Review: Learning Bimodal Structures in Audio-Visual Data

CSE 704 : Readings in Joint Visual, Lingual and Physical Models and Inference Algorithms

Suren Kumar

Vision and Perceptual Machines Lab
106 Davis Hall
UB North Campus

surenkum@buffalo.edu

January 28, 2014

Summary

What? Learn bimodally informative structures from audio-visual signals

Why? Understand complex relationship between the inputs to different sensory modalities

How? Represent audio-video signals as sparse sum of kernels consisting of audio-waveform and a spatio-temporal visual basis.

Motivation: Biological Evidence, Evolution

Literature

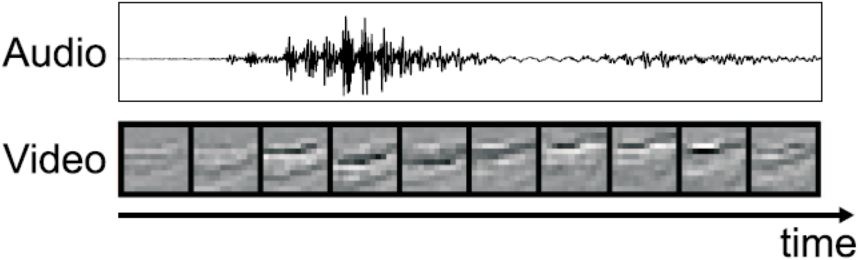
- Detect synchronous co-occurrences of transient structures in different modalities.
    1. Extract fixed and predefined unimodal features in audio and video stream separately
    2. Analyze correlation between the resulting feature representation

What is the problem with it?

Literature

- Detect synchronous co-occurrences of transient structures in different modalities.
  1. Extract fixed and predefined unimodal features in audio and video stream separately
  2. Analyze correlation between the resulting feature representation

What is the problem with it?



Basic Steps

- Capture bimodal signal structure by shift-invariance sparse generative model.
- Unsupervised learning for forming an overcomplete dictionary

p^{th} norm of a vector x in Euclidean space is defined as
$||x||_p = (\sum_{i=1}^{n} |x_i|^p)^{\frac{1}{p}}$

Matching Pursuit

Consider a dictionary $\mathcal{D} = (g)_{g \in \mathcal{D}}$ in a Hilbert space H.
The dictionary is countable, normalized ($||g|| = 1$), complete, but not orthogonal and possibly redundant.
Sparse Approximation Problem: Given $N > 0$ and $f \in H$, construct an N-term combination $f_N = \sum_{k=1,2,..N} c_k g_k$ with $g_k \in \mathcal{D}$ which approximates f "at best", and study how fast f_N converges to f.
Sparse Recovery Problem: if f has an unknown representation $f = \sum_g c_g g$ with $(c_g)_{g \in \mathcal{D}}$ a (possibly) sparse sequence, recover this sequence exactly or approximately from the data of f.
Building an optimal sparse representation of arbitrary signals is NP-hard problem.

Matching Pursuit

Matching pursuit is a greedy approximation to the problem.
Initialization $f_0 = 0$
Projection Step: At step $k-1$, projection step is the approximation of f
$f_{k-1} = Span\{g_1, ..., g_{k-1}\}$
Selection Step: Choice of next element based on residual
$r_{k-1} = f - f_{k-1}$
$g_k = \arg\max_{g \in \mathcal{D}} | < r_{k-1}, g > |$

Convolutional Generative Model

Audio visual data $s = (a, v)$, $a(t)$, $v(x, y, t)$
Dictionary $\{\phi_k\}$, $\phi_k = (\phi_k^{(a)}(t), \phi_k^{(v)}(x, y, t))$
Each atom can be translated to any point in space and time using operator $T_{(p,q,r)}$
$T_{(p,q,r)} = (\phi_k^{(a)}(t - r), \phi_k^{(v)}(x - p, y - q, t - r))$
Thus an audio-visual signal can be represented as
$s \approx \sum_{k=1}^{K} \sum_{i=1}^{n_k} c_{k_i} T_{(p,q,r)_{k_i}} \phi_k$, $c_{k_i} = (c_{k_i}^{(a)}, c_{k_i}^{(v)})$

Coding Find the coefficients (sparse)

Learning Learning the dictionary

Audio-Visual Matching Pursuit

Transient substructures that co-occur simultaneously are indicative of common underlying physical cause.

Discrete audio-visual translation τ

$$\mathcal{T}^{(\nu^{(a)},\nu^{(v)})}_{(p,q,r)} = \left(\mathcal{T}_\alpha, \mathcal{T}_{(p,q,\beta)}\right) = \mathcal{T}_{(p,q,\alpha,\beta)}$$

$$\begin{aligned} \alpha &= \operatorname{nint}(r/\nu^{(a)}) \in \mathbb{Z} \\ \beta &= \operatorname{nint}(r/\nu^{(v)}) \in \mathbb{Z}. \end{aligned}$$

Signal Approximation. Start with $R^0 s = s$

$$R^0 s = \left(\hat{c}_0^{(a)}\, \mathcal{T}_{\alpha_0}\phi_0^{(a)}, \hat{c}_0^{(v)}\, \mathcal{T}_{(p,q,\beta)_0}\phi_0^{(v)}\right) + R^1 s$$

$$\begin{aligned} \hat{c}_0^{(a)} &= \langle a, \mathcal{T}_{\alpha_0}\phi_0^{(a)}(t)\rangle \\ \hat{c}_0^{(v)} &= \langle v, \mathcal{T}_{(p,q,\beta)_0}\phi_0^{(v)}(x,y,t)\rangle. \end{aligned}$$

Audio-Visual Matching Pursuit

The function ϕ_0 and its spatio-temporal translation are chosen maximizing similarity measures $C(R^0 s, \phi)$
When to Stop?

Audio-Visual Matching Pursuit

The function ϕ_0 and its spatio-temporal translation are chosen maximizing similarity measures $C(R^0 s, \phi)$

When to Stop?

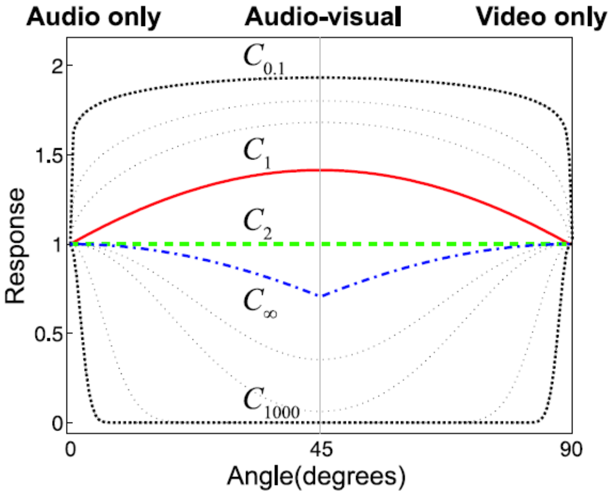
Number of iterations N or the maximum value of C between residual and dictionary elements falls below a threshold.

- Similarity measure should reflect properties of human perception
- Unaffected by small relative time-shifts

$$C_\rho(R^n s, \phi) = \|\langle R^n a, \mathcal{T}_\alpha \phi^{(a)} \rangle\|^\rho + \|\langle R^n v, \mathcal{T}_{(p,q,\beta)} \phi^{(v)} \rangle\|^\rho$$
$$\text{subject to} \quad \alpha \in [F \cdot (\beta - 1) + 1, F \cdot \beta] \, .$$

$$\{\alpha, \beta\} = \underset{\beta \in \mathbb{Z},\, \alpha \in [F \cdot (\beta-1)+1, F \cdot \beta]}{\arg\max} C_\rho(R^n s, \phi) \, ,$$

Choosing the norm



Learning

Find the kernel functions from a given set of audio-visual data.

$$p(s|\mathcal{D}) = \int p(s|\mathcal{D}, c)p(c)dc \approx p(s|\mathcal{D}, c^\star)p(c^\star)\,.$$

Assuming the noise to be gaussian

$$\log p(s|\mathcal{D}) \approx \frac{-1}{2\sigma_N^2} \left\| s - \sum_{k=1}^{K} \sum_{i=1}^{n_k} \hat{c}_{k_i} \mathcal{T}_{(p,q,\alpha,\beta)_{k_i}} \phi_k \right\|^2 .$$

Learning

Find the kernel functions from a given set of audio-visual data.

$$p(s|\mathcal{D}) = \int p(s|\mathcal{D}, c)p(c)dc \approx p(s|\mathcal{D}, c^\star)p(c^\star)\,.$$

Assuming the noise to be gaussian

$$\log p(s|\mathcal{D}) \approx \frac{-1}{2\sigma_N^2} \left\| s - \sum_{k=1}^{K} \sum_{i=1}^{n_k} \hat{c}_{k_i} \mathcal{T}_{(p,q,\alpha,\beta)_{k_i}} \phi_k \right\|^2 .$$

$$\begin{aligned} \frac{\partial \log(p(s|\mathcal{D}))}{\partial \phi_k} &\approx \frac{-1}{2\sigma_N^2} \frac{\partial}{\partial \phi_k} \left\{ s - \sum_{k=1}^{K} \sum_{i=1}^{n_k} \hat{c}_{k_i} \left\{ s - \hat{s} \right\}_{\mathcal{T}_{(p,q,\alpha,\beta)_{k_i}}} \right\}^2 \\ &= \frac{1}{\sigma_N^2} \sum_{i=1}^{n_k} \hat{c}_{k_i} \left\{ s - \hat{s} \right\}_{\mathcal{T}_{(p,q,\alpha,\beta)_{k_i}}} , \end{aligned} \qquad (9)$$

Learning

Update:

$$\phi_k[j] = \phi_k[j-1] + \eta\delta\phi_k$$

Learning

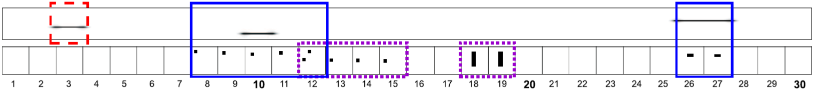
Key Idea : Update only one atom ϕ_k

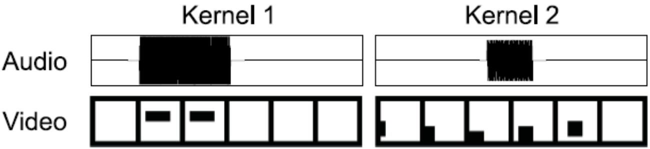
$$\frac{\partial \log(p(s|\mathcal{D}))}{\partial \phi_k} \approx \frac{-1}{2\sigma_N^2}\frac{\partial}{\partial \phi_k}\left\{s - \sum_{k=1}^{K}\sum_{i=1}^{n_k} \hat{c}_{k_i}\,\{s-\hat{s}\}_{\mathcal{T}_{(p,q,\alpha,\beta)_{k_i}}}\right\}^2$$
$$= \frac{1}{\sigma_N^2}\sum_{i=1}^{n_k} \hat{c}_{k_i}\,\{s-\hat{s}\}_{\mathcal{T}_{(p,q,\alpha,\beta)_{k_i}}}\,, \qquad (9)$$

Update each function with principal component of the residual errors. In general has faster convergence compared to GA.

Synthetic Data

Audio - 3 sine waves and video has four black shapes

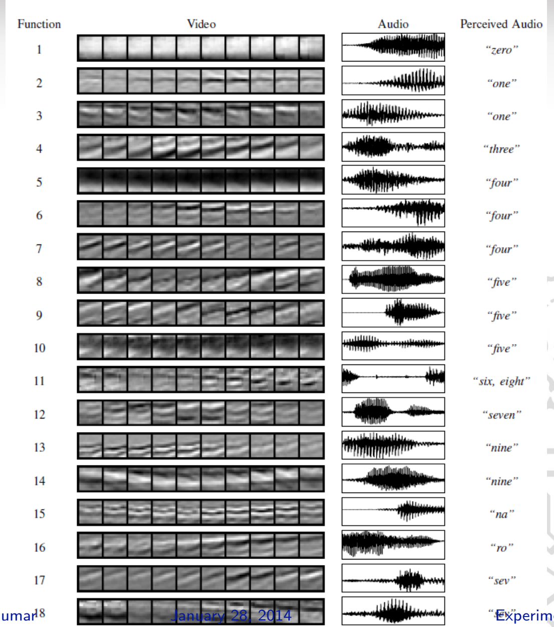




Audio-Visual Speech

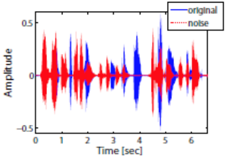
Speaker uttering the digits from zero to nine in english
Study convergence for 1 and 2 norm with GA and K-SVD

- C_1 encodes joint audio-visual structures better.
- K-SVD is faster compared to GA.


Function
Video
Audio
Perceived Audio
1 "zero"
2 "one"
3 "one"
4 "three"
5 "four"
6 "four"
7 "four"
8 "five"
9 "five"
10 "five"
11 "six, eight"
12 "seven"
13 "nine"
14 "nine"
15 "na"
16 "ro"
17 "sev"
18

Sound Source Localization

CUAVE - Visual Distractor and Acoustin Distractor



- Filter Audio
- Corresponding audio filter audio function and store the maximum projection
- Filter with corresponding video function and store maximum projection.
- Cluster video position

Summary

- Audio - Visual Matching Pursuit
- Similarity measures and learning algorithms
- Testing for speaker localization with distractors.

Reference

- http://www.math.tamu.edu/ popov/Learning/Cohen.pdf