

MIT



Informative Subspaces: High-level Function from Low-level Fusion

John Fisher

Oxygen Workshop, January, 2002



MIT

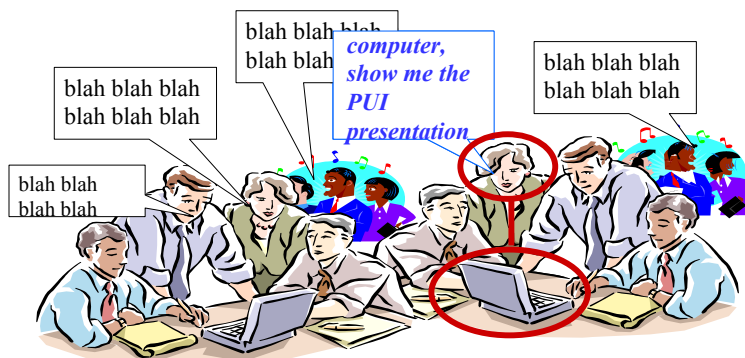
Overview



- **Motivation for *early* fusion of audio and video.**
 - Why is it hard?
 - Why is it useful?
- **A statistical model consistent with the approach.**
- **Information theoretic perspective.**
- **Algorithmic description**
- **Applications of the method**
 - Video localization of speaker
 - Audio Enhancement
 - Audio/video synchrony



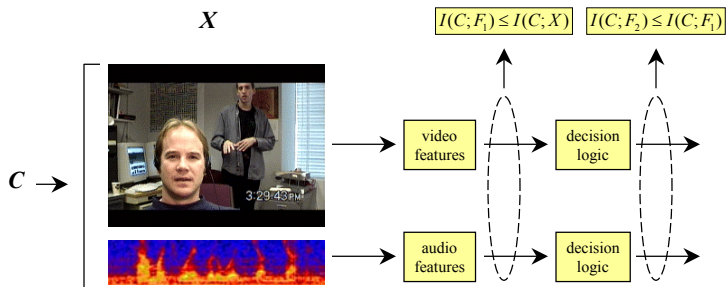
- Who is there? (presence, identity)
- Which person said that? (audiovisual grouping)
- Where are they? (location)
- What are they looking / pointing at? (pose, gaze)
- What are they doing? (activity)



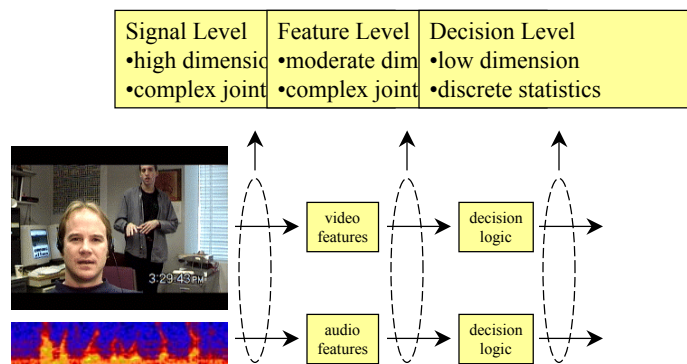
A perceptual link between user and device:

- detect and recognize user
- confirm that utterance was from user



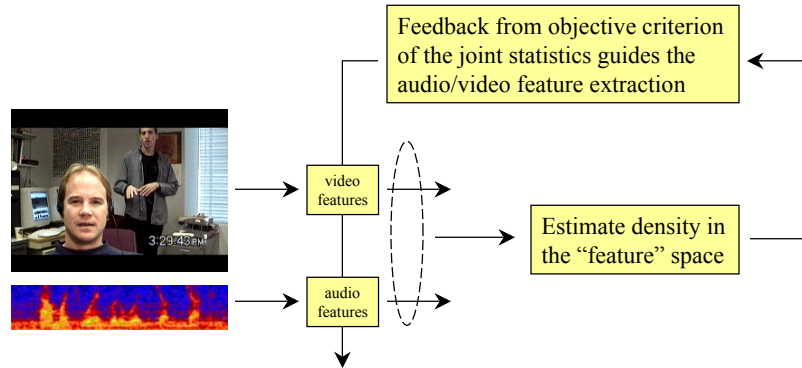


Data Processing Inequality : processing data can only destroy information



- Data fusion implies estimating the joint statistics of two or more data sources.
- The point of “fusion” is dictated by a tradeoff between robustness/simplicity.





- How might we (implicitly) model the joint statistics?
- What objective criterion is appropriate adapting the mapping from signals to features?

Sounds and motions which are consistent may be attributed to a common cause...



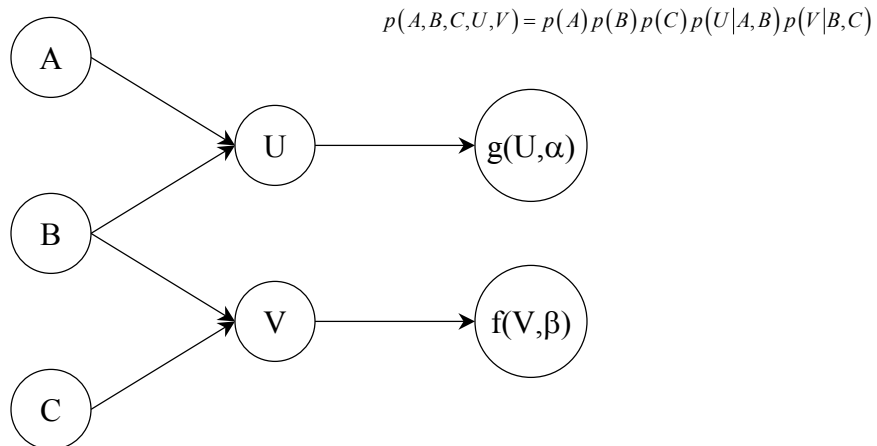
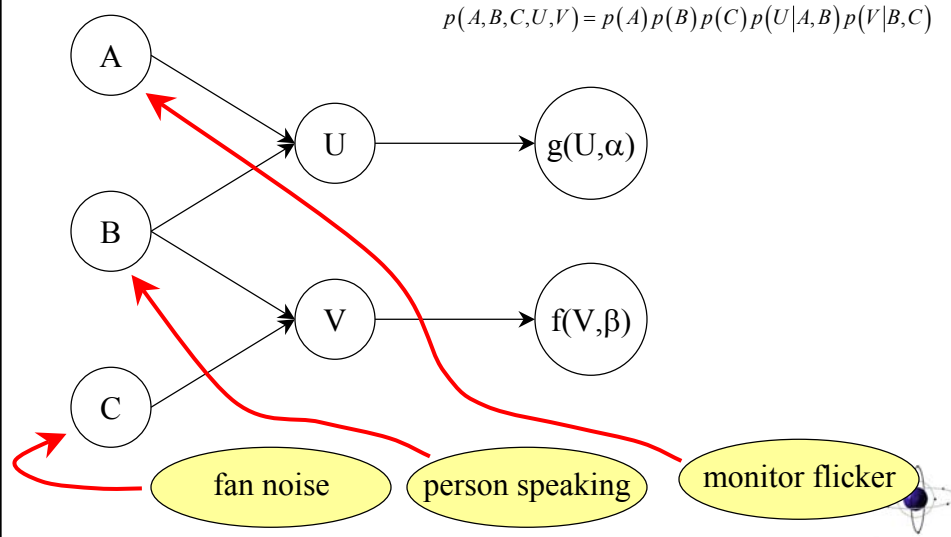
consistent



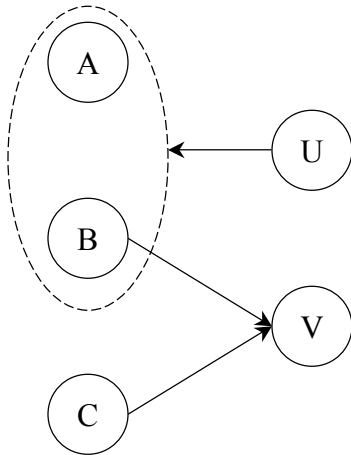
not

Question: How do we quantify “consistent”?





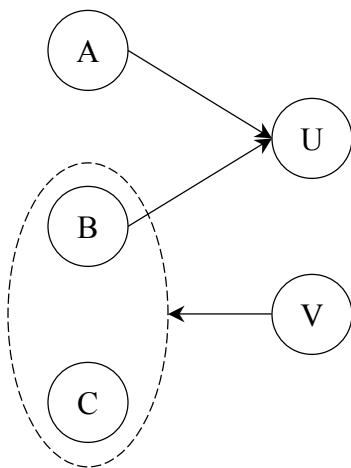
U conveys information about V through the joint of (A,B)



$$\begin{aligned}
 p(A,B,C,U,V) &= p(A)p(B)p(C)p(U|A,B)p(V|B,C) \\
 &= p(U)p(A,B|U)p(C)p(V|B,C)
 \end{aligned}$$

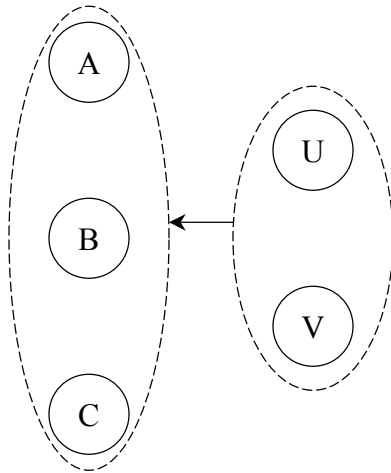


V conveys information about U through the joint of (B,C)

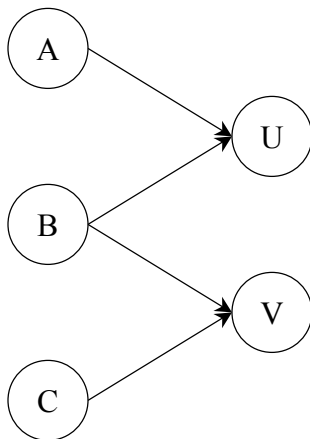


$$\begin{aligned}
 p(A,B,C,U,V) &= p(A)p(B)p(C)p(U|A,B)p(V|B,C) \\
 &= p(U)p(A,B|U)p(C)p(V|B,C) \\
 &= p(V)p(B,C|V)p(A)p(U|A,B)
 \end{aligned}$$





$$\begin{aligned}
 p(A, B, C, U, V) &= p(A)p(B)p(C)p(U|A, B)p(V|B, C) \\
 &= p(U)p(A, B|U)p(C)p(V|B, C) \\
 &= p(V)p(B, C|V)p(A)p(U|A, B) \\
 &= p(U, V)p(A, B, C|U, V)
 \end{aligned}$$



Suppose a partitioning of U and V *exists* such that:

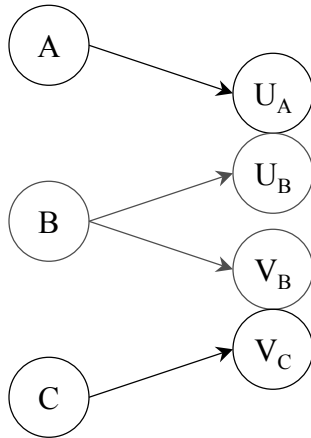
$$p(U, A, B) = p(A)p(B)p(U_A|A)p(U_B|B)$$

$$p(V, B, C) = p(B)p(C)p(V_B|B)p(V_C|C)$$

then...

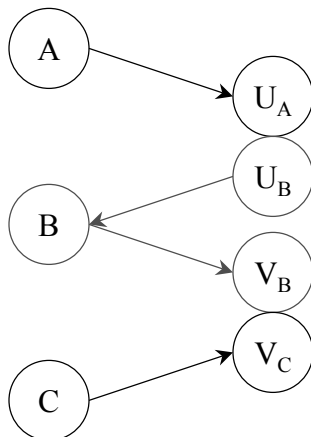
Bear in mind we still have the task of finding it.





becomes

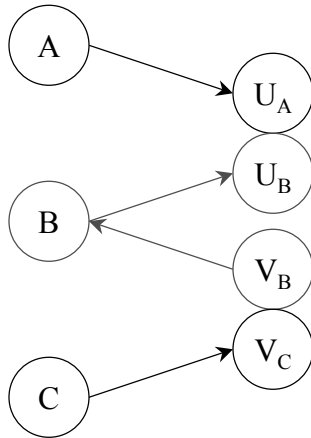
$$p(A, B, C) = p(A)p(B)p(C) \\ p(U_A|A)p(U_B|B)p(V_B|B)p(V_C|C)$$



or

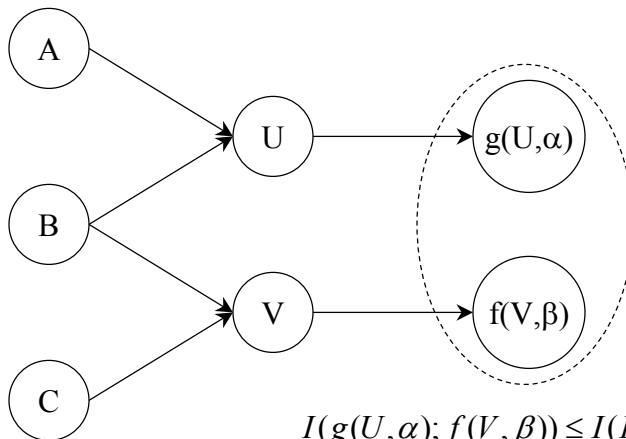
$$p(A, B, C) = p(A)p(B)p(C) \\ p(U_A|A)p(U_B|B)p(V_B|B)p(V_C|C) \\ = p(U_B)p(B|U_B)p(V_B|B) \\ p(A)p(C)p(U_A|A)p(V_C|C)$$





or

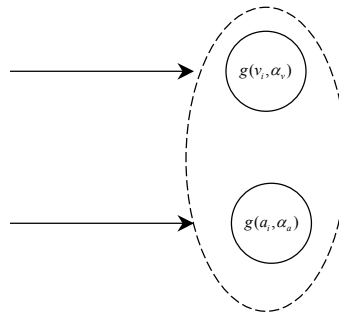
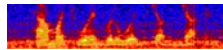
$$\begin{aligned}
 p(A, B, C) &= p(A)p(B)p(C) \\
 &= p(U_A|A)p(U_B|B)p(V_B|B)p(V_C|C) \\
 &= p(U_B)p(B|U_B)p(V_B|B) \\
 &= p(A)p(C)p(U_A|A)p(V_C|C) \\
 &= p(V_B)p(B|V_B)p(U_B|B) \\
 &= p(A)p(C)p(U_A|A)p(V_C|C)
 \end{aligned}$$



$$\begin{aligned}
 I(g(U, \alpha); f(V, \beta)) &\leq I(B; f(V, \beta)) \\
 &\leq I(B; g(U, \alpha))
 \end{aligned}$$



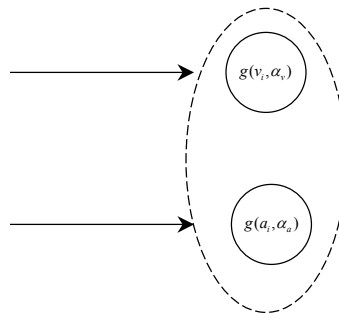
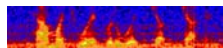
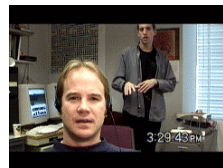
Fusion of Audio/Video Measurement using Mutual Information



- Choose the mapping parameters such that the mutual information between the extracted features is maximized (i.e. project onto a maximally informative subspace).



Fusion of Audio/Video Measurement using Mutual Information

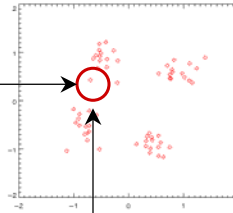


- By maximizing MI, we are summarizing the common information in the measurements, (i.e. which is related to their common cause).
- From the information theory perspective, the joint of the feature variables is a proxy for the “observable” part of their common cause.

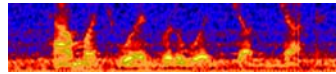




$$f_{v_i} = h_v^T V_i$$



$$f_{a_i} = h_a^T A_i$$



- Treat each image/audio frame in the sequence as a sample of a random variable.
- Projections optimize the *joint* audio/video statistics in the lower dimensional feature space.

- Mutual information quantifies the reduction in uncertainty (on average) about one random variable achieved by observing another.

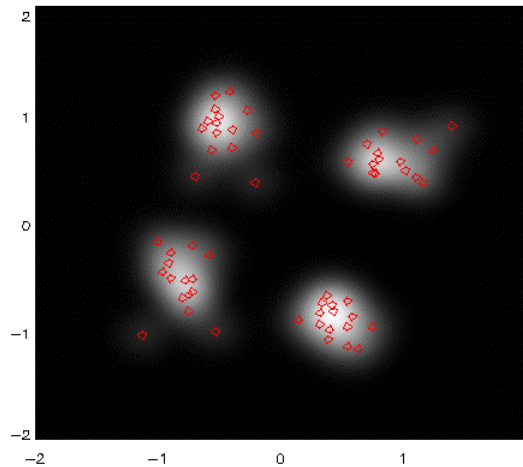
$$\begin{aligned} I(\theta, y) &= H(\theta) + h(y) - h(\theta, y) \\ &= H(\theta) - H(\theta|y) \\ &= h(y) - h(y|\theta) \end{aligned}$$

- The entropy terms depend on whether the random variable is discrete or continuous.

$$\begin{aligned} H(z) &= -\sum_i \log(p_z(z_i)) p_z(z_i) \quad , z \text{ discrete} \\ h(z) &= -\int_{\Omega_z} \log(p_z(z)) p_z(z) dz \quad , z \text{ continuous} \end{aligned}$$



Entropy vs. Moments (i.e. correlation)



Approximating Differential Entropy

- Substitute approximation into integral and simplify

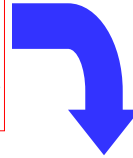
$$H(p) = \int_{\Omega_y} p(y) \log p(y) dy$$

$$\hat{H}(p) = H(p) + D(p\|q) - \int_{\Omega_y} \frac{1}{2q(y)} (p(y) - q(y))^2 dy$$

- Consequently, maximizing this approximation to entropy is equivalent to minimizing the chi-squared distance between the density, p , and the expansion density, q .

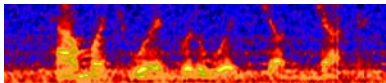


$$\begin{aligned}\hat{H}(\hat{p}) &= \log V_{\Omega_y} - \int_{\Omega_y} \frac{V_{\Omega_y}}{2} (\hat{p}(y) - p_u(y))^2 dy \\ &= \log V_{\Omega_y} - \int_{\Omega_y} \frac{V_{\Omega_y}}{2} \left(\frac{1}{N} \sum_{i=1}^N \kappa(y - y_i; h_N) - p_u(y) \right)^2 dy \\ &= \log V_{\Omega_y} - \int_{\Omega_y} \frac{V_{\Omega_y}}{2} \left(\frac{1}{N} \sum_{i=1}^N \kappa(y - g(x_i; \alpha); h_N) - p_u(y) \right)^2 dy\end{aligned}$$



Gradient of approximation can be computed exactly by evaluation of N functions at N sample locations.

$$\begin{aligned}\frac{\partial}{\partial \alpha} \hat{H} &= -\frac{V_{\Omega_y}}{N} \sum_{i=1}^N \left[\varepsilon_i \frac{\partial}{\partial \alpha} g(x_i; \alpha) \right] \\ \varepsilon_i &= f_r(y_i) - \frac{1}{N} \sum_{j \neq i} \kappa_a(y_i - y_j; h_N) \\ f_r(y) &= p_u(y) * \kappa_z(y; h_N) \\ \kappa_a(y; h_N) &= \kappa(y; h_N) * \kappa_z(y; h_N)\end{aligned}$$



- Which pixels are “related” to the associated audio?
- Joint statistics of video and audio modalities are not well modeled by parametric forms.
- Slaney and Covell (NIPS '00) demonstrate that canonical correlations (a second-order statistical measure) do not successfully detect audio/video synchrony using spectral representations.
- Classical sensor fusion approaches are formulated as joint Bayesian estimation problems, which is equivalent to MI in the non-parametric case.



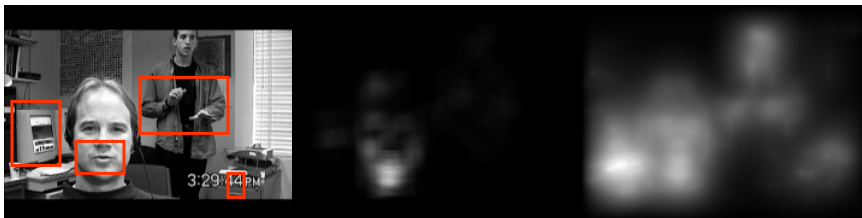
Detecting (change) motion is not enough



- Red squares indicate regions with large pixel variance
- Variance image of sequence at left
- Magnitude of MAX MI video projection shown at center
- Inspection of the learned video projection coefficients tells us which pixels are associated with the audio signal.









Pixel Representation vs. Motion Representation



- Similar result using an optic flow representation [Anandan '89] of motion in the video





-  •Left channel
-  •Right channel
-  •Wiener (left)
-  •Wiener (right)
-  •MI (left)
-  •MI (right)

In this experiment, regions of the video are selected for enhancement (e.g. face detector, manually).



	Wiener filter	Pixel-Periodogram Representation	Optical Flow-Periodogram Representation
SPG (male voice)	10.43 dB	8.9 dB	9.2 dB
SPG (female voice)	10.5 dB	5.7 dB	5.6 dB





MI: 0.68



0.61



0.19



0.20

Compute confusion matrix for 8 subjects:

	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

No errors!

No training!

Also can use for
audio/visual
temporal
alignment....

*While sounds and motions may be consistent
with each other, they are not always
attributable to a common cause...*



But: We can incorporate prior models when available?



- **We've motivated a tractable and computationally feasible information theoretic approach to signal level fusion.**
- **The method**
 - starts from a simple assumption (consistency at the signal level)
 - made little use of prior models
 - the statistical framework provides a clear methodology for incorporating prior models when their use is appropriate.
- **Tracking, tracking, tracking...**
- **Audio-video synchrony is a straightforward application of the approach, the fact that we learn a generative subspace model allows us to do more...**

