speaker. However, the technique often failed to converge, as a consequence of the under-determined system of equations 1. To improve on the method, we thus introduce a prior on h_v and h_a in the form of L_2 penalties. Additionally, we constrain h_v to have minimum average output energy when convolved with images in the sequence. This penalty was proposed by Mahalanobis *et al* [9] for optimized correlator design. The difference being that in their case the outputs were designed explicitly while here they are derived from the MI optimization step of equation 6. The adaptation criterion, maximized via coordinate descent, is then:

$$J = \hat{I}\left(Y^v;Y^a\right) - \alpha_v h_v^T h_v - \alpha_u h_a^T h_a - \beta h_v \bar{R}_v h_v \quad (9)$$

where $\bar{R}_v$ is average autocorrelation function of the images in the sequence and can be combined efficiently with the L_2 penalty term. The scalar weighting terms α_v, α_u, β, were set using a data dependent heuristic for all experiments. The last term is more easily computed in the frequency domain (see [9]) and is equivalent to pre-whitening the images using the inverse of the average power spectrum. Pre-whitening accentuates edges in the input image. Moving edges (lips, chin, etc.) to convey the most information about the audio.

6. EMPIRICAL RESULTS

In our experiments we demonstrate speaker localization in the video and the measurement the A/V consistency. Simple techniques which check only for the presence of a face (or moving face) would fail when a person off-camera spoke a command. With an eye toward interchangeable devices, we are interested in the case where no prior voice or appearance model is available.

We collected A/V data from eight subjects. Images were collected at 30 frames/s (360x240 pixels). The audio was filtered to 10KHz. Subjects uttered the phrase "How's the weather in Taipei?", (2-2.5 seconds of data). Video frames were processed as is, while the audio was converted to a sequence of periodogram using a window length of 2/30 seconds. After computing h_v and h_a as described, $I(Y^a;Y^v)$, estimated from samples was used to measure consistency. Reported values are normalized with respect to the maximum possible value (i.e. two uniformly distributed random variables which perfectly predict one another). We assume no significant head movement on the part of the speaker.

Figure 2a shows results from two difference sequences. The top row corresponds to an A/V sequence of a single speaker and a video monitor while the bottom row is from an A/V sequence with one speaker (on left) and one moving person who is not speaking. Column (a) is an image from the sequences while column (b) is the magnitude of the image after prewhitening. Column (c) is an image of the pixel-wise standard deviations of the image sequence, note that motion distractors have more energy than the speaker in both cases. Column (d) shows the magnitude of the resulting h_v (as an image). In both cases it is the facial features of the speaker that are highlighted. In addition to video localization of the audio, we also check A/V consistency. Such a test is useful when the person to which a system is visually attending is not the person who actually spoke. Having learned an MI-optimized projection, we estimate the MI and

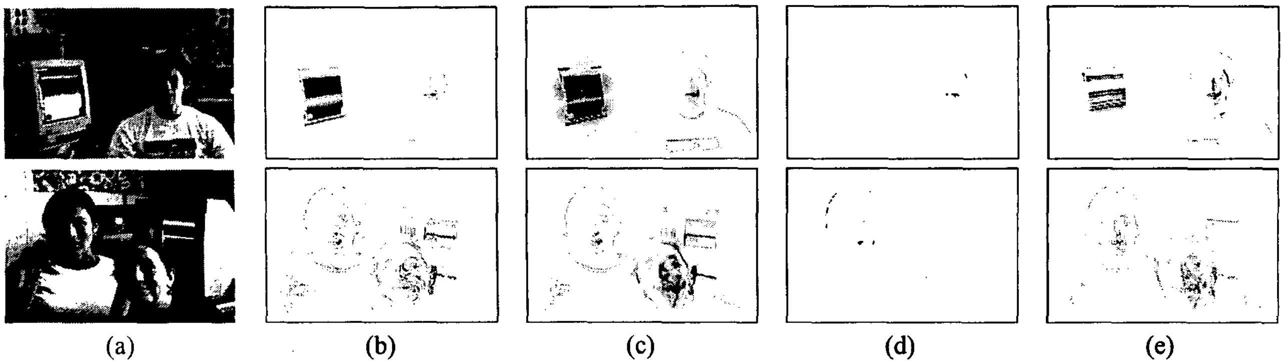

Fig. 2. Results from two A/V sequences: Top sequence contains one speaker and flickering monitor. Bottom sequence contains one speaker (left) and person with random mouth movements: (a) image from sequence, (b) prewhitened image (magnitude), (c) pixel standard deviation image, (d) image of h_v (correct audio), (e) image of h_v (incorrect audio).

Table 1. Summary of results over eight video sequences. The columns indicate which audio sequence was used while the rows indicate which video sequence was used.

	a1	a2	a3	a4	a5	a6	a7	a8
v1	**0.68**	0.19	0.12	0.05	0.19	0.11	0.12	0.05
v2	0.20	**0.61**	0.10	0.11	0.05	0.05	0.18	0.32
v3	0.05	0.27	**0.55**	0.05	0.05	0.05	0.05	0.05
v4	0.12	0.24	0.32	**0.55**	0.22	0.05	0.05	0.10
v5	0.17	0.05	0.05	0.05	**0.55**	0.05	0.20	0.09
v6	0.20	0.05	0.05	0.13	0.14	**0.58**	0.05	0.07
v7	0.18	0.15	0.07	0.05	0.05	0.05	**0.64**	0.26
v8	0.13	0.05	0.10	0.05	0.31	0.16	0.12	**0.69**

use it estimate to quantify the A/V consistency. Column (e) of figure 2 shows h_v when an alternate audio is compared to the image sequences. Note that the speaker is not localized in either case. Furthermore, the MI estimate drops from 0.68 to 0.05 (top sequence) and from 0.61 to 0.32 (bottom sequence) when the incorrect audio is used. Finally, data collected from six additional subjects is used to perform an eight-way test. Each video sequence was compared to each audio sequence; summarized in table 1. No attempt was made to align the mismatched audio sequences. The previous sequences correspond to subjects 1 and 2 in the table. In every case the matching A/V pairs exhibit the highest MI.

7. DISCUSSION

We have presented an information theoretic approach to the problem of finding cross-modal correspondence. A statistical formulation of joint signal generation was proposed, showing that maximizing MI detected the correspondences, via nformative subspaces. We proposed new priors on projection coefficients. Our approach was applied to the problem of audio-visual localization and verification, detecting the correspondence between the speech and appearance of a human speaker without a prior model in either domain. In all the cases presented, our technique correctly matched video with the corresponding audio from a particular individual, and localized the user's face in a video frame.

8. REFERENCES

[1] John W. Fisher III, Trevor Darrell, William T. Freeman, and Paul. Viola, "Learning joint statistical models for audio-visual fusion and segregation," in *Advances in Neural Information Processing Systems 13*, 2000.

[2] Uwe Meier, Rainer Stiefelhagen, Jie Yang, and Alex Waibel, "Towards unrestricted lipreading," in *Second International Conference on Multimodal Interfaces (ICMI99)*, 1999.

[3] G. Wolff, K. V. Prasad, D. G. Stork, and M. Hennecke, "Lipreading by neural networks: Visual preprocessing, learning and sensory integration," in *Proc. of Neural Information Proc. Sys. NIPS-6*, 1994, pp. 1027–1034.

[4] John Hershey and Javier Movellan, "Using audio-visual synchrony to locate sounds," in *Advances in Neural Information Processing Systems 12*, S. A. Solla, T. K. Leen, and K-R. Mller, Eds., 1999, pp. 813–819.

[5] Malcolm Slaney and Michele Covell, "Facesync: A linear operator for measuring synchronization of video facial images and audio tracks," in *Advances in Neural Information Processing Systems 13*, T. K. Leen, Thomas G. Dietterich, and Volker Tresp, Eds., 2000.

[6] Thomas M. Cover and Joy A. Thomas, *Elements of Information Theory*, John Wiley & Sons, Inc., New York, 1991.

[7] J.W. Fisher and J.C. Principe, "Unsupervised learning for nonlinear synthetic discriminant functions," in *Proc. SPIE, Optical Pattern Recognition VII*, D. Casasent and T. Chao, Eds., 1996, vol. 2752, pp. 2–13.

[8] E. Parzen, "On estimation of a probability density function and mode," *Ann. of Math Stats.*, vol. 33, pp. 1065–1076, 1962.

[9] A. Mahalanobis, B. Kumar, and D. Casasent, "Minimum average correlation energy filters," *Applied Optics*, vol. 26, no. 17, pp. 3633–3640, 1987.