Audio, Video, and their Joint Processing with Applications to Teleconferencing

Cha Zhang and Zhengyou Zhang
Communication and Collaboration Systems Group
Microsoft Research
7/11/2011

Microsoft Research

Telepresence

Bell Labs Video Conferencing System (1967)

HP Halo (2005)

Polycom RPX (2006)

Cisco Telepresence (2006)

And Tandberg, Telanetix, Teliris, …

Microsoft Research

Telepresence System Objectives

(Wainhouse Research 2007)

- High quality audio and video
- Simplicity
- High reliability
- Environmental excellence

Microsoft Research

Telepresence System Objectives

(Wainhouse Research 2007)

- High quality audio and video
  - Audio: Clear, low noise, low latency, echo-free, spatial audio
  - Video: Clear, life size, low noise, low latency, eye contact, 3D
- Simplicity
- High reliability
- Environmental excellence

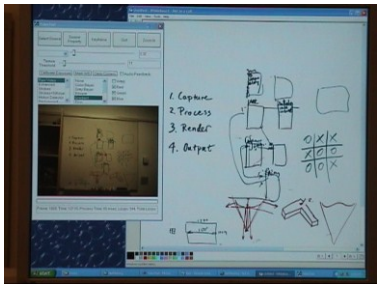
Microsoft Research

Related Systems at MSR

- Interactive whiteboard
- RingCam/RoundTable
- Interactive mobile teleconferencing
- Accelerated instant replay
- Personal telepresence station
- Viewport
- Embedded social proxies

Microsoft Research

Interactive Whiteboard



Microsoft Research

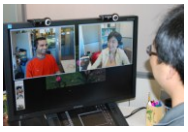
RingCam and RoundTable

Interactive Mobile Teleconferencing

Accelerated Instant Replay

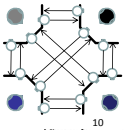
Personal Telepresence Station






Video

Full Mesh Geometry

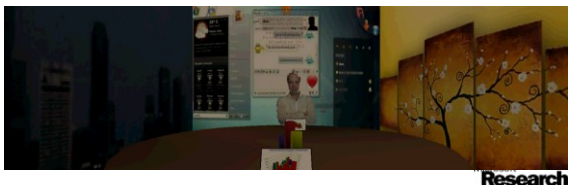


10

Microsoft Research

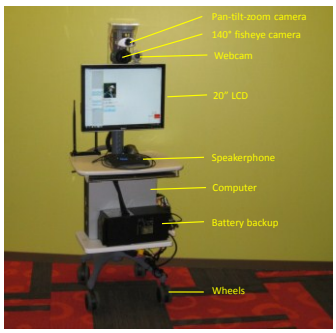
Viewport





Microsoft Research

Embodied Social Proxies



Video

12

Microsoft Research

Outline

- Introduction
- Sound source localization with compact mic arrays
- Spatial sound and multi-channel echo cancellation
- Video-based tracking
- Other video processing techniques
- Conclusions and future work

Outline

- Introduction
- Sound source localization with compact mic arrays
  - A maximum likelihood solution
  - Room modeling for enhancing range sensitivity
  - Joint audio/visual speaker localization
- Spatial sound and multi-channel echo cancellation
- Video-based tracking
- Other video processing techniques
- Conclusions and future work

Microsoft RoundTable

• Panoramic view from 5 high-resolution cameras
• Circular 6-microphone array for speech capturing and sound source localization
• Automated speaker switching

SSL Based on Time Difference of Arrival (TDOA)

- Cross-correlation

$$R_{ik}(\tau) = \int x_i(t) x_k(t-\tau) dt,$$

when $\tau = \frac{d}{c}$, one gets the max value. In freq. domain:

$$R_{ik}(\tau) = \int X_i(\omega) X_k^*(\omega) e^{j\omega\tau} d\omega,$$

In case of multiple mics, solve the following using hypothesis testing:

$$\begin{aligned}\mathcal{R}(\mathbf{s}) &= \sum_{i=1}^{P}\sum_{k=1}^{P}\int X_i(\omega)X_k^*(\omega)e^{j\omega(\tau_i-\tau_k)}d\omega, \\ &= \int \Big[\sum_{i=1}^{P} X_i(\omega)e^{j\omega\tau_i}\Big]\Big[\sum_{k=1}^{P} X_k(\omega)e^{j\omega\tau_k}\Big]^* d\omega, \\ &= \int \Big|\sum_{i=1}^{P} X_i(\omega)e^{j\omega\tau_i}\Big|^2 d\omega,\end{aligned}$$

Steered Response Power (SRP)

Microsoft Research

SSL Based on TDOA (cont.)

- Generalized cross-correlation

$$\mathcal{R}(\mathbf{s}) = \sum_{i=1}^{P}\sum_{k=1}^{P}\int \Psi_{ik}(\omega) X_i(\omega) X_k^*(\omega) e^{j\omega(\tau_i-\tau_k)} d\omega,$$

  - SRP-PHAT

$$\Psi_{ik}(\omega) = \frac{1}{|X_i(\omega)X_k^*(\omega)|} = \frac{1}{|X_i(\omega)||X_k(\omega)|}$$

$$\mathcal{R}(\mathbf{s}) = \int \Big|\sum_{i=1}^{P} \frac{X_i(\omega)e^{j\omega\tau_i}}{|X_i(\omega)|}\Big|^2 d\omega,$$

  - ML based SRP (weight optimal for 2-mic only)

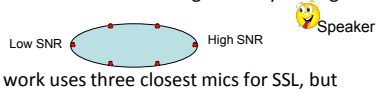
$$\Psi_{ij}(\omega) = \frac{|X_i(\omega)||X_j(\omega)|}{|N_i(\omega)|^2|X_j(\omega)|^2 + |N_j(\omega)|^2|X_i(\omega)|^2}.$$

Microsoft Research

Motivation

- RoundTable uses *directional* microphones
  - All mics point outwards
  - For sound capture, experimental results show:

    Directional mics + mic selection > omni-direcitonal mics + beamforming
  - Signal to noise ratio varies significantly among mics



  - Early work uses three closest mics for SSL, but performance is not great

Microsoft Research

Signal Model

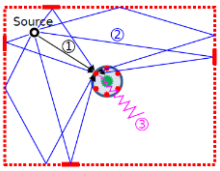
- Model $\mathbf{X}(\omega) = \underbrace{S(\omega)\mathbf{G}(\omega)}_{①} + \underbrace{S(\omega)\mathbf{H}(\omega)}_{②} + \underbrace{\mathbf{N}(\omega)}_{③},$

Received signal: $\mathbf{X}(\omega) = [X_1(\omega), \cdots, X_P(\omega)]^T,$

Direct propagation gain: $\mathbf{G}(\omega) = [\alpha_1(\omega)e^{-j\omega\tau_1}, \cdots, \alpha_P(\omega)e^{-j\omega\tau_P}]^T,$

Environment response: $\mathbf{H}(\omega) = [H_1(\omega), \cdots, H_P(\omega)]^T,$

Additive noise: $\mathbf{N}(\omega) = [N_1(\omega), \cdots, N_P(\omega)]^T.$



Microsoft Research

Noise Modeling

- Define:

$$\mathbf{N}^c(\omega) = S(\omega)\mathbf{H}(\omega) + \mathbf{N}(\omega),$$

and assume it follows a zero-mean, independent between frequencies, joint Gaussian distribution

$$p(\mathbf{N}^c(\omega)) = \rho \exp\Big\{-\frac{1}{2}[\mathbf{N}^c(\omega)]^H \mathbf{Q}^{-1}(\omega)\mathbf{N}^c(\omega)\Big\},$$

$$\begin{aligned}\mathbf{Q}(\omega) &= E\{\mathbf{N}^c(\omega)[\mathbf{N}^c(\omega)]^H\} \\ &= E\{\mathbf{N}(\omega)\mathbf{N}^H(\omega)\} + |S(\omega)|^2 E\{\mathbf{H}(\omega)\mathbf{H}^H(\omega)\}\end{aligned}$$

Microsoft Research

Estimate Noise Covariance Matrix

$$\begin{aligned}\mathbf{Q}(\omega) &= E\{\mathbf{N}^c(\omega)[\mathbf{N}^c(\omega)]^H\} \\ &= E\{\mathbf{N}(\omega)\mathbf{N}^H(\omega)\} + |S(\omega)|^2 E\{\mathbf{H}(\omega)\mathbf{H}^H(\omega)\}\end{aligned}$$

- First term:

$$E(N_i(\omega)N_j^*(\omega)) = \lim_{K\to\infty}\frac{1}{K}\sum_{k=1}^{K} N_{ik}(\omega)N_{jk}^*(\omega)$$

- Second term (Wang 97, Rui 04):

$$|S(\omega)|^2 E\{\mathbf{H}(\omega)\mathbf{H}^H(\omega)\} \approx \mathrm{diag}(\lambda_1, \cdots, \lambda_P)$$

$$\begin{aligned}\lambda_i &= E\{|H_i(\omega)|^2|S(\omega)|^2\} \\ &\approx \gamma(|X_i(\omega)|^2 - E\{|N_i(\omega)|^2\})\end{aligned}$$

$0 < \gamma < 1$ is an empirical parameter

Microsoft Research

Formulation and Solution

- Maximize the likelihood of the received signals

$$p(\mathbf{X}|S,\mathbf{G},\mathbf{Q}) = \prod_{\omega} p(\mathbf{X}(\omega)|S(\omega),\mathbf{G}(\omega),\mathbf{Q}(\omega)),$$

- Results:

$$S(\omega) = \frac{\mathbf{G}^H(\omega)\mathbf{Q}^{-1}(\omega)\mathbf{X}(\omega)}{\mathbf{G}^H(\omega)\mathbf{Q}^{-1}(\omega)\mathbf{G}(\omega)}$$

and the ML based SSL algorithm shall maximize:

$$\int_{\omega} \frac{[\mathbf{G}^H(\omega)\mathbf{Q}^{-1}(\omega)\mathbf{X}(\omega)]^H \mathbf{G}^H(\omega)\mathbf{Q}^{-1}(\omega)\mathbf{X}(\omega)}{\mathbf{G}^H(\omega)\mathbf{Q}^{-1}(\omega)\mathbf{G}(\omega)} d\omega$$

through hypothesis testing.

Microsoft Research

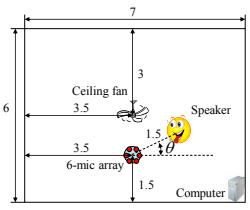
Some High-Level Messages

- It's a very general ML framework for multi-microphone SSL
- It does not require known microphone gain pattern (gain is estimated by the algorithm)
- It can weight microphones differently based on their noise variances
- It can handle certain amount of reverberation
- SRP-PHAT is a special case of the framework when SNR is extremely high
- It is equivalent to using a minimum variance distortionless response (MVDR) beamformer for SSL
- For detailed derivations, see Zhang 07.

Microsoft Research

A Synthetic Environment

- An omni-directional mic array in a virtual room
- Two simulated additive noise sources (ceiling fan and computer) at adjustable SNR
- Adjustable reverberation level
- Speaker is placed at different locations (θ=0,36°,72°,...,). Average SSL error rate of all locations is reported.



Microsoft Research

Results

Reverberation = 100 ms

Input SNR	SRP-PHAT		Proposed ML SSL	
	< 2°	< 10°	< 2°	< 10°
25 dB	**93.28%**	**94.41%**	92.36%	93.82%
20 dB	**83.35%**	84.92%	83.19%	**84.96%**
15 dB	79.24%	81.15%	**80.30%**	**82.35%**
10 dB	73.87%	76.71%	**76.29%**	**78.75%**
5 dB	66.59%	70.87%	**71.94%**	**75.30%**
0 dB	54.02%	60.78%	**64.19%**	**69.57%**

Reverberation = 500 ms

Input SNR	SRP-PHAT		Proposed ML SSL	
	< 2°	< 10°	< 2°	< 10°
25 dB	**61.53%**	**75.67%**	60.76%	74.57%
20 dB	**60.14%**	**73.88%**	59.79%	73.23%
15 dB	54.76%	**67.14%**	**54.78%**	66.97%
10 dB	52.45%	64.66%	**53.40%**	**65.22%**
5 dB	49.53%	61.96%	**50.80%**	**62.86%**
0 dB	45.11%	57.91%	**47.47%**	**60.14%**

Microsoft Research

Real-World Data

- 99 meetings captured in more than 50 rooms with various number of attendees
- 4 minutes each
- Speaker direction manually labeled (6706 frames in total)
- Input SNR ranges from 5 dB to 25 dB
- Report results on the percentage of frames that are within 6° (~20 pixels) and 14° (~40 pixels)

26

Microsoft Research

Results

Conventional Approach (closest 3 mics)		SRP-PHAT (6 mics)		Proposed method (6 mics)	
< 6°	< 14°	< 6°	< 14°	< 6°	< 14°
80.55%	86.85%	81.73%	88.13%	**83.49%**	**90.13%**

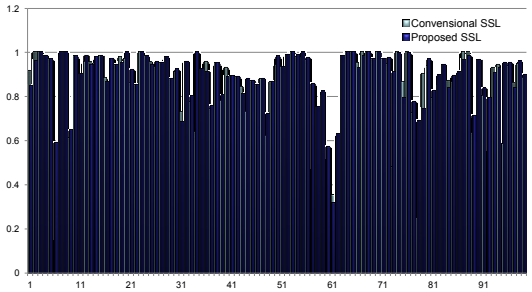
15.1% reduction in error

24.9% reduction in error

27

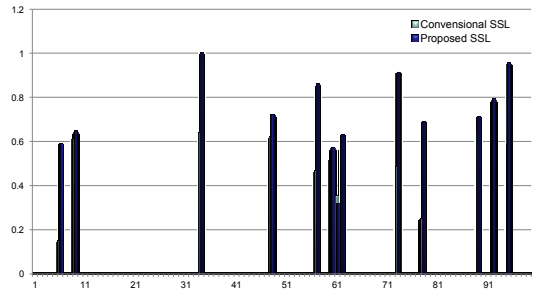
Microsoft Research

Results (cont.)



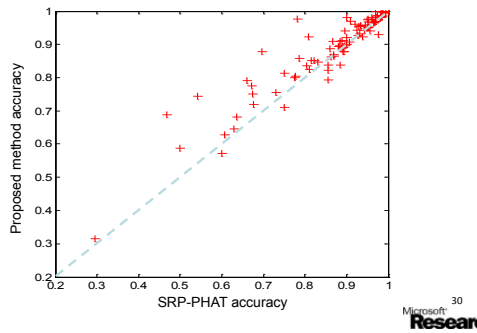
Microsoft Research

Results (cont.)



Microsoft Research

Results (cont.)



Microsoft Research

References

- M. Brandstein and H. Silverman, "A practical methodology for speech localization with microphone arrays," *Computer, Speech, and Language*, vol. 11, no. 2, pp. 91–126, 1997.
- T. Gustafsson, B. Rao, and M. Trivedi, "Source localization in reverberant environments: performance bounds and ml estimation," in *Proc. of ICASSP*, 2001.
- K. Harmanci, J. Tabrikian, and J. L. Krolik, "Relationships between adaptive minimum variance beamforming and optimal source localization," IEEE Trans. on Signal Processing, vol. 48, no. 1, pp. 1–12, 2000.
- C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-24, no. 4, pp. 320–327, 1976.
- Y. Rui and D. Florencio, "Time delay estimation in the presence of correlated noise and reverberation," in *Proc. of ICASSP*, 2004.
- Y. Rui, D. Florencio,W. Lam, and J. Su, "Sound source localization for circular arrays of directional microphones," in *Proc. of ICASSP*, 2005.
- H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in *Proc. of IEEE ICASSP*, 1997.
- C. Zhang, D. Florencio, D. Ba and Z. Zhang, "Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings," *IEEE Trans. on Multimedia*, pp. 538-548, Vol. 10, No. 3, Apr. 2008.
- C. Zhang, Z. Zhang and D. Florencio, "Maximum likelihood sound source localization for multiple directional microphones," in Proc. *ICASSP* 2007.

Microsoft Research

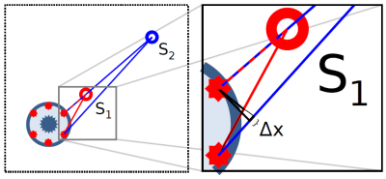
Outline

- Introduction
- Sound source localization with compact mic arrays
  - A maximum likelihood solution
  - Room modeling for enhancing range sensitivity
  - Joint audio/visual speaker localization
- Spatial sound and multi-channel echo cancellation
- Video-based tracking
- Other video processing techniques
- Conclusions and future work

Microsoft Research

Motivation

- Compact mic array has poor resolution in range and elevation for SSL
- Reverberation degrades SSL performance



Microsoft Research

Modeling the Reverberation

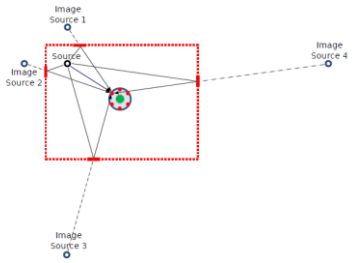
- Energy approximation
  - Wang 97, Rui 04, Zhang 07

$$|S(\omega)|^2 E\{\mathbf{H}(\omega)\mathbf{H}^H(\omega)\} \approx \mathrm{diag}(\lambda_1, \cdots, \lambda_P)$$
$$\lambda_i = E\{|H_i(\omega)|^2|S(\omega)|^2\} \approx \gamma(|X_i(\omega)|^2 - E\{|N_i(\omega)|^2\})$$

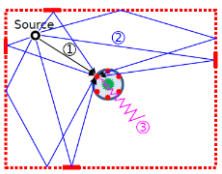
- Measure/estimate full room impulse response (Weinstein 94)
  - Very challenging
  - RIR change rapidly and significantly with the position and orientation of the source

Our Approach: Room Modeling

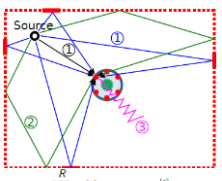
- The image model (Allen 79)



A New Signal Model



$$X_i(\omega) = \underbrace{\alpha_i(\omega)\, e^{-j\omega\tau_i} S(\omega)}_{①} + \underbrace{H_i(\omega)\, S(\omega)}_{②} + \underbrace{N_i(\omega)}_{③}$$



$$X_i(\omega) = \underbrace{\sum_{r=0}^{R} \alpha_i^{(r)}(\omega)\, e^{-j\omega\tau_i^{(r)}} S(\omega)}_{①} + \underbrace{H_i(\omega)\, S(\omega)}_{②} + \underbrace{N_i(\omega)}_{③}$$

Extend the ML Solution

- Consider: $X_i(\omega) = \underbrace{\sum_{r=0}^{R} \alpha_i^{(r)}(\omega)\, e^{-j\omega\tau_i^{(r)}}\, S(\omega)}_{①} + \underbrace{H_i(\omega)\, S(\omega)}_{②} + \underbrace{N_i(\omega)}_{③}$.
- Let: $G_i(\omega) = \sum_{r=0}^{R} \alpha_i^{(r)}(\omega)\, e^{-j\omega\tau_i^{(r)}} = g_i(\omega) e^{-j\varphi_i(\omega)}$

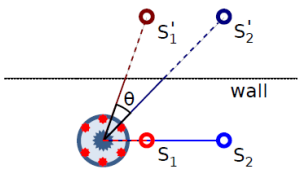
$$g_i(\omega) = \left| \sum_{r=0}^{R} \alpha_i^{(r)}(\omega)\, e^{-j\omega\tau_i^{(r)}} \right| \quad e^{-j\varphi_i(\omega)} = \frac{\sum_{r=0}^{R} \alpha_i^{(r)}(\omega)\, e^{-j\omega\tau_i^{(r)}}}{\left| \sum_{r=0}^{R} \alpha_i^{(r)}(\omega)\, e^{-j\omega\tau_i^{(r)}} \right|}$$

- Approximate: $e^{-j\varphi_i(\omega)} \approx \frac{\sum_{r=0}^{R} \frac{\rho_i^{(r)}}{r_i^{(r)}} e^{-j\omega\tau_i^{(r)}}}{\left| \sum_{r=0}^{R} \frac{\rho_i^{(r)}}{r_i^{(r)}} e^{-j\omega\tau_i^{(r)}} \right|}$ known
- Same solution as before

Microsoft Research

Range Discrimination

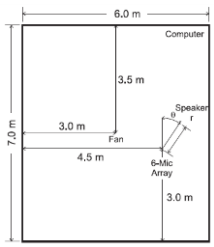
- The reflected sound source can help range discrimination
- Reverberation is now our friend!



Microsoft Research

Exp. Results – Synthetic Data

- Synthetic room
- Speaker at 1.3m to the array, tested at 0° elevation, and azimuth $\theta = 0°, 36°, 72°, \ldots$
- Reverberation T_{60} at 250 ms and 500 ms



Microsoft Research

Exp. Results – Synthetic Data

Closest Mic. SNR	VAD Voice Frames	ML-SSL $\Delta\theta > 2°$	ML-SSL $\Delta\phi > 1°$	ML-SSL $\Delta r > .15\,m$	R-ML-SSL $\Delta\theta > 2°$	R-ML-SSL $\Delta\phi > 1°$	R-ML-SSL $\Delta r > .15\,m$
25 dB	48	0.4 %	92.0 %	84.6 %	0.4 %	0.2 %	0.2 %
20 dB	47	0.2 %	90.6 %	81.6 %	0.2 %	0.0 %	0.0 %
15 dB	44	1.1 %	88.2 %	80.9 %	1.1 %	0.9 %	0.4 %
10 dB	38	3.4 %	85.6 %	81.6 %	3.4 %	2.9 %	2.1 %
5 dB	29	15.0 %	86.0 %	90.1 %	15.0 %	14.3 %	6.9 %
0 dB	16	24.8 %	85.1 %	94.3 %	24.8 %	22.2 %	15.5 %

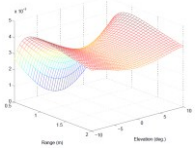
Table I
ERROR RATES FOR SYNTHETIC DATA, $T_{60} = 250\,ms$.

Closest Mic. SNR	VAD Voice Frames	ML-SSL $\Delta\theta > 2°$	ML-SSL $\Delta\phi > 1°$	ML-SSL $\Delta r > .15\,m$	R-ML-SSL $\Delta\theta > 2°$	R-ML-SSL $\Delta\phi > 1°$	R-ML-SSL $\Delta r > .15\,m$
25 dB	47	1.3 %	94.3 %	87.5 %	1.3 %	1.9 %	1.2 %
20 dB	45	1.1 %	93.2 %	86.3 %	1.1 %	1.7 %	1.5 %
15 dB	41	1.4 %	93.8 %	86.8 %	1.4 %	1.1 %	0.9 %
10 dB	38	2.6 %	94.1 %	88.5 %	2.6 %	2.6 %	1.6 %
5 dB	33	6.5 %	92.3 %	89.0 %	6.5 %	5.4 %	2.2 %
0 dB	23	21.4 %	91.0 %	93.0 %	21.4 %	19.3 %	12.4 %

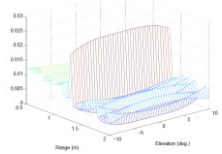
Table II
ERROR RATES FOR SYNTHETIC DATA, $T_{60} = 500\,ms$.

Microsoft Research

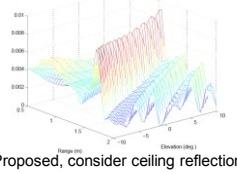
Exp. Results – Synthetic Data



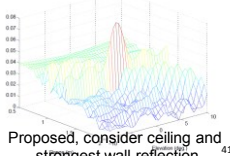
ML-SSL



Proposed, consider strongest wall reflection



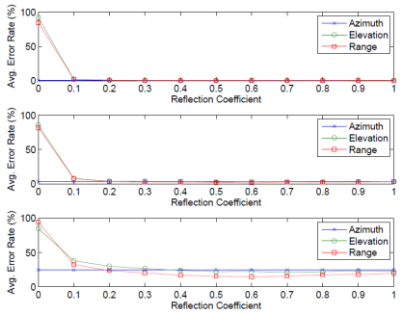
Proposed, consider ceiling reflection



Proposed, consider ceiling and strongest wall reflection

Microsoft Research

Sensitivity to Wall Reflection Coef.



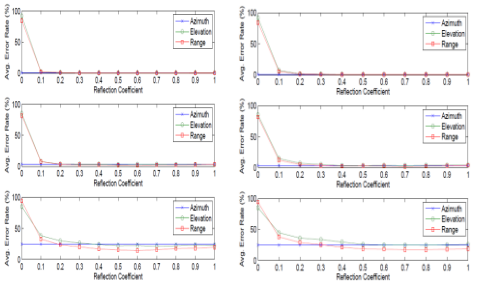
25 dB SNR

10 dB SNR

0 dB SNR

Microsoft Research

Sensitivity to Number of Walls

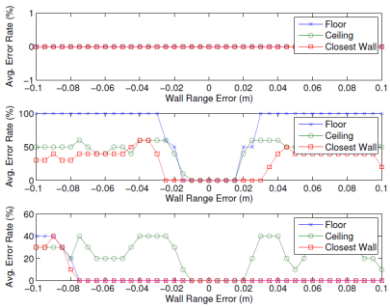


All walls considered

Only considers floor, ceiling and closest wall

Microsoft Research

Sensitivity to Wall Distance Error



Azimuth errors

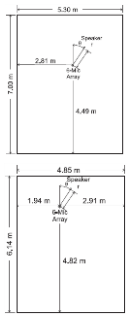
Elevation errors

Range errors

Microsoft Research

Exp. Results – Real Room Data

- Room 1
  - Distance to wall estimated using RoundTable device
  - 3 Closest reflectors used
- Room 2
  - Distance to wall measured by an ultrasonic range finder (1cm res.)
  - Reflection coef. set to 0.3



Microsoft Research

Exp. Results – Real Room Data

Speaker Position	# of Frames	ML-SSL			R-ML-SSL		
		$\Delta\theta > 10°$	$\Delta\phi > 5°$	$\Delta r > 0.15$	$\Delta\theta > 10°$	$\Delta\phi > 5°$	$\Delta r > 0.15$
$\theta \approx 0°, \phi \approx 2.5°, r \approx 1.70\,m$	11	0%	0%	55%	0%	0%	0%
$\theta \approx 60°, \phi \approx 2.5°, r \approx 1.47\,m$	21	0%	5%	95%	0%	9%	9%
$\theta \approx 120°, \phi \approx 2.5°, r \approx 1.37\,m$	9	0%	0%	33%	0%	11%	11%

Room 1

Speaker Position	# of Frames	ML-SSL			R-ML-SSL		
		$\Delta\theta > 10°$	$\Delta\phi > 5°$	$\Delta r > 0.15$	$\Delta\theta > 10°$	$\Delta\phi > 5°$	$\Delta r > 0.15$
$\theta \approx 60°, \phi \approx 26°, r \approx 1.85\,m$	77	0%	0%	100%	0%	5%	1%
$\theta \approx 60°, \phi \approx 18°, r \approx 2.40\,m$	71	0%	0%	87%	0%	3%	16%
$\theta \approx 120°, \phi \approx 32°, r \approx 1.60\,m$	72	3%	3%	96%	3%	3%	3%
$\theta \approx 120°, \phi \approx 22°, r \approx 2.15\,m$	68	0%	3%	90%	0%	2%	3%
$\theta \approx 180°, \phi \approx 15°, r \approx 3.10\,m$	82	1%	10%	99%	1%	2%	27%
$\theta \approx 180°, \phi \approx 12°, r \approx 4.00\,m$	76	0%	24%	96%	0%	1%	17%
$\theta \approx 240°, \phi \approx 36°, r \approx 1.45\,m$	75	12%	12%	96%	12%	9%	13%
$\theta \approx 300°, \phi \approx 40°, r \approx 1.40\,m$	84	7%	7%	98%	7%	6%	34%

Room 2

Microsoft Research

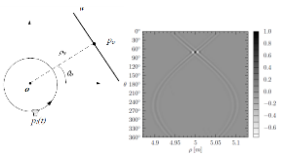
How to Estimate Room Model

- Visual methods
  - Range sensors
  - Landmark based (Kimber 09)
- Acoustic methods
  - Acoustic imaging (Moebus 07, O'Donovan 08)
  - Image model (Allen 79)
    - Single mic (Aprea 09, Antonacci 10, Dokmanic 11)
    - Mic arrays (Filos 10, Canclini 11)

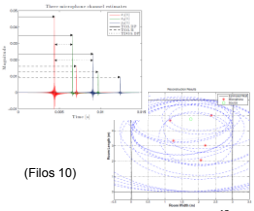
Microsoft Research

Image Model based Methods

- Two steps:
  - Measure room impulse responses (RIRs)
  - Infer room model



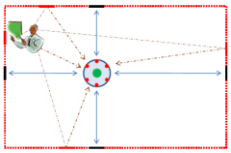
(Aprea 09)



(Filos 10)

Microsoft Research

Representation



$$R = \{(a_i, d_i, \theta_i, \phi_i)\}_{i=1}^{6}$$

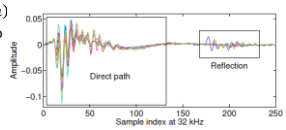
Reflection coef. of wall i.

3D position of wall i w.r.t. array center

Microsoft Research

RIR Modeling

- Signal observed at mic m:
  - $y_m(n) = h_m(n) * s(n) + u_m(n)$
  - $h_m(n)$: RIR from array center to mic m
- Only consider early reflections



$h_m^{(dp)}(n)$: direct path IR from array center to mic m

$h_m^{(r_i,\theta_i,\phi_i)}(n)$: IR at mic m from perfect wall at (r_i, θ_i, ϕ_i) (WIR)

R : number of modeled walls

Then

$$h_m(n) \approx h_m^{(dp)}(n) + \sum_{i=1}^{R} \rho^{(i)} h_m^{(r_i,\theta_i,\phi_i)}(n) + v_m(n)$$

Microsoft Research

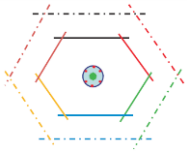
Infer the Wall Locations

- In vector form:

$$\mathbf{h} \approx \mathbf{h}^{(dp)}(n) + \sum_{i=1}^{R} \rho^{(i)} \mathbf{h}^{(r_i,\theta_i,\phi_i)} + \mathbf{v}$$

- Solve:

$$\min_{\mathbf{a}} \|\mathbf{h}_{room} - \mathbf{H}\mathbf{a}\|_2^2 + \lambda \|\mathbf{a}\|_1$$

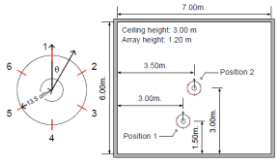


- Note we obtain a sparse solution for $\boldsymbol{a}$ due to L_1 optimization.
- Solved using LASSO (Tibshirani 96).

Microsoft Research

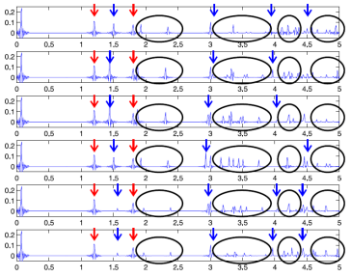
Exp. Results – Synthetic Data

- 6x7x3m room, 300ms reverberation
- Reflection coef. of 0.77
- SNR ~20dB
- Position 2 is more challenging



Microsoft Research

Exp. Results – Synthetic Data



Blue arrows: 1st order reflections with walls
Red arrows: 1st order reflections with floor/ceiling
Ellipses: high order reflections

Microsoft Research

Exp. Results – Synthetic Data

Ground Truth			Estimates		
r (m)	θ (°)	ϕ (°)	r (m)	θ (°)	ϕ (°)
1.200	0	-90	1.200	0	90
1.800	0	90	1.800	0	90
4.500	0	0	4.502	0	0
4.000	90	0	4.000	90	0
1.500	180	0	1.500	180	0
3.000	270	0	3.005	270	0

Position 1

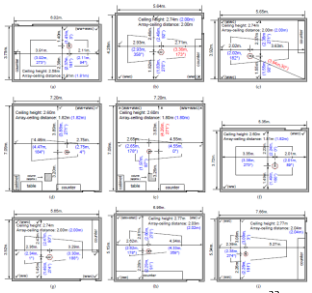
Ground Truth			Estimates		
r (m)	θ (°)	ϕ (°)	r (m)	θ (°)	ϕ (°)
1.200	0.0	-90	1.200	0.0	90
1.800	0.0	90	1.800	0.0	90
3.000	0.0	0	3.001	0.0	0
3.500	90.0	0	3.501	90.0	0
3.000	180.0	0	3.001	180.0	0
3.500	270.0	0	3.501	270.0	0

Position 2

Microsoft Research

Exp. Results – Real Room Data

- 9 conference rooms
- RoundTable device
  - WIR measured in anechoic chamber
  - Device on the table, no floor



Microsoft Research

Exp. Results – Real Room Data

Room	Ground Truth			Estimates			
	r (m)	θ (°)	ϕ (°)	r (m)	θ (°)	ϕ (°)	Refs
A	1.81	0	90	1.81	0	90	–
	1.40	0	0	1.41	3	0	111
	2.11	90	0	2.11	94	0	111
	2.36	180	0	2.37	181	0	100
	3.91	270	0	3.92	273	0	100
B	2.00	0	90	2.00	0	90	–
	2.93	0	0	2.93	358	0	101
	2.46	90	0	2.46	92	0	100
	2.71	180	0	3.36	173	0	001
	1.82	270	0	1.83	270	0	101
C	2.00	0	90	2.00	0	90	–
	3.63	0	0	–	–	–	–
	–	–	–	3.94	30	0	100
	1.55	90	0	1.55	90	0	101
	2.02	180	0	2.02	182	0	102
	2.37	270	0	2.38	271	0	101
D	1.82	0	90	1.82	0	90	–
	2.75	0	0	2.75	4	0	011
	3.20	90	0	–	–	–	–
	4.46	180	0	4.47	184	0	100
	3.89	270	0	3.88	271	0	101
E	1.80	0	90	1.80	0	90	–
	4.55	0	0	4.55	0	0	111
	3.26	90	0	–	–	–	–
	2.65	180	0	2.65	178	0	111
	3.83	270	0	4.26	271	0	022
F	1.81	0	90	1.82	0	90	–
	2.23	0	0	2.24	0	0	101
	2.01	90	0	2.01	89	0	001
	1.49	180	0	1.49	180	0	110
	3.35	270	0	3.36	273	0	010
G	2.00	0	90	2.00	0	90	–
	2.36	0	0	2.34	1	0	111
	2.47	90	0	2.47	90	0	101
	3.29	180	0	3.30	188	0	101
	1.45	270	0	1.46	274	0	111
H	2.03	0	90	2.02	0	90	–
	4.34	0	0	4.33	359	0	010
	2.28	90	0	2.29	91	0	110
	2.26	180	0	2.62	178	0	121
	2.87	270	0	2.86	270	0	121
I	2.04	0	90	2.04	0	90	–
	3.08	0	0	3.09	3	0	101
	5.27	90	0	–	–	–	–
	2.26	180	0	2.26	181	0	101
	2.39	270	0	2.38	274	0	102

Most walls are identified correctly, with a typical resolution of 1 cm.
Some distant walls are not correctly detected, though they won't impact SSL performance.

Microsoft Research

References

- J. Allen and D. Berkley, "Image method for efficiently simulating small room acoustics," *J. Acoust. Soc. Am*, vol. 65, no. 4, pp. 943–950, 1979.
- F. Antonacci, A. Sarti, and S. Tubaro, "Geometric reconstruction of the environment from its response to multiple acoustic emissions," in *Proc. of ICASSP*, 2010.
- D. Aprea, F. Antonacci, A. Sarti, and S. Tubaro, "Acoustic reconstruction of the geometry of an environment through acquisition of a controlled emission," in *Proc. of EUSIPCO*, 2009.
- D. Ba, F. Ribeiro, C. Zhang and D. Florencio, "L1 Regularized Room Modeling with Compact Microphone Arrays", in *Proc. of ICASSP*, 2010.
- A. Canclini, P. Annibale, F. Antonacci, A. Sarti, R. Rabenstein, and S. Tubaro, "From direction of arrival estimates to localization of planar reflectors in a two dimensional geometry," in *Proc. of ICASSP*, 2011.
- I. Dokmanic, Y.M. Lu, and M. Vetterli, "Can one hear the shape of a room: the 2-D polygonal case," in *Proc. of ICASSP*, 2011.
- J. Filos, E.A.P. Habets, and P.A. Naylor, "A Two-Step Approach to Blindly Infer Room Geometries," in *Proc. of the International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2010.
- D. Kimber, C. Chen, E.G. Rieffel, J. Shingu, and J. Vaughan, "Marking up a world: visual markup for creating and manipulating virtual models," in *Proc. of IMMERSCOM*, 2009.
- M. Moebus and A.M. Zoubir, "Three-Dimensional Ultrasound Imaging in Air using a 2D Array on a Fixed Platform," in *Proc. of ICASSP*, 2007.
- A. O'Donovan, R. Duraiswami, and D. Zotkin, "Imaging concert hall acoustics using visual and audio cameras," in *Proc. of ICASSP*, 2008.

Microsoft Research

References

- F. Ribeiro, C. Zhang, D. Florencio, and D. Ba, "Using Reverberation to Improve Range and Elevation Discrimination for Small Array Sound Source Localization," *IEEE Transactions on Audio, Speech and Language Processing*, pp. 1781-1792, Vol. 18, No. 7, Sep. 2010.
- Y. Rui and D. Florencio, "Time delay estimation in the presence of correlated noise and reverberation," in *Proc. of ICASSP*, 2004.
- R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society*, Series B (Methodological), vol. 58, no. 1, pp. 267–288, 1996.
- E. Weinstein, A. Oppenheim, M. Feder, and J. Buck, "Iterative and sequential algorithms for multisensor signal enhancement," *IEEE Trans. on Signal Processing*, vol. 42, no. 4, pp. 846–859, 1994.
- H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in *Proc. of IEEE ICASSP*, 1997.
- C. Zhang, D. Florencio, D. Ba and Z. Zhang, "Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings," *IEEE Trans. on Multimedia*, pp. 538-548, Vol. 10, No. 3, Apr. 2008.

Microsoft Research

Outline

- Introduction
- Sound source localization with compact mic arrays
  - A maximum likelihood solution
  - Room modeling for enhancing range sensitivity
  - Joint audio/visual speaker localization
- Spatial sound and multi-channel echo cancellation
- Video-based tracking
- Other video processing techniques
- Conclusions and future work

Microsoft Research

Motivation

- SSL only is still not accurate enough
  - Reverberation
  - Noise
- Can we use the panoramic video for help?

Microsoft Research

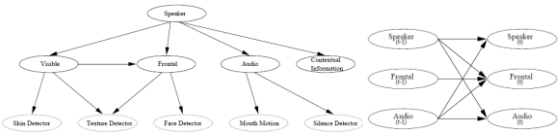
Previous Work

- SSL only speaker pointing (Polycom iPower 900)
  - Sensitive to reflections and noises
- SSL + face detection decision level fusion (DLF)
  - Very popular (Kapralos 02, Yoshimi 02)
  - RingCam used this approach (Cutler 02)

Microsoft Research

Previous Work

- Graphic models for audio/visual fusion (Pavlović 01, Beal 02)
  - Intuitive way to fuse multimodal data
  - Inference could be time-consuming

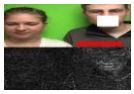


(Pavlović 01)

Microsoft Research

Previous Work

- Audio visual synchrony (Cutler00, Hershey00, Nock03)
  - Use mutual information between AV signals
  - Search for high correlation between AV



Light pixel represents high mutual information between AV (Nock03)

- Speaker tracking (Vermaak 01, Chen 04)
  - Fuse audio/video information using particle filters
  - Still need detection first

Microsoft Research

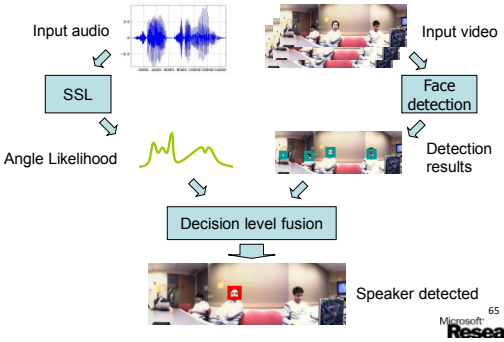
Why Is The Problem Hard?

- People don't always look at the camera
- There can be many meeting attendees
- Color calibration is hard, hence skin color-based face detection doesn't work
- Extremely low computation budget (100 MIPS)
- Image resolution is low, face can be as small as 10x10 pixel

64

Microsoft Research

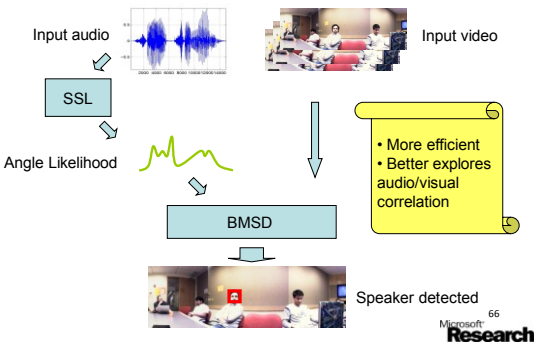
Decision Level Fusion



65

Microsoft Research

BMSD



66

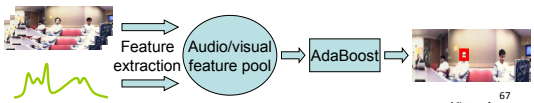
Microsoft Research

AdaBoost

- Adaptively focus on the difficult examples
- Boost the weak learner to a strong classifier by linear combination

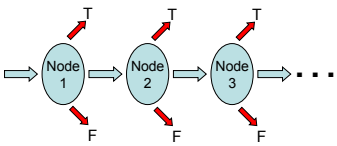
$$H(x) = \Sigma \alpha_t * h_t(x)$$

- Greedily pick the best feature according to the current sample weight distribution
- Can be used for automatic feature selection



Microsoft Research

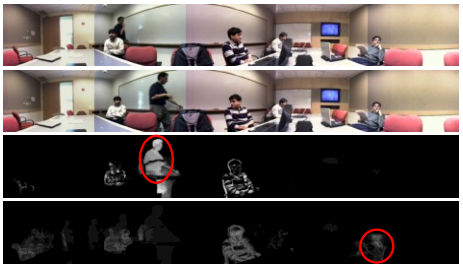
Attention Cascade



- Each node can contain one or multiple features
- Most negative examples are rejected at early stages
- Extremely efficient

Microsoft Research

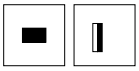
Video Feature Extraction



$$M_t = min\left(|I_t - I_{t-1}|, |I_t - I_{t-2}|\right)$$
$$R_t = \alpha M_t + (1-\alpha) R_{t-1}$$

Microsoft Research

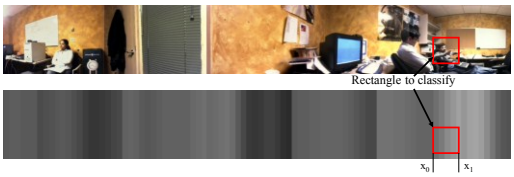
Video Feature Extraction (cont.)



Haar features (Viola 01)
on all three images.

Microsoft Research

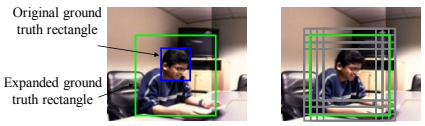
Audio Feature Extraction



1. $\frac{L^l_{\max}-L^g_{\min}}{L^g_{\max}-L^g_{\min}}$	2. $\frac{L^l_{\min}-L^g_{\min}}{L^g_{\max}-L^g_{\min}}$	3. $\frac{L^l_{avg}-L^g_{\min}}{L^g_{\max}-L^g_{\min}}$
4. $\frac{L^l_{\text{mid}}-L^g_{\min}}{L^g_{\max}-L^g_{\min}}$	5. $\frac{L^l_{\max}}{L^l_{\min}}$	6. $\frac{L^l_{\max}}{L^l_{avg}}$
7. $\frac{L^l_{\min}}{L^l_{avg}}$	8. $\frac{L^l_{\text{mid}}}{L^l_{avg}}$	9. $\frac{L^l_{\max}-L^l_{\min}}{L^l_{avg}}$
10. $\frac{L^l_{\max}}{L^g_{\max}}$	11. $\frac{L^l_{\min}}{L^g_{\max}}$	12. $\frac{L^l_{avg}}{L^g_{\max}}$
13. $\frac{L^l_{\text{mid}}}{L^g_{\max}}$	14. $\frac{L^l_{\max}-L^l_{\min}}{L^g_{\max}}$	15. $L^g_{\max}-L^l_{\max}<\epsilon$

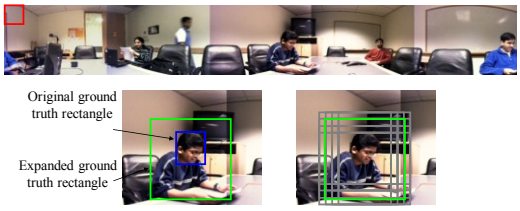
Microsoft Research

Training and Testing



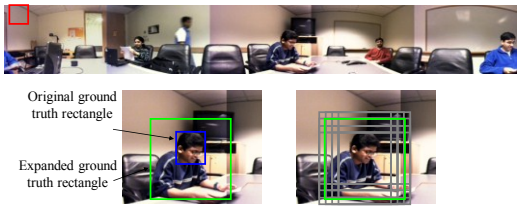
Microsoft Research

Training and Testing



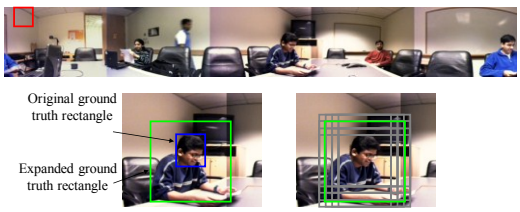
Microsoft Research

Training and Testing



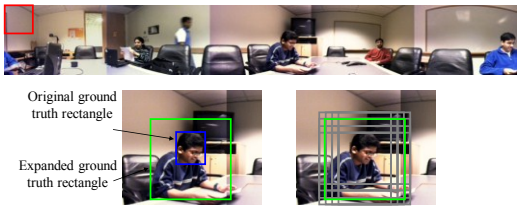
Microsoft Research

Training and Testing



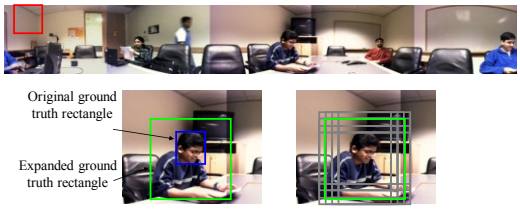
Microsoft Research

Training and Testing



Microsoft Research

Training and Testing



Microsoft Research

Merge of Detected Windows

- Multiple detections could exist for a frame
- Output a single speaker location
- Two methods to merge the windows
  - Projection and merge (PAM)
    - Project to azimuth axis, accumulate histogram
    - Find peak of histogram
    - Average nearby windows
  - Top N merge (TNM)
    - Select top N windows with highest confidence
    - Apply PAM above

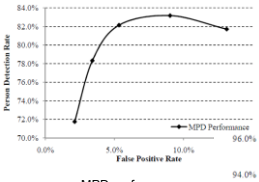
Microsoft Research

Experiments

- Training: 93 meetings, 3204 labeled frames in 10 different conference rooms
- Testing: 82 meetings, 1502 labeled frames in 10 meeting rooms (different from training)
- Speaker detection in each frame independently
- Compare SSL-only, DLF (SSL+MPD) and BMSD

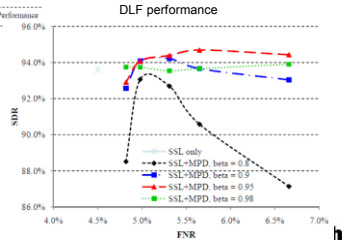
Microsoft Research

Decision Level Fusion

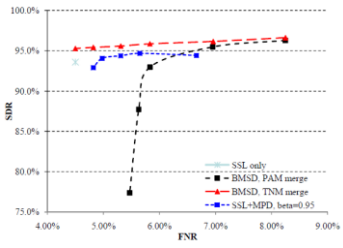


MPD performance

DLF performance

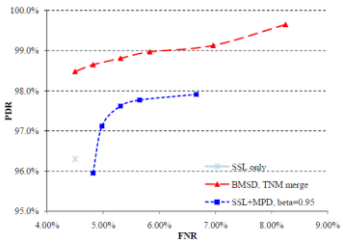


Speaker Detection Performance



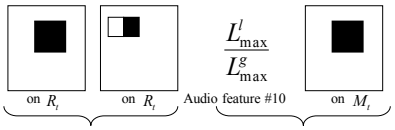
Microsoft Research

Person Detection Performance



Microsoft Research

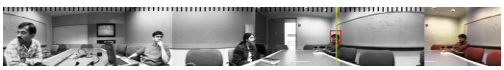
Features Selected by AdaBoost

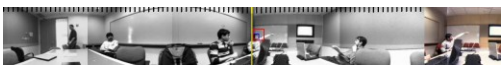


- BMSD much faster than DLF
  - Audio features help pruning
  - Audio features are extremely fast to compute

Microsoft Research

Examples





Microsoft Research

References

- M. Beal, H. Attias, and N. Jojic, "Audio-video sensor fusion with probabilistic graphical models," in *Proc. of ECCV*, 2002.
- Y. Chen and Y. Rui, "Real-time speaker tracking using particle filter sensor fusion," *Proceedings of the IEEE*, vol. 92, no. 3, pp. 485–494, 2004.
- R. Cutler and L. Davis, "Look who's talking: speaker detection using video and audio correlation," in *Proc. of IEEE ICME*, 2000.
- R. Cutler, Y. Rui, A. Gupta, J. Cadiz, I. Tashev, L. He, A. Colburn, Z. Zhang, Z. Liu, and S. Silverbert, "Distributed meetings: a meeting capture and broadcasting system," in *Proc. ACM Conf. on Multimedia*, 2002.
- J. Hershey and J. Movellan, "Audio vision: using audio-visual synchrony to locate sounds," in *Advances in Neural Information Processing Systems*, 2000.
- B. Kapralos, M. Jenkin, and E. Milios, "Audio-visual localization of multiple speakers in a video teleconferencing setting," York University, Canada, Tech. Rep., 2002.
- H. Nock, G. Iyengar, and C. Neti, "Speaker localisation using audio-visual synchrony: an empirical study," in *Proc. of CIVR*, 2003.
- V. Pavlović, A. Garg, J. Rehg, and T. Huang, "Multimodal speaker detection using error feedback dynamic Bayesian networks," in *Proc. of IEEE CVPR*, 2001.
- J. Vermaak, M. Gangnet, A. Black, and P. P´erez, "Sequential Monte Carlo fusion of sound and vision for speaker tracking," in *Proc. of IEEE ICCV*, 2001.
- B. Yoshimi and G. Pingali, "A multimodal speaker detection and tracking system for teleconferencing," in *Proc. ACM Conf. on Multimedia*, 2002.
- C. Zhang, P. Yin, Y. Rui, R. Cutler, P. Viola, X. Sun, N. Pinto and Z. Zhang, "Boosting-Based Multimodal Speaker Detection for Distributed Meeting Videos," *IEEE Trans. on Multimedia*, pp. 1541-1552, Vol. 10, No. 8, Dec. 2008.

Microsoft Research

Outline

- Introduction
- Sound source localization with compact mic arrays
- Spatial sound and multi-channel echo cancellation
  - Head phone based spatial sound
  - Loudspeaker spatial sound
  - Multi-channel echo cancellation
- Video-based tracking
- Other video processing techniques
- Conclusions and future work

Microsoft Research

Motivation

- Current audio conferencing systems
  - Monaural → adequate for 1-to-1
  - Poor when #people > 2
- Why poor?
  - All the voice streams are intermixed into a single one
  - Huge cognitive load: *Do 2 things simultaneously*
    - *Associate voice signals to the speaker*
    - *Comprehend what is being discussed*

Microsoft Research

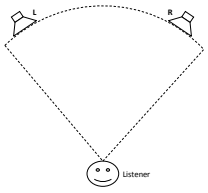
Benefits of Spatial Audio

- Human's cocktail party effect
  - Selective attention
  - Only spend effort on comprehension
- Brain rejects incoherent signals at two ears
  - Reverberation & noise are disregarded (not for mono!)
- Benefits (Baldis 01)
  - Memory
  - Focal Assurance
  - Perceived Comprehension
  - Listener's Preference

89
Microsoft Research

Multiparty Spatial Audio Conferencing

- Virtual seating



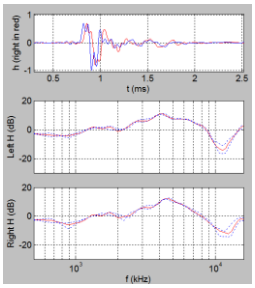
90
Microsoft Research

Headphone based Spatial Sound

- Situations requiring privacy
  - Open offices, public spaces, home office, etc
  - On-the-go: mobile devices
- Modern hardware
  - high-quality stereo ear-canal headphones
  - Compact Bluetooth headsets (monaural)
- Controlled environment than loudspeakers
- Easy job for AEC!

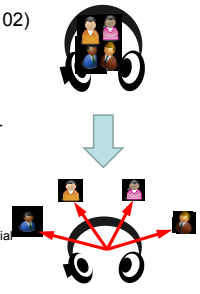
Background

- HRTF/HRIR (Algazi 01)
  - Linear response of each ear to
    - Each azimuth & elevation
    - Captured in an anechoic room
  - Contributing factors
    - Head shadow
    - Pinna/concha/torso
    - Air absorption
  - Gives sense of directivity
    - Individualization is hard



Background

- Headphone "virtualization" (Zlzer 02)
  - Input: regular audio samples
  - Output: simulating a listening environment
    - Localization & externalization
- Model the signal path (Schroeder 62, Stautner 82)
  - transmitter, medium & listener
  - Room response
    - Most important factor for externalization
    - Classical solution: various forms of artificial "reverb"
      - e.g. Schroeder's original comb filters



Challenges & Solution

- Unique requirements for conferencing
  - True room response is subtle
  - HRTF calibration is difficult
  - Sensitivity to human voice
- Higher realism is the solution
  - Capture “combined” transfer function
  - One or few models suffices
  - Localized parameterization possible
  - More computation but acceptable today

94

Microsoft Research

Offline Measurement

- Overall transfer function

$$Y(\omega) = H_l H_r H_t X(\omega) + N$$

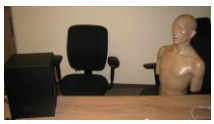
- Probe signal
  - Exponential chirp

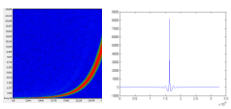
$$x(t) = \sin k_t e^{k_2 t}$$

  - de-convolution
    - At high SNR:

$$\hat{h} = y * x(-t) \approx h * x(t) * x(-t) \approx h$$

  - TBD
    - Comparison & Noise analysis



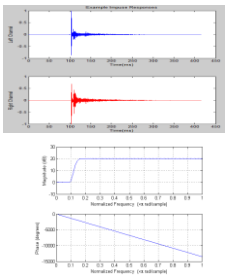


95

Microsoft Research

Post-processing

- Sounds bad if use as is
- Alignment & truncation
- Equalization
  - Match the spectrum of the original and processed signal
- Acoustic ratio
  - Between the direct component & the rest
  - Parameterization & personalization



96

Microsoft Research

Online Synthesis

- Partitioning the filters in time
  - Short filter h^s
    - Location dependent, up to 10ms
    - Source, HRTF & early reflections
  - Long filter h^l
    - Location independent, up to 100+ ms
    - Reverberation: omni-directional
- Lower complexity implementation

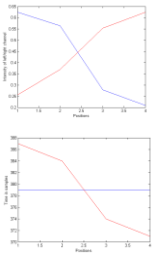
$$y_i(t;i) = x_i(t) * h_i^s(t;i) + h^l(t) * \sum_j x_j(t)$$

$$y_l(t) = \sum_i y_l(t;i)$$

Microsoft Research

Animation

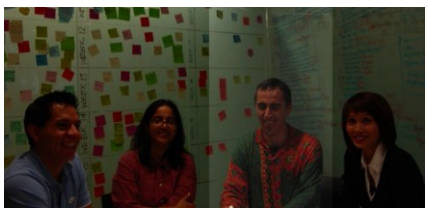
- Important for added realism
  - Natural head motion while talking
  - With video, help resolve HRTF mismatch
- Classical limitation of non-model based approach
  - Non-parameterized $h_i^s(t;i) \rightarrow h_i^s(t;\phi)$
  - Too many to measure
- Efficient implementation
  - Extract and interpolate three cues: AR, IID, ITD
  - Locally modify transfer function
    $h_i^s(t;\phi) = \Delta\alpha h_i^s(t - \Delta\theta; i)$
  - TBD: smooth, large range animation



Microsoft Research

Demo

- Please put on headphone



Microsoft Research

References

- V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio*, 2001.
- J. Baldis, "Effects of spatial audio on memory, compression, and preference during desktop conferences," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2001, pp. 166-173.
- W. Chen and Z. Zhang, "Highly realistic audio spatialization for multiparty conferencing using headphones," in *Proc. of MMSP*, 2009.
- M. R. Schroeder, "Natural-sounding artificial reverberation," *J. Audio Eng. Soc.*, vol. 10, no. 3, pp. 219–233, 1962.
- J. Stautner and M. Puckette, "Designing multichannel reverberators," *Computer Music J.*, vol. 6, no. 1, pp. 52–65, 1982.
- U. Zlzer, Ed., *DAFX: Digital audio effects.* John Wiley & Sons, LTD, 2002, ch. 3.

Microsoft Research

Outline

- Introduction
- Sound source localization with compact mic arrays
- Spatial sound and multi-channel echo cancellation
  - Head phone based spatial sound
  - Loudspeaker spatial sound
  - Multi-channel echo cancellation
- Video-based tracking
- Other video processing techniques
- Conclusions and future work

Microsoft Research

Motivation

- People do not like to wear headphones in a remote meeting
- Need a loud speaker solution
  - Challenges
    - User's head moves around
    - Requires multi-channel echo cancellation

Microsoft Research

Amplitude Panning

- Well known method (Pullki 97)
- Delay and Gain Modulation
  - Delay

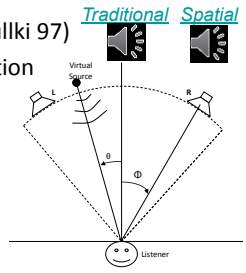
    $\Delta_R = D - D\cos(\lambda(\Phi - \Theta))$

    $\Delta_L = D - D\cos(\lambda(\Phi + \Theta))$

    $D=0.45ms \quad 1 \le \lambda \le \pi/(2\Phi)$
  - Gain

    $G_R = \cos(\lambda(\Phi - \Theta)/2)$

    $G_L = \cos(\lambda(\Phi + \Theta)/2)$

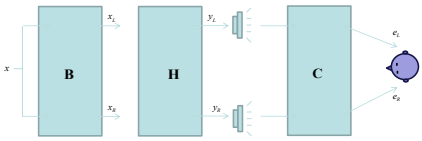


- Only create spatial sound between the loudspeakers

Microsoft Research

HRTF Based Schemes

- HRTF based method through crosstalk cancellation
  - Binaural sound reproduction over **loudspeaker** with crosstalk canceller

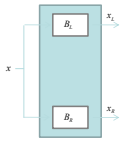


- Organization
  - B : Binaural synthesis
  - H : crosstalk canceller : inversion of C
  - C : transmission path or acoustic channel

Microsoft Research

Binaural Synthesis

- Locating sound images around the listener by filtering the monaural input signal with the HRTF for a given angle of incidence.



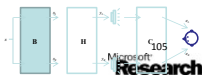
$$\mathbf{x} = \begin{bmatrix} x_L \\ x_R \end{bmatrix} = \begin{bmatrix} B_L \\ B_R \end{bmatrix} x = \mathbf{B}x$$

x : input monaural signal

x_L and x_R : binaural signals for left and right

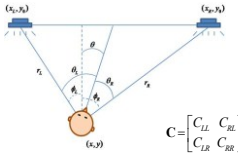
B_L and B_R : HRTF for left and right ear

- Convolution with pre-measured HRTF



Microsoft Research

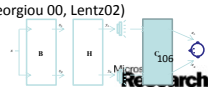
The Dynamic Acoustic Channel



$$\begin{cases} \theta & \text{: head rotation} \\ \theta_L, \theta_R & \text{: azimuths to loudspeakers} \\ (x, y) & \text{: head position} \end{cases}$$

$$\mathbf{C} = \begin{bmatrix} C_{LL} & C_{RL} \\ C_{LR} & C_{RR} \end{bmatrix} \Rightarrow \mathbf{C}_d = \begin{bmatrix} \frac{1}{r_L} z^{-d_L} C_{LL} & \frac{1}{r_R} z^{-d_R} C_{RL} \\ \frac{1}{r_L} z^{-d_L} C_{LR} & \frac{1}{r_R} z^{-d_R} C_{RR} \end{bmatrix}$$

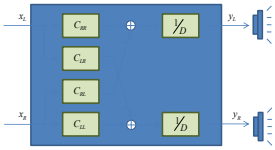
- Adjusting azimuth for HRTFs according to tracking information (Gardner 97)
- Involving time delay and level attenuation
- Tracking using electromagnetic trackers (Georgiou 00, Lentz02)
- Tracking using webcams (Kyriakakis 98, López 99, Lentz 06, Kim 08)



Microsoft Research

Dynamic Crosstalk Cancellation

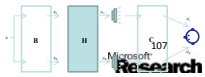
- Inverse the acoustic transfer matrix **C**



$$\mathbf{H} = \mathbf{C}^{-1} = \begin{bmatrix} C_{LL} & C_{RL} \\ C_{LR} & C_{RR} \end{bmatrix}^{-1} = \begin{bmatrix} C_{RR} & -C_{RL} \\ -C_{LR} & C_{LL} \end{bmatrix} \frac{1}{D},$$

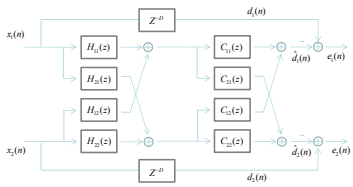
$$D = C_{LL} C_{RR} - C_{LR} C_{RL}.$$

- Direct inverse is unstable
  - **H** is adaptively obtained by least mean square (LMS) method



Microsoft Research

Crosstalk Cancellation with LMS

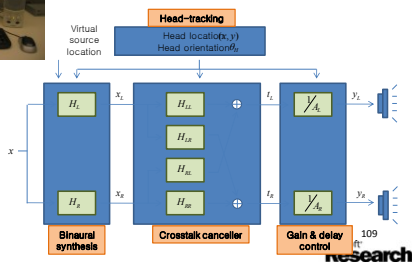


$$\begin{cases} C_{ml}(z) & \text{: Acoustic channels} \\ H_{nk}(z) & \text{: CTX filters} \end{cases}$$

108

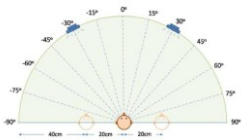
Microsoft Research

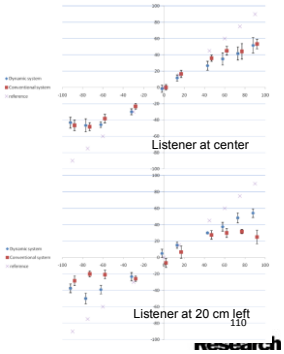
Our Complete System



Listening Test

- Pink noise sound source
- Subjects asked to identify virtual sound direction, which is compared with the ground truth

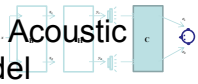




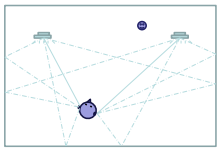
Adding a Room Model

- Crosstalk canceller suffers from room reverberation. HRTFs only represent the *direct path* of acoustic channel.
- Room model?
  - Simplified room model often sufficient (e.g., SSL)
  - Acoustic measurement possible;

Room Model Based Acoustic Channel Model



- New acoustic channel model
  - Summation of the responses for the individual rays from 6 walls as well as direct path from real loudspeaker.



$$C_{nm} = \sum_{k=0}^{N_r} \frac{\beta_k}{r_{mk}} z^{-\Delta_{mk}} C_{nm}(\angle_k)$$

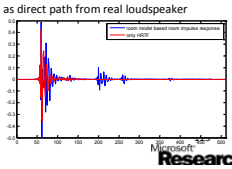
β_k : reflection coeff. for k-th image of m-speaker
r_k : distance from k-th image of m-speaker
$-\Delta_k$: delay from k-th image of m-speaker
$C_{nm}(\angle_k)$: HRTF from k-th image of m-speaker to n-ear

$$\mathbf{C} = \begin{bmatrix} C_{LL} & C_{RL} \\ C_{LR} & C_{RR} \end{bmatrix} \Rightarrow \mathbf{C} = \begin{bmatrix} \sum_{k=0}^{N_r} \frac{1}{r_{LL}} z^{-\Delta_{LL}} C_{LL}(\angle_k) & \sum_{k=0}^{N_r} \frac{1}{r_{RL}} z^{-\Delta_{RL}} C_{RL}(\angle_k) \\ \sum_{k=0}^{N_r} \frac{1}{r_{LR}} z^{-\Delta_{LR}} C_{LR}(\angle_k) & \sum_{k=0}^{N_r} \frac{1}{r_{RR}} z^{-\Delta_{RR}} C_{RR}(\angle_k) \end{bmatrix}$$

Microsoft Research

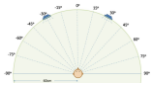
Room Model Based Crosstalk Canceller

- Procedure
  - Room model estimation
    - Obtaining reflection coefficients and configuration of walls
  - Calculating each ray individually
    - Taking account of attenuation by reflection and distance of the path
  - Determining HRTFs
    - Corresponding to direction of the incoming ray.
  - Summation of the responses
    - For the individual rays from 6 walls as well as direct path from real loudspeaker
- Example of transmission channel based on room model
  - C_{RR}
  - Including first reflections from 6 walls

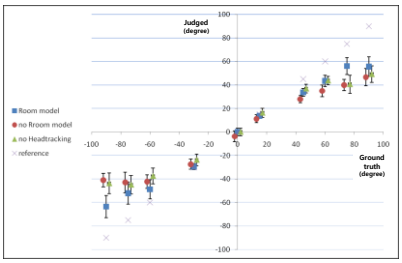


Microsoft Research

User Study

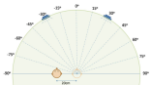


- Listener at center in reverberant room

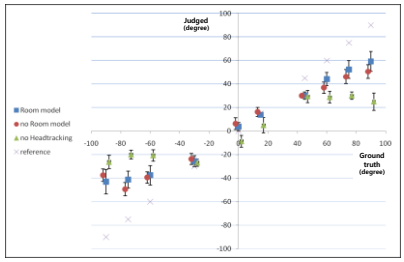


Microsoft Research

User Study

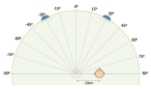


- Listener at 20cm left in reverberant room

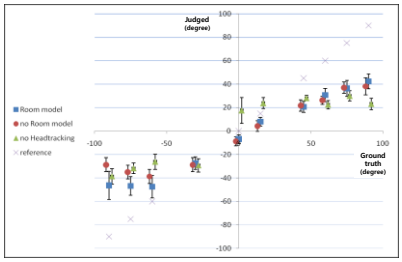


Microsoft Research

User Study



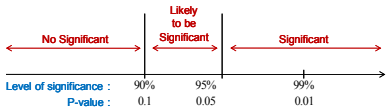
- Listener at 20cm right in reverberant room



Microsoft Research

User Study

- Student t-test
  - Assessing whether the means of two groups are statistically different from each other
  - Comparing absolute value of difference between ground-truth and judged azimuth
    - $|reference_i - Judged_{i,n}|$, where i and n are azimuth index and subject index



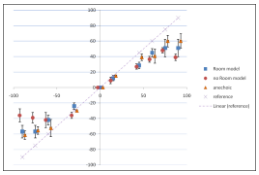
- T-test Results
  - Room-model vs. no Room-model $p = 0.0007893 < 0.05$
    - ➔ There is significant difference between the two method
  - Head-tracking vs. no Head-tracking $p = 0.0000000483052 < 0.05$
    - ➔ There is significant difference between the two method.

Microsoft Research

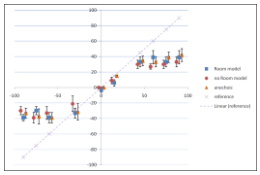
User Study

- Results comparison between 5 good listeners and 5 bad listeners (user at center, no head tracking)

Best 5



worst 5



Differences for virtual sources located outside of loudspeaker
Differences for 3 types
Best 5 : Anechoic > Room model > no Room model
Worst 5 : Anechoic = Room model = no Room model

References

- W. Gardner, "3-D audio using loudspeakers," Ph.D. thesis, Massachusets Institute of Technology, 1997.
- P. Georgiou, A. Mouchtaris, I. Roumeliotis, and C. Kyriakakis, "Immersive Sound Rendering Using Laser-Based Tracking", *Proc. 109th Convention of the Audio Eng. Soc.*, Paper 5227, 2000.
- S. Kim, D. Kong, and S. Jang, "Adaptive Virtual Surround Sound Rendering System for an Arbitrary Listening Position," *J. Audio Eng. Soc.*, Vol. 56, No. 4, 2008.
- C. Kyriakakis and T. Holman, "Video-based head tracking for improvements in multichannel loudspeaker audio," *105th Audio Engineering Society*, San Francisco, CA, 1998.
- T. Lentz, G. Behler, "Dynamic Crosstalk Cancellation for Binaural Synthesis in Virtual Reality Environments," *J. Audio Eng. Soc.*, Vol. 54, Issue 4, pp. 283-294, 2006.
- T. Lentz, O. Schmitz, "Realisation of an adaptive cross-talk cancellation system for a moving listener," *21st AES Conference on Architectural Acoustics and Sound Reinforcement*, 2002.
- J. Lopez and A. Gonzalez, "3-D Audio With Dynamic Tracking for Multimedia Environtments," 2nd COST-G6Workshop on Digital Audio Effects, 1999.
- V. Pullki, "Virtual Sound Source Positioning Using Vector Base Amplitude Panning," *J. Audio Eng. Soc.*, Vol. 45, pp. 456-466, 1997.
- M. Song, C. Zhang, D. Florencio and H.-G. Kang, "Personal 3D Audio System with Loudspeakers," *International Workshop on Hot Topics in 3D*, in conjunction with ICME, 2010.
- M. Song, C. Zhang, D. Florencio and H.-G. Kang, "Enhanced Binaural Loudspeaker Audio System with Room Modeling," in *Proc. of MMSP*, 2010.

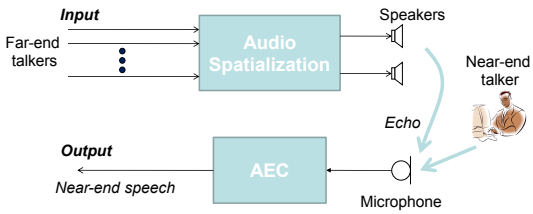
Outline

- Introduction
- Sound source localization with compact mic arrays
- Spatial sound and multi-channel echo cancellation
  - Head phone based spatial sound
  - Loudspeaker spatial sound
  - Multi-channel echo cancellation
- Video-based tracking
- Other video processing techniques
- Conclusions and future work

Challenges and Previous Work

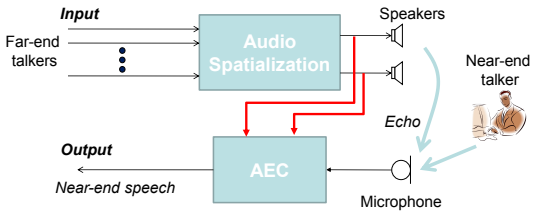
- Stereo ACE suffers from the mis-convergence problem (Sonfhi 95)
- In order to de-correlate the speaker signals
  - Adding nonlinearity (Benesty 98)
  - Adding noise (Sonfhi95, Gilloire 98)
- Adaptive filtering
  - NLMS (Yensen 01)
  - Kalman filtering (Enzner 06)
  - Constrained Kalman filtering (Zhang 08)

Multichannel AEC



- **Question:** *Which reference signals to use?*

Approach 1: Use Speaker Signals



- **Our solution:** Kalman Filtering (general RLS)
- Possible problem: Correlation between speaker signals

Problem Statement: Speaker based

- Speaker Signals:
- Speaker's room response: *L*-tap filter
- Microphone input: Echo

where

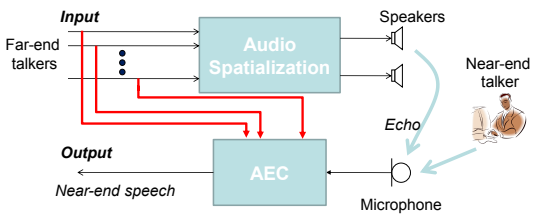
Multichannel AEC: Speaker-based

- Kalman Filter (Generalized RLS)

where is the aging factor and changes over time

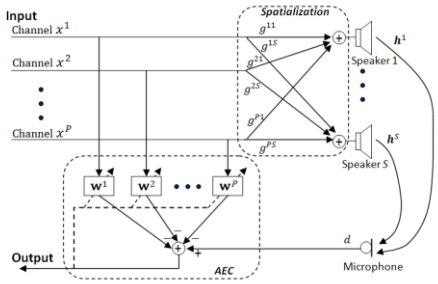
- When , Kalman Filter ⇨ Traditional RLS
- *We model ambient noise.*

Approach 2: Use Far-End Channels



- Cancel each individual far-end speech
- **Our solution:** Constrained Kalman filtering

Multichannel AEC: CKF Diagram



Microsoft Research

Problem Statement

- Remote channels: $\{X^i \mid i = 1, \ldots, P\}$
- Spatialization on S speakers: $Y^s = \sum_{i=1}^{P} G^{is} X^i$
- Speaker's room response: L-tap filter

$$\boldsymbol{H}_t^s = [H_t^s, H_{t-1}^s, \ldots, H_{t-L-1}^s]^T$$

- Microphone input: Echo

$$D_t = \sum_{i=1}^{P} \sum_{s=1}^{S} G^{is} \left(H_t^s X_t^i + H_{t-1}^s X_{t-1}^i + \cdots + H_{t-L-1}^s X_{t-L-1}^i \right) = \sum_{i=1}^{P} \sum_{s=1}^{S} G^{is} \boldsymbol{H}_t^{s\,T} \boldsymbol{X}_t^i$$

Microsoft Research

Problem Statement (cont'd)

- Determine the echo cancellers:
  - one per remote channel i: L-tap filter

$$\boldsymbol{W}_t^i = [W_t^i, W_{t-1}^i, \ldots, W_{t-L-1}^i]^T$$

such that echo is cancelled, i.e.,

$$D_t - \sum_{i=1}^{P} \boldsymbol{W}_t^{i\,T} \boldsymbol{X}_t^i = 0$$

- Constraint: **W_t^is are not mutually independent**

$$D_t = \sum_{i=1}^{P} \sum_{s=1}^{S} G^{is} \boldsymbol{H}_t^{s\,T} \boldsymbol{X}_t^i \quad \Rightarrow$$

Microsoft Research

Constrained Kalman Filtering

- State Vector: Echo cancellers + Speaker RIR filters
  $$\boldsymbol{s}_t = \left[\boldsymbol{w}_t^{1T}, \ldots, \boldsymbol{w}_t^{pT}, \boldsymbol{h}_t^{1T}, \ldots, \boldsymbol{h}_t^{sT}\right]^T$$
- System equation:
- Observation equation: $D_t = \boldsymbol{A}_t^T \boldsymbol{s}_t + v_t$ with
- Constraint: $\boldsymbol{C}\boldsymbol{s}_t = \boldsymbol{0}$ with $\boldsymbol{C} = \begin{bmatrix} -\boldsymbol{1} & & & G^{11}\boldsymbol{1} & \cdots & G^{1S}\boldsymbol{1} \\ & \ddots & & \vdots & \ddots & \vdots \\ & & -\boldsymbol{1} & G^{P1}\boldsymbol{1} & \cdots & G^{PS}\boldsymbol{1} \end{bmatrix}$
- New observation equation: observation + constraint
  with $\boldsymbol{Y}_t = [D_t, 0, \ldots, 0]^T$ $\boldsymbol{B}_t = \begin{bmatrix} \boldsymbol{A}_t^T \\ \boldsymbol{C} \end{bmatrix}$ $\boldsymbol{v}_t = \begin{bmatrix} v_t \\ \boldsymbol{u}_t \end{bmatrix}$

Microsoft Research

Constrained Kalman Filtering (cont'd)

- Assumptions
  $E[\boldsymbol{n}_t] = \boldsymbol{0}$ $E[\boldsymbol{n}_t \boldsymbol{n}_t^T] = \boldsymbol{Q}_t$
  $E[\boldsymbol{v}_t] = \boldsymbol{0}$ $E[\boldsymbol{v}_t \boldsymbol{v}_t^T] = \boldsymbol{R}_t = \begin{bmatrix} \sigma_t^2 & \boldsymbol{0}^T \\ \boldsymbol{0} & \end{bmatrix}$
  Tuning parameter to control how hard the constraint be satisfied
- Equations
  $$\boldsymbol{S}_t^- = \boldsymbol{S}_{t-1}$$
  $$\boldsymbol{P}_t^- = \boldsymbol{P}_{t-1} + \boldsymbol{Q}_t$$
  $$\boldsymbol{K}_t = \boldsymbol{P}_t^- \boldsymbol{B}_t^H (\boldsymbol{B}_t \boldsymbol{P}_t^- \boldsymbol{B}_t^H + \boldsymbol{R}_t)^{-1}$$
  $$\boldsymbol{S}_t = \boldsymbol{S}_t^- + \boldsymbol{K}_t (\boldsymbol{Y}_t - \boldsymbol{B}_t \boldsymbol{S}_t^-)$$
  $$\boldsymbol{P}_t = (\mathbf{I} - \boldsymbol{K}_t \boldsymbol{B}_t) \boldsymbol{P}_t^-$$

Microsoft Research

Benefits of Constrained KF

- The constraint is taken care of automatically, and can be imposed *with varying degrees*.
- All channels are taken into account simultaneously.
  →Overlapping far-end talking is not an issue
- The AEC for each channel is updated continuously because of the constraint, even if it is inactive.
  →AEC's are always up to date
- Ambient noise can be time varying.
  → Use a separate noise tracker

Microsoft Research

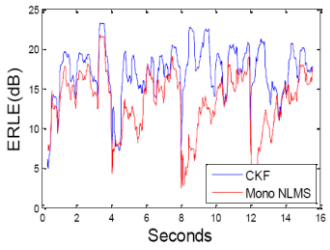
Comparison with Prior Art

- T.N. Yensen, R.A. Goubran, and I. Lambadaris, "Synthetic Stereo Acoustic Echo Cancellation Structure for Multiple Participant VoIP Conferences", IEEE Transactions on Speech and Audio Processing, Vol. 9, No. 2, pp. 168-174, Feb. 2001.
- Same: One canceller per remote channel
- Differences:
  - Constrained vs. independent cancellers
    - Additional canceller is initiated before being active
    - A canceller is updated even if it is not active
  - Frequency vs. time domain
  - KF (RLS) vs. NLMS

133 Microsoft Research

Experimental Results

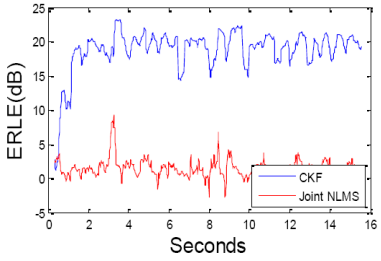
- Simulation setup:
  - 4 remote talkers at [-30˚, 30˚, 0˚, -45˚]
  - Each talks for 4s
  - Noise: -20dB
  - Fixed RIR
- Comparison
  - Constrained KF
  - Multiple mono NLMS



134 Microsoft Research

Experiment: Overlapping Talkers

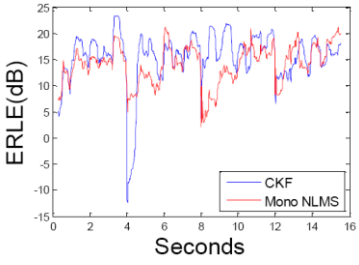
- Two simultaneous remote talkers



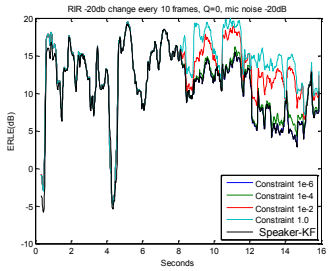
135 Microsoft Research

Experiment: Changing RIR

- -30dB change in RIR every 0.5 seconds

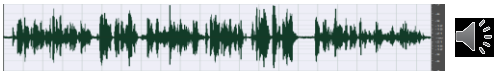


Different Constraints

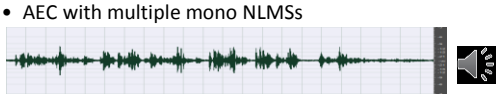


Experiment with real data

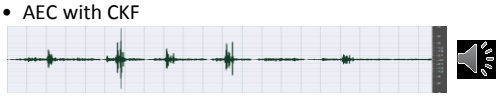
- Original recording with near-end talker



- AEC with multiple mono NLMSs

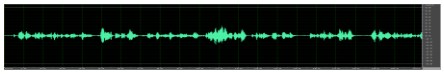


- AEC with CKF

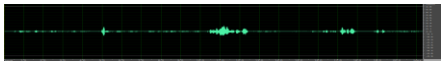


Experiment with real data (2)

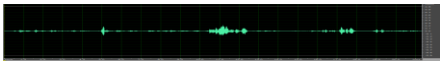
- Original Mic Input with Near-end Speech



- AEC with Far-end Channel based CKF

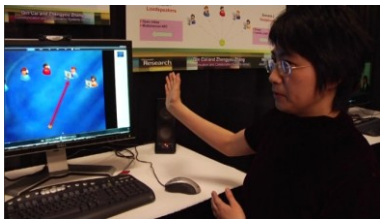


- AEC with Speaker Based KF



Microsoft Research

Demo Video



Microsoft Research

References

- J. Baldis, "Effects of spatial audio on memory, compression, and preference during desktop conferences," In *Proc. of the CHI Conference on Human Factors in Computing Systems*, 2001.
- J. Benesty, D. Morgan, and M. Sonfhi, "A better understanding and an improved solution to the problem of stereophonic acoustic echo cancellation," *IEEE Trans. on Speech and Audio Processing*, vol. 6, pp. 156-165, 1998.
- G. Enzner, and P. Vary, "Frequency-domain adaptive Kalman filter for acoustic echo control in hands-free telephones," *Signal Processing*, vol. 86, pp. 1140-1156, Oct, 2006.
- A. Gilloire, and V. Turbin, "Using auditory properties to improve the behavior of stereophonic acoustic cancellers," in *Proc. of ICASSP*, 1998.
- M. Sonfhi, D. Morgan, and J. Hall, "Stereophonic acoustic echo cancelation – An overview of the fundamental problem," *IEEE Speech Processing Letter*, vol. 2, pp. 148-151, 1995.
- T. N. Yensen, R. A. Goubran, and I. Lambadaris, "Synthetic Stereo Acoustic Echo Cancellation Structure for Multiple Participant VoIP Conference," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 2, pp. 168-174, Feb. 2001.
- Z. Zhang, Q. Cai and J. Stokes, "Multichannel Acoustic Echo Cancellation in Multiparty Spatial Audio Conferencing with Constrained Kalman Filtering," in *Proc. Of IWAENC*, 2008.

Microsoft Research

Outline

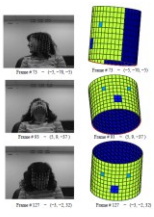
- Introduction
- Sound source localization with compact mic arrays
- Spatial sound and multi-channel echo cancellation
- Video-based tracking
  - Multi-camera head pose tracking
  - Depth camera based facial expression tracking
  - Depth camera based Skeleton tracking
- Other video processing techniques
- Conclusions and future work

Motivation

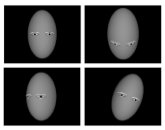
- Head Pose is very important for tele-conferencing
  - Face Relighting
  - Gaze Correction
  - Adaptive Displays
  - Video driven avatar animation, etc.

Single Camera Head Pose Tracking

- Model-based (see Lepetit 05 for a survey)



Cylinder (Aggarwal 05)



Ellipsoids (Moon 01)

3D face model (Wang 06, Vacchetti 04)

Motivation

- Single Camera Tracking has limitation on
  - Narrow Range of Tracked Poses
  - Occlusions
- Multiple Camera Tracking
  - Existing Setup in Video Conferencing
  - Multiple Views of the Face

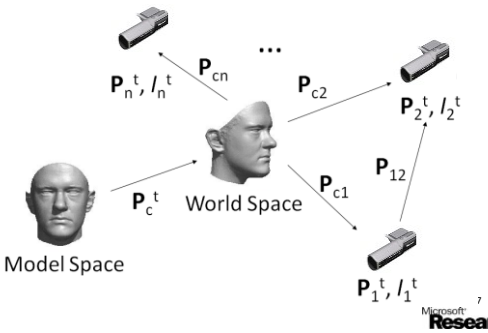
Microsoft Research

Challenges

- Our Multi-Cam Tracking System
  - No Pre-calibration of Camera Placements
  - Real Time Processing
  - Drift Free
  - Wider Range of Tracked Poses
  - Continuous Tracking w/ Occlusion
  - Recover after Tracking is Lost

Microsoft Research

Model-based Approach



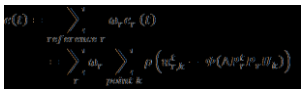
Microsoft Research

Tracking by References

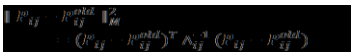
Microsoft Research

Generic Tracking Scheme

- Suitable for both single and multiple cameras
- References span on both Temporal and Spatial domains

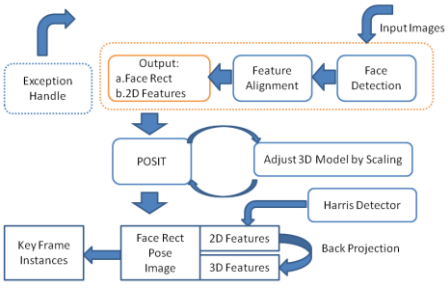


- Camera Relative Geometry converges over time



Microsoft Research

Tracking Initialization



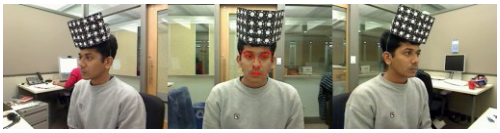
Microsoft Research

Tracking Algorithm

- **Input: Image , previous frames , & key frames**
- **Output: Current Pose .**
  - Select Pose of Reference frames on each camera
  - Initialize Relative Pose between Ref Camera *c* and other camera *i* .
  - For each camera *I*, do
    - *For each pose of ref. frame , do*
    - *Find 2D feature matches between and , then back-project in to get 3D points in the model **U**.*
    - *Discard outliers by applying **RANSAC***
    - *end*
  - Apply POSIT for all inlier 3D-2D matches
  - End

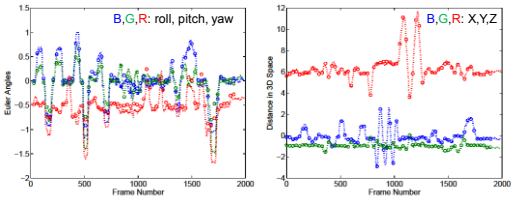
Microsoft Research

Results (3 Cameras)



Microsoft Research

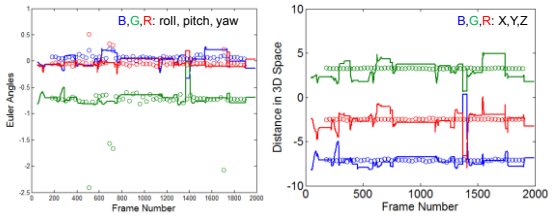
Comp. w/ Ground Truth I (No Occlusion)



….. Tracked Pose from our Algorithm.
o Ground Truth from Calibration.

Microsoft Research

Relative Pose Comp. w/ Ground Truth II



Results (Occlusion)

Fair Comparison (no Occlusion)

# of Cams	Median Angle Error (degree)	Median Translation Error	Mean Angle Error (degree)	Mean Translation Error
1	5.9754	0.4138	19.0132	0.5504
2	5.4878	0.3481	11.8039	0.4885
3	4.4671	0.3404	11.0915	0.4634
3 (w/ pre camera calibration)	4.0827	0.3132	11.0577	0.4540

References

- G. Aggarwal, A. Veeraraghavan, and R. Chellappa, "3D Facial Pose Tracking in Uncalibrated Videos," *Lecture Notes in Computer Science*, 3776:515, 2005.
- Q. Cai, A. Sankaranarayanan, Q. Zhang, Z. Zhang, and Z. Liu, "Real Time Head Pose Tracking from Multiple Cameras with a Generic Model," in *IEEE Workshop on Analysis and Modeling of Faces and Gestures*, 2010.
- V. Lepetit and P. Fua. "Monocular model-based 3d tracking of rigid objects: A survey," *Foundations and Trends in Computer Graphics and Vision*, 1(1):1–89, October 2005.
- H. Moon, R. Chellappa, and A. Rosenfeld, "3d object tracking using shape-encoded particle propagation," in *IEEE International Conference on Computer Vision*, 2001.
- L. Vacchetti, V. Lepetit, and P. Fua, "Stable Real-Time 3D Tracking Using Online and Offline Information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1385–1391, 2004.
- Q. Wang, W. Zhang, X. Tang, and H. Shum, "Real-Time Bayesian 3-D Pose Tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, 16(12):1533, 2006.

Microsoft Research

Outline

- Introduction
- Sound source localization with compact mic arrays
- Spatial sound and multi-channel echo cancellation
- Video-based tracking
  - Multi-camera head pose tracking
  - Depth camera based facial expression tracking
  - Depth camera based Skeleton tracking
- Other video processing techniques
- Conclusions and future work

Microsoft Research

Deformable Face Tracking

- Many applications
  - Human computer interaction
  - Performance-driven facial animation
  - Face recognition
- Challenging
  - Limited number of features on the face
  - Dozens of parameters to estimate

Microsoft Research

Related Works

- Video based fitting
  - Active appearance models (AAM) (Xiao04)
  - 3D morphable models (Blanz99)
  - 3D deformable model + active shape model (ASM) (Vogler07)
  - Feature based (Zhang08)
- 3D scans based fitting
  - Iterative closest point (ICP) + 3D deformable model (Zhang04, Wang04, Weise09, Weise11)

Microsoft Research

Commodity Depth Cameras

SwissRanger

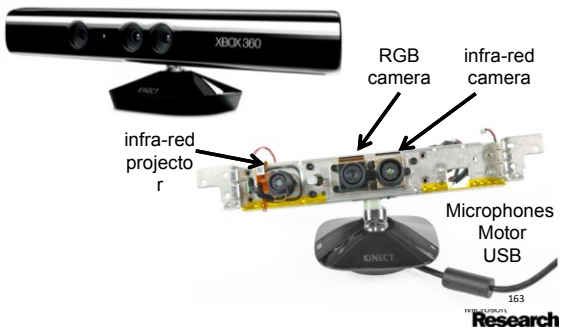
Canesta

ZCam

Kinect

Microsoft Research

Kinect Sensor



Microsoft Research

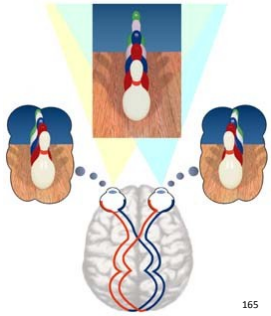
Non-commercial Windows software development kit

Access to deep Kinect system information

- Depth data
- Audio
- Direct control of the Kinect sensor
- System API

Human Stereo Vision

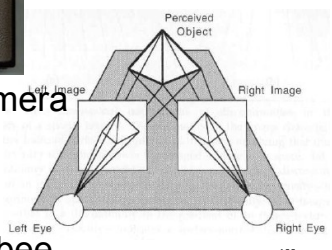
Difference in your two eyes gives you the ability to perceive your surrounding environment in **3 Dimensions**



Computer Stereo Vision

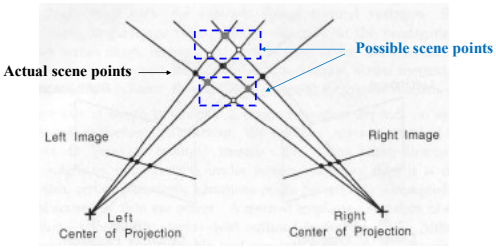
dak Stereo Camera

ntGrey Bumblebee

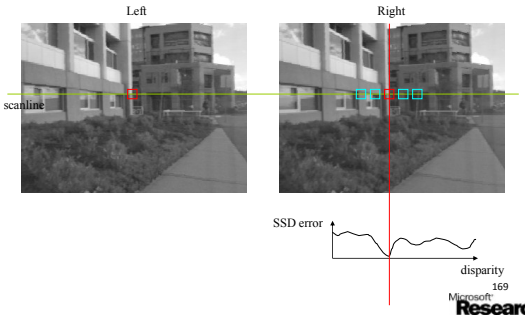


Stereo Images

Correspondence Problem

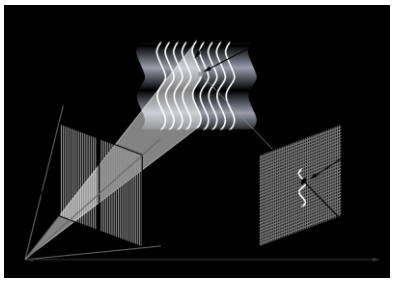


Correspondence by Correlation



How it works?

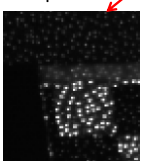
- **Structured light 3D scanner**



Microsoft Research

How it works? **Kinect Sensor**

- Modified structured light 3D scanner
  - IR projector
  - IR camera
  - Random pattern



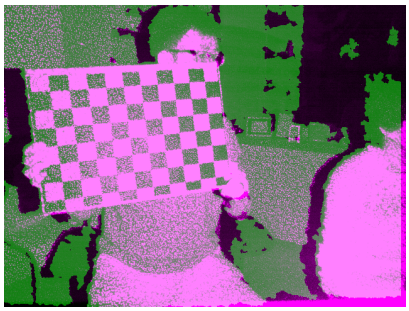
Microsoft Research

Matching & Depth Map

- Correlation

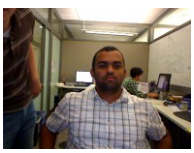
Microsoft Research

Overlay of Depth Map on IR Image



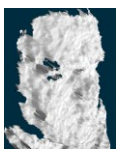
Microsoft Research

Captured Image



Texture

Depth



Depth enlarged

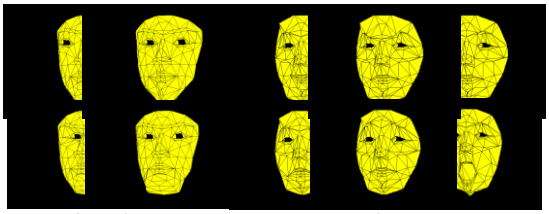
Microsoft Research

Challenge and Solution

- Depth information should help tracking
- Challenge: noisy depth input
- Solution:
  - Maximum likelihood solution with arbitrary noise covariance matrices
  - Regularization

Microsoft Research

Linear Deformable Model



Static deformations Action deformations

(**Artist rendered** linear deformable model)

$$\begin{bmatrix} q_1 \\ \vdots \\ q_K \end{bmatrix} = \begin{bmatrix} p_1 \\ \vdots \\ p_K \end{bmatrix} + \boldsymbol{A}\begin{bmatrix} r_1 \\ \vdots \\ r_K \end{bmatrix} + \boldsymbol{B}\begin{bmatrix} s_1 \\ \vdots \\ s_K \end{bmatrix}, \text{ where } \boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_1 \\ \vdots \\ \boldsymbol{A}_K \end{bmatrix}, \boldsymbol{B} = \begin{bmatrix} \boldsymbol{B}_1 \\ \vdots \\ \boldsymbol{B}_K \end{bmatrix}$$

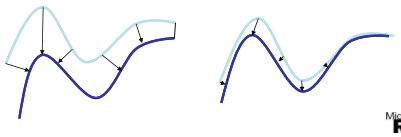
Microsoft Research

Maximum Likelihood DMF

- Formulation, $(\boldsymbol{q}_k, \boldsymbol{g}_k)$ correspondence pair:

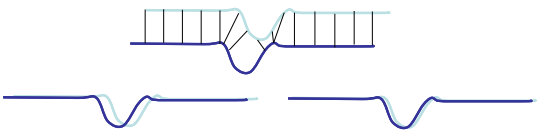
$$\boldsymbol{R}(\boldsymbol{p}_k + \boldsymbol{A}_k\boldsymbol{r} + \boldsymbol{B}_k\boldsymbol{s}) + \boldsymbol{t} = \boldsymbol{g}_k + \boldsymbol{x}_k$$
$$\boldsymbol{x}_k \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{x}_k})$$

- Iterative closest point
  - Assume closest points correspond
  - Compute transformation
  - Iterate until convergence



Microsoft Research

Point-to-Point and Point-to-Plane Distance



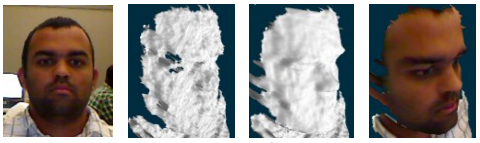
point-to-point (slow convergence) point-to-plane (fast convergence)

- In both cases, we have solution for arbitrary noise covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}_k}$
  - See paper for details

Microsoft Research

Model Initialization

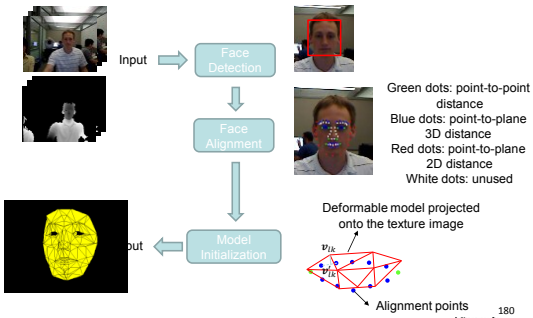
- Initialization
  - Assume multiple neutral face frames available
  - Action deformations set to zero
  - Jointly solve shape deformations and rotation/translation of each frame



single frame | 30 frames aligned and averaged | textured model

Microsoft Research

Model Initialization



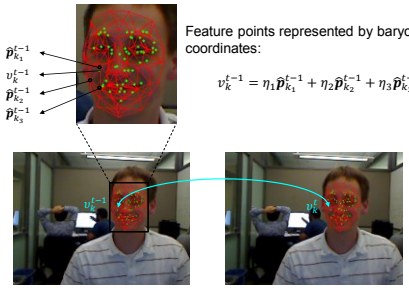
Microsoft Research

Face Tracking

- Tracking
  - Shape deformations fixed
  - Based on feature point correspondence
  - Solve for action deformation, rotation and translation
  - Regularization
    - l_2 norm constraining the difference between neighboring frames' action deformations
    - l_1 norm constraining the number of non-zero action deformation parameters

Microsoft Research

Face Tracking



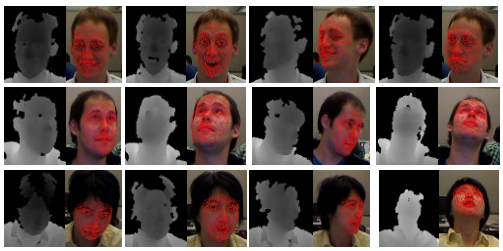
Feature points represented by barycentric coordinates:

$$v_k^{t-1} = \eta_1 \hat{p}_{k_1}^{t-1} + \eta_2 \hat{p}_{k_2}^{t-1} + \eta_3 \hat{p}_{k_3}^{t-1}$$

Texture image at time t-1 | Texture image at time t

Microsoft Research

Tracking Results



Top to bottom: Seq #1 (810 frames), Seq #2 (681 frames), Seq #3 (300 frames)

Video

Microsoft Research

Qualitative Results

Median tracking error in pixels

	ID+l_2	ID+l_1	ID+l_2+l_1	NM+l_2	NM+l_1	NM+l_2+l_1
Seq #1	3.56	2.88	2.78	2.85	2.69	2.66
Seq #2	4.48	3.78	3.71	4.30	3.64	3.55
Seq #3	3.98**L**	3.91	3.91	3.92**L**	3.91	3.50

ID: use identity covariance matrix for sensor noise
NM: use the proposed noise modeling scheme
l_2: quadratic constraint between successive frames
l_1: sparse constraint on the action transforms
L: lost tracking in the middle and never recover

Microsoft Research

References

- V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *Proc. of SIGGRAPH*, 1999.
- Q. Cai, D. Gallup, C. Zhang and Z. Zhang, "3D Deformable Face Tracking with a Commodity Depth Camera," in Proc. of ECCV, 2010.
- Y. Wang, X. Huang, C. Lee, S. Zhang, Z. Li, D. Samaras, D. Metaxas, A. Elgammal, P. Huang, "High resolution acquisition, learning and transfer of dynamic 3-D facial expressions," in *Proc. of EUROGRAPHICS*, 2004.
- T. Weise, S. Bouaziz, H. Li, M. Pauly, "Realtime performance-based facial animation," *Proc. of SIGGRAPH*, 2011.
- T. Weise, H. Li, L. Gool, M. Pauly: "Face/off: Live facial puppetry," in *Symposium on Computer Animation*, 2009.
- C. Vogler, Z. Li, A. Kanaujia, S. Goldenstein, D. Metaxas, "The best of both worlds: Combining 3d deformable models with active shape models," in *Proc. of ICCV*, 2007.
- J. Xiao, S. Baker, I. Matthews, T. Kanade, "Real-time combined 2d+3d active appearance models," in *Proc. of CVPR*, 2004.
- L. Zhang, N. Snavely, B. Curless, S. Seitz, "Spacetime faces: high-resolution capture for modeling and animation," in *Proc. of SIGGRAPH*, 2004.
- W. Zhang, Q. Wang, X. Tang, "Real time feature based 3-D deformable face tracking," in *Proc. of ECCV*, 2008.

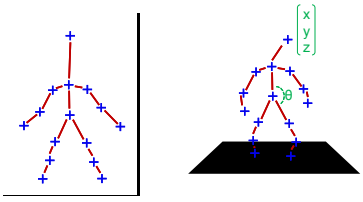
Microsoft Research

Outline

- Introduction
- Sound source localization with compact mic arrays
- Spatial sound and multi-channel echo cancellation
- Video-based tracking
  - Multi-camera head pose tracking
  - Depth camera based facial expression tracking
  - Depth camera based Skeleton tracking
- Other video processing techniques
- Conclusions and future work

Microsoft Research

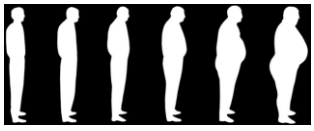
Human Pose Estimation



Kinect tracks 20 body joints in real time.

Microsoft Research

Why Is It Hard?

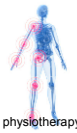



Microsoft Research

Why Do We Want It?

interaction

analysis




And applications to be determined…

Microsoft Research

Motion Capture

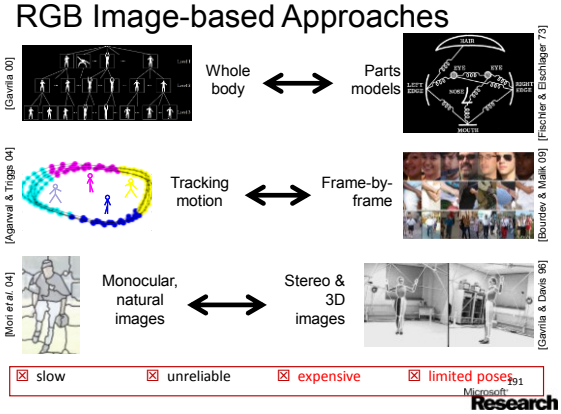

[Vicon]


[Xsens]

- ☑ very accurate
- ☑ high frame rate
- ☒ suit / sensors
- ☒ expensive
- ☒ large space
- ☒ calibration

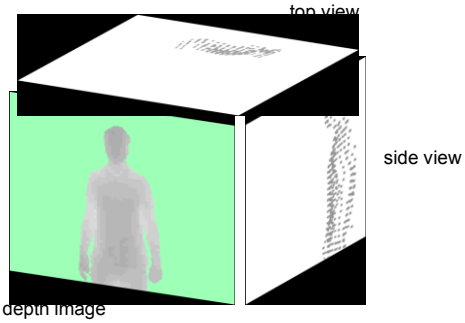
Microsoft Research

RGB Image-based Approaches



Microsoft Research

Depth Cameras



Microsoft Research

RGB vs. Depth for Pose Estimation

RGB

- ☒ Only works well lit
- ☒ Background clutter
- ☒ Scale unknown
- ☒ Clothing, skin colour

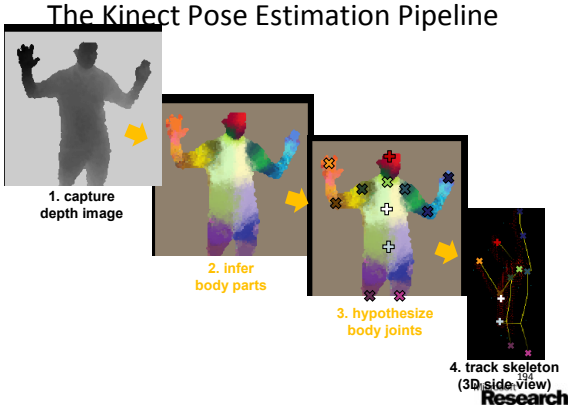
much easier with depth!

DEPTH

- ☑ Works in low light
- ☑ Person 'pops' out from bg
- ☑ Scale known
- ☑ Uniform texture
  - ☑ easy to simulate
- ☒ Shadows, missing pixels

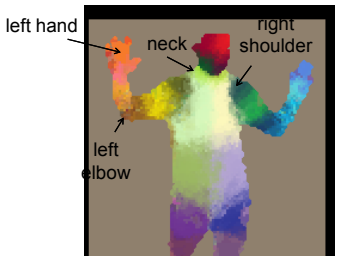
Microsoft Research

The Kinect Pose Estimation Pipeline



Microsoft Research

Body Part Recognition



("left" = player left with camera acting as mirror)

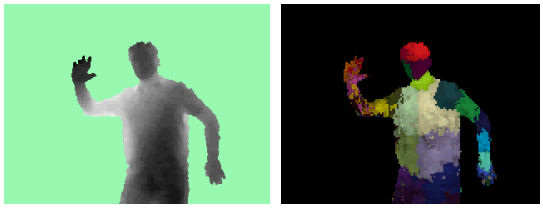
Microsoft Research

Body Part Recognition

- No temporal information
  - Frame-by-frame
- Local pose estimate of parts
  - Independently treat
    - Each pixel & each body joint
  - Reduces training data and computation time
- Very fast
  - Depth image features
  - Decision forest classifier

Microsoft Research

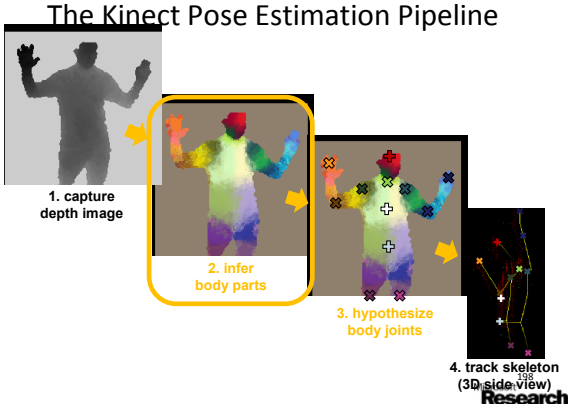
Body Part Recognition



input depth image | inferred body parts

Microsoft Research

The Kinect Pose Estimation Pipeline



Microsoft Research

Classifying Pixels

- Compute $\mathrm{P}(c_i \mid w_i)$
  - pixels $i = (x, y)$
  - body part c_i
  - image window w_i

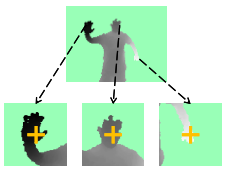


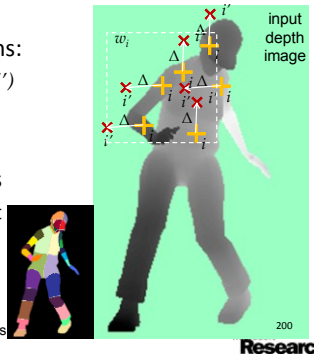
image windows move with classifier

- Discriminative approach
- Learn classifier $\mathrm{P}(c_i \mid w_i)$ from training data

Microsoft Research

Fast Depth Image Features

- Depth comparisons:
  - $f(i\,;\,\Delta) = d(i) - d(i')$ where $i' = i + \Delta$
- Background pixels
  - d = large constant



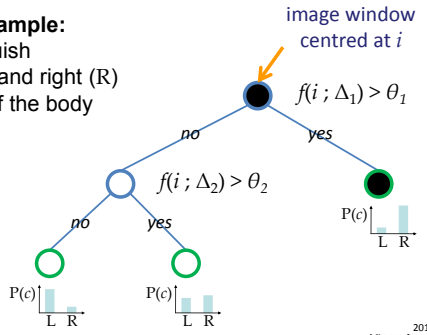
desired
body parts

200

Microsoft Research

Decision Tree Classification

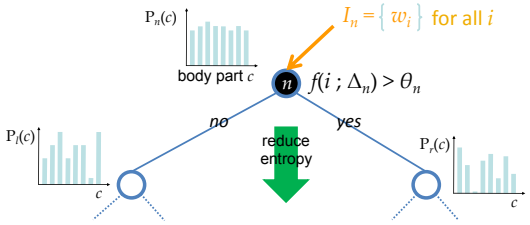
Toy example:
distinguish
left (L) and right (R)
sides of the body



201

Microsoft Research

Training Decision Trees (Breiman 84)



Take (Δ, θ) that maximises information gain:

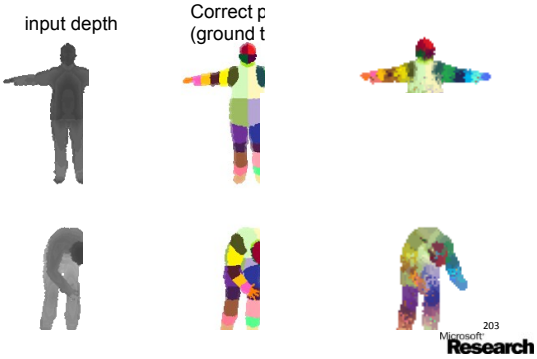
$$\Delta E = -\frac{|I_l|}{|I_n|}E(I_l) - \frac{|I_r|}{|I_n|}E(I_r)$$

Goal: drive entropy at leaf nodes to zero

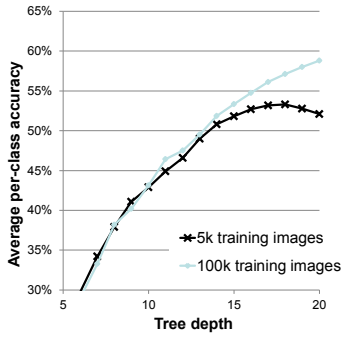
202

Microsoft Research

Depth of Trees



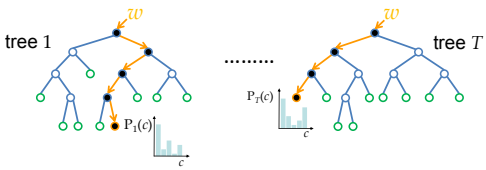
Depth of Trees



Decision Forests

(Amit 97, Breiman 01, Geurts 06)

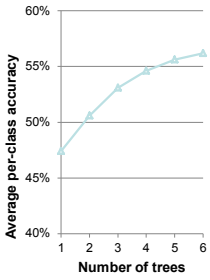
- Single trees tend to over-fit
- Train forest – ensemble of trees:



  - different random subset of images
  - average tree posteriors

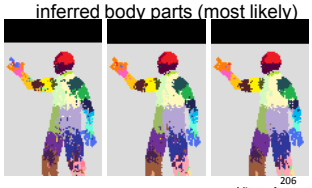
$$P(c|w) = \sum_{t=1}^{T} P_t(c|w)$$

Number of Trees



th

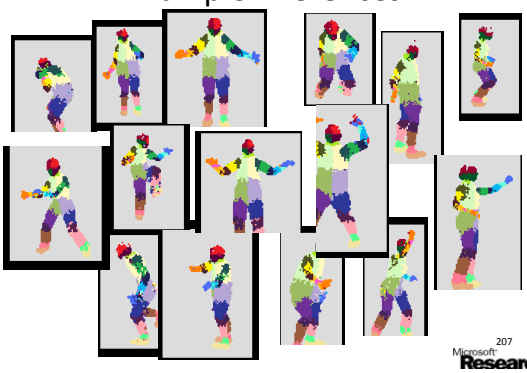
inferred body parts (most likely)



206

Microsoft Research

Example Inferences

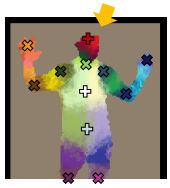


207

Microsoft Research

Body Parts to Joint Hypotheses

- Depth image & probability mass
- Localize body parts in 3D
  - global centroid of prob. mass
  - local modes of density (mean shift)
- Map body parts to skeletal joints
  - many parts map directly to joints



3. hypothesize body joints

208

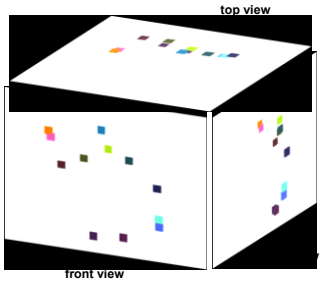
Microsoft Research

3D Joint Hypotheses

NB No tracking or smoothing!

input depth image

inferred body parts & overlaid joint hypotheses

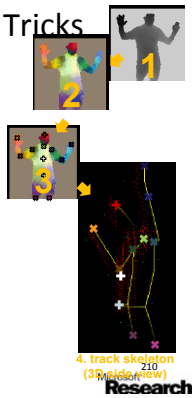


3D joint hypotheses

Microsoft Research

… and Some Other Tricks

- Exploit
  - 3D joint hypotheses
  - kinematic constraints
  - temporal coherence
- Predict
  - full skeleton
  - invisible joints
  - multi-player



Microsoft Research

References

- Y. Amit and D. Geman, "Shape quantization and recognition with randomized trees," *Neural Computation*, pp. 1545–1588, Vol. 9, No. 7, 1997.
- L. Breiman, "Random forests," *Machine Learning*, pp. 5–32, Vol. 45, No. 1, 2001.
- L. Breiman, J. Friedman, C. Stone and R. Olshen, *Classification and Regression Trees*, Chapman and Hall/CRC, 1984.
- P. Geurts, D. Ernst and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, pp. 3-42, Vol. 63, No. 1, 2006.
- J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, A. Blake, "Real-time human pose recognition in parts from single depth images," *Proc. of CVPR*, 2011.

Microsoft Research

Outline

- Introduction
- Sound source localization with compact mic arrays
- Spatial sound and multi-channel echo cancellation
- Video-based tracking
- Other video processing techniques
  - Learning-based perceptual image quality enhancement
  - Monocular 3D and Foreground/background segmentation
  - Joint tracking and multiview video compression
- Conclusions and future work

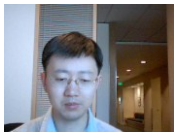
Motivation

- Two factors affecting image quality during video conferencing:
  - Camera device
  - Lighting condition
- Observation
  - Camera devices are getting better
  - Lighting conditions will stay the same
- Goal:
  - Improving virtual lighting condition

Previous work

- Brightness and contrast improvement
  - Hardware:
    - Automatic gain control
    - Face tracking -> gain control parameters
  - Image processing techniques (exposure correction):
    - Bhukhanwala 94, Messina 03, Saitoh 99, Shi 04

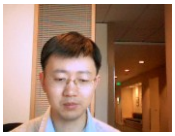
Beyond exposure



Original

Good exposure

But pale, unpleasant



Enhanced

Warm, appealing

Microsoft Research

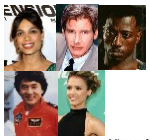
Color tone

- Stage lighting design
  - Color scheme is critical
    - The perceived look of the host / actors
    - Mood of the stage
- How does a color affect the look of a face
  - Subjective
  - No mathematical formula
  - Learning from experts

Microsoft Research

Learning-based color tone mapping

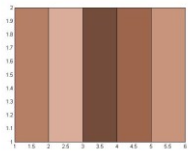
- Idea:
  - Learning color statistics from images taken by professional photographers (in good lighting conditions)
  - Color tone mapping for an input image
- Data collection:
  - 400 celebrity images



Microsoft Research

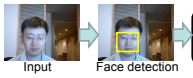
Training

- Mixtures of Gaussians
  - Feature:
    - mean and variance on the face region
      $v_i = (m_{i,r}, m_{i,g}, m_{i,b}, \sigma_{i,r}, \sigma_{i,g}, \sigma_{i,b}), i = 1,\ldots,400$
    - 5 mixture components $(\mu_j, \Sigma_j), j = 1,\ldots,5$



Microsoft Research

Color tone mapping



Source mean/std vector
$v = (m_1, m_2, m_3, \sigma_1, \sigma_2, \sigma_3)$

Target mean/std vector= weighted sum of cluster centers
$\bar{v} = (\bar{m}_1, \bar{m}_2, \bar{m}_3, \bar{\sigma}_1, \bar{\sigma}_2, \bar{\sigma}_3) = \sum_{j=1}^{5} w_j(v) * \mu_j$

$$w_j = \frac{1/D_j(v)}{\sum_{l=1}^{5} 1/D_l(v)}$$

$$D_j(v) = \sqrt{(v-\mu_j)^T \Sigma_j^{-1} (v-\mu_j)} \Rightarrow \text{Mahalanobis distance}$$

Microsoft Research

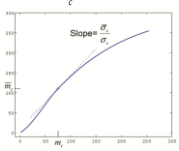
Color tone mapping function

$y = f_c(x)$: Color tone mapping function for channel c

Piecewise cubic spline with constraints:

$$f_c(m_c) = \bar{m}_c \quad f_c(0) = 0 \quad f_c(255) = 255$$

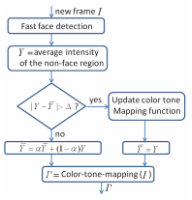
$$f_c'(m_c) = \frac{\bar{\sigma}_c}{\sigma_c} \quad f_c'(0) = 0.5\frac{\bar{m}_c}{m_c} \quad f_c'(255) = 0.5\frac{255-\bar{m}_c}{255-m_c}$$



Microsoft Research

Application to video sequence

- Image intensities change over time:
  - Automatic gain control
  - Lighting change
- Global intensity change detect
  - Updating color tone mapping function whenever a global intensity change is detected



Microsoft Research

Computation reduction

- Color table lookup:
  - Each color tone mapping functions $f_c(x)$ is stored as a color lookup table with 256 entries
  - Color tone mapping for each pixel becomes 3 simple table lookup.
- CPU usage:
  - 5% on a 3.2GHz PC with video resolution 320x240 and frame rate 30fps

Microsoft Research

Experiment results

Microsoft Research

User Study

- 16 video sequences
  - Offices with various lighting conditions
  - Eight different webcams
    - Logitech, Creative, Veo, Dynex, Microsoft LiveCam, etc.
- User is asked to look at the side-by-side view for each sequence, and rate each one with a MOS score:
  - 1: very bad quality
  - 2: bad quality
  - 3: acceptable
  - 4: good quality
  - 5 very good quality

Microsoft Research

User study results

- 18 users responded
- Average improvement:
  - **2.55 ⇨ 3.30**

Sequence ID	Original Video	Enhanced Video
1	2.89	3.50
2	2.94	3. 33
3	1.94	2.67
4	1.61	2.83
5	2.11	3.00
6	3.22	4.17
7	3.11	3.72
8	1.89	2.72
9	2.11	2.17
10	1.94	2.00
11	2.11	3.44
12	3.28	3.72
13	2.65	4.18
14	2.33	3.72
15	3.00	3.89
16	3.72	3.67
Average	**2.55**	**3.30**

Microsoft Research

Extend to Hardware Solution

- Observation:
  - Camera auto exposure responds to environment
  - Local Lighting for target (face) may not be sufficient
- Control exposure based on target
  - Extreme exposure leads to noise
  - Lack of color tone adjustment

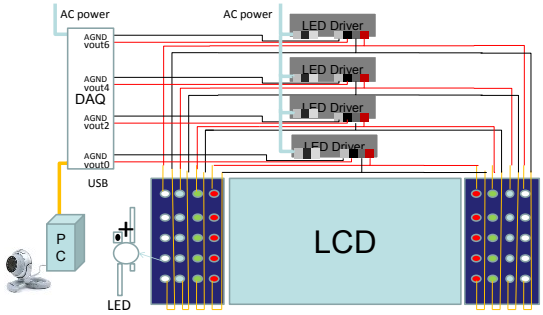
Microsoft Research

Our Solution

- Active lighting
  - Use computer controlled lights
  - Automatically determine the appropriate lighting to optimize face image quality

Microsoft Research

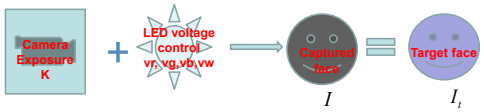
Hardware Configuration



Microsoft Research

Software: Optimization Problem

- Image is a function of exposure and light



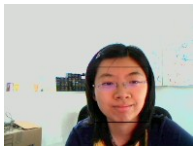
- Goal: Find the right exposure k and voltages **v** to minimize the difference between captured image and target image

$$\min_{k,\mathbf{v}} \|I(k,\mathbf{v}) - I_t\|$$

$$\mathbf{v} = [v_r, v_g, v_b, v_w]^T$$

Microsoft Research

Experiment Results

Video

Microsoft Research

References

- S. A. Bhukhanwala and T. V. Ramabadram, "Automated global enhancement of digitized photographs," *IEEE Trans. on Consumer Electronics*, 40(1), 1994.
- Z. Liu, C. Zhang and Z. Zhang, "Learning-Based Perceptual Image Quality Improvement for Video Conferencing," in *Proc. of ICME*, 2007.
- G. Messina, A. Castorina, S. Battiato, and A. Bosco, "Image quality improvement by adaptive exposure correction techniques," In *Proc. of ICME*, 2003.
- F. Saitoh, "Image contrast enhancement using genetic algorithm," in *IEEE International Conference on SMC*, 1999.
- C. Shi, K. Yu, J. Li, and S. Li, "Automatic image quality improvement for videoconferencing," in *Proc. of ICASSP*, 2004.
- M. Sun, Z. Liu, J. Qiu, Z. Zhang, M. Sinclair, "Active lighting for Video Conferencing," *IEEE Trans. on CSVT*, Vol. 19, No. 12, pp.1819-1829, Dec. 2009.

Microsoft Research

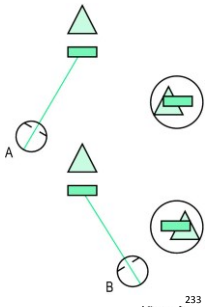
Outline

- Introduction
- Sound source localization with compact mic arrays
- Spatial sound and multi-channel echo cancellation
- Video-based tracking
- Other video processing techniques
  - Learning-based perceptual image quality enhancement
  - Monocular 3D and Foreground/background segmentation
  - Joint tracking and multiview video compression
- Conclusions and future work

Microsoft Research

What is Motion Parallax

- Apparent difference in the direction of movement or speed produced when the subject moves relative to his environment



Microsoft Research

Track User Movement

- Motion tracking systems (Welch and Foxlin 02)
  - Mechanical sensing
  - Inertial sensing
  - Acoustic sensing
  - Magnetic sensing
  - Optical sensing
  - Radio frequency sensing
- For our applications, we use
  - Face tracking via webcams

Microsoft Research

Mono 3D

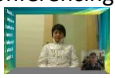
(Pastoor 99, Suenaga 05, Harrison 08)

- Pseudo-3D effects for teleconferencing
  - Box framing

Microsoft Research

Mono 3D

- Pseudo-3D effects for teleconferencing
  - Box framing
  - Layered video



Microsoft Research

Video Segmentation

- Very challenging problem
- Existing scheme using:
  - Stereo (Kolmogorov 05)
  - Mono video (Criminisi 06, Sun 06, Yu 07)
  - Depth camera

Microsoft Research

Fg/Bg Segmentation with Depth Camera

- We explore the use of depth camera for live video foreground/background segmentation

ZCam system specification: Full RGB resolution: 640*480;
Operation range: 0.5-2.5 meter
Depth data rate: 30fps (320*240)

Microsoft Research

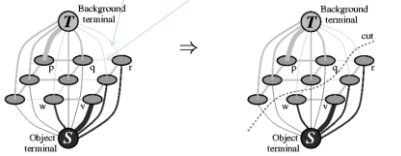
Sample Video

Microsoft Research

Graph-Cut Based Segmentation

- We build a graph model for graph-cut segmentation.

$$P(s|I) = \frac{1}{Z}\exp\left\{\sum_{i\in V}\sum_{k}\lambda_k\Phi_k(s_i,I) + \sum_{<i,j>\in E}\Psi(s_i,s_j,I)\right\} = \frac{1}{Z}\exp\{E(s,I)\}$$



(Boykov and Kolmogorov 03)

Microsoft Research

The Cues for Segmentation

- ZCam provides color images, depth and depth confidence.

$$P(s|I) = \frac{1}{Z}\exp\left\{\sum_{i\in V}\sum_{k}\lambda_k\Phi_k(s_i,I) + \sum_{<i,j>\in E}\Psi(s_i,s_j,I)\right\} = \frac{1}{Z}\exp\{E(s,I)\}$$

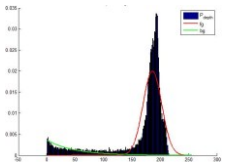
Microsoft Research

Data Term

- Data term in the graph

$$\Phi_{color}(s_i, I) = -\ln\left\{P_{color}(I_i|s_i) \cdot P_{\text{confidence}}(s_i)\right\}$$
$$\Phi_{depth}(s_i, I) = -\ln\left\{P_{depth}(I_i|s_i) \cdot P_{\text{confidence}}(s_i)\right\}$$

The color likelihood is computed from the previous segmentation mask.
The depth likelihood is generated from depth histogram.



$$P_{depth}(I_i^Z|s_i = 1) = G(I_i^Z|_{\mu,\sigma})$$

$$P_{depth}(I_i^Z|s_i = 0) = \gamma e^{-\gamma I_i^Z}$$

242

Microsoft Research

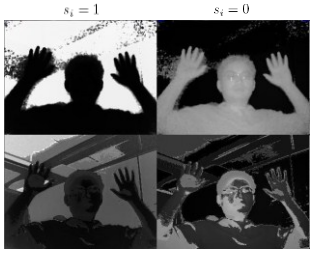
Data Term

- Data term in the graph

$s_i = 1$ $\quad$ $s_i = 0$

$\Phi_{depth}(s_i, I)$

$\Phi_{color}(s_i, I)$



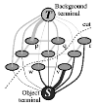
$$\Phi_{color}(s_i, I) = -\ln\left\{P_{color}(I_i|s_i) \cdot P_{\text{confidence}}(s_i)\right\}$$
$$\Phi_{depth}(s_i, I) = -\ln\left\{P_{depth}(I_i|s_i) \cdot P_{\text{confidence}}(s_i)\right\}$$

243

Microsoft Research

Link Term

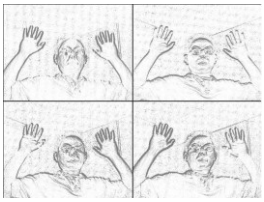
- Link term in the graph



$$\Psi(s_i, s_j, I) = [s_i \neq s_j] \cdot \frac{1}{dist(i,j)} \cdot e^{\frac{-\|I_i - I_j\|^2}{2\beta^2}}$$

Horizontal link term

Vertical link term



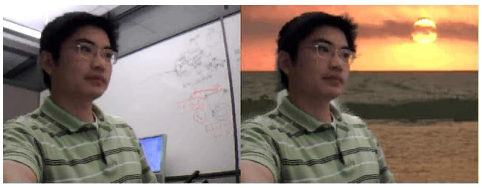
Diagonal link term

Diagonal link term

244

Microsoft Research

Results



Results showed with background replacement.

References

- Y. Boykov and V. Kolmogorov, "Computing geodesics and minimal surfaces via graph cuts", in *Proc. of ICCV*, 2003.
- A. Criminisi, G. Cross, A. Blake, and V. Kolmogorov, "Bilayer segmentation of live video," in *Proc. of CVPR*, 2006.
- C. Harrison and S. E. Hudson, "Pseudo-3d video conferencing with a generic webcam," in *Proc. of the 10th IEEE International Symposium on Multimedia*, 2008.
- V. Kolmogorov, A. Criminisi, A. Blake, G. Cross, and C. Rother, "Bilayer segmentation of binocular stereo video," in *Proc. of CVPR*, 2005.
- S. Pastoor, J. Liu, and S. Renault, "An experimental multimedia system allowing 3-d visualization and eye-controlled interaction without userworn devices," *IEEE Trans. on Multimedia*, vol. 1, no. 1, pp. 41–52, March 1999.
- T. Suenaga, Y. Matsumoto, and T. Ogasawara, "3d display based on motion parallax using non-contact 3d measurement of head position," *in Proc. of the 17th Australia conference on Computer-Human interaction*, 2005.
- J. Sun, W. Zhang, X. Tang, and H. Y. Shum, "Background cut," in *Proc. of ECCV*, 2006.
- G. Welch and E. Foxlin, "Motion tracking: No silver bullet, but a respectable arsenal," *IEEE Computer Graphics and Applications*, vol. 22, no. 6, pp. 24–38, Nov.-Dec. 2002.
- T. Yu, C. Zhang, M. Cohen and Y. Rui, "Monocular Video Foreground/Background Segmentation by Tracking Spatial-Color Gaussian Mixture Models", *IEEE Workshop on Motion and Video Computing*, 2007.
- C. Zhang, Z. Yin and D. Florencio, "Improving Depth Perception with Motion Parallax and Its Application in Teleconferencing," *Proc. of MMSP*, 2009.

Outline

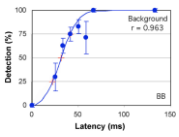
- Introduction
- Sound source localization with compact mic arrays
- Spatial sound and multi-channel echo cancellation
- Video-based tracking
- Other video processing techniques
  - Learning-based perceptual image quality enhancement
  - Monocular 3D and Foreground/background segmentation
  - Joint tracking and multiview video compression
- Conclusions and future work

Motion Parallax

- Improve 3D perception
  - More realistic 3D experience when combined with 3D displays
  - Works with legacy 2D displays

Motion Parallax Delay Sensitivity

- Rendering video from exact viewer perspective is a must. This has impact on:
  - Realism
  - 3D clues
  - Comfort/sickness
- Motion parallax should be kept below **30** ms.

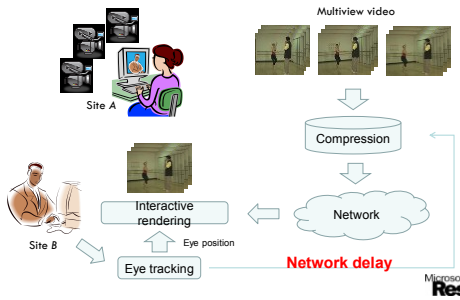


from: Stephen Ellis

- Motion Parallax based Rendering needs to be **LOCAL**.

Enable Motion Parallax for Multiview Video based 3D Teleconferencing

- Eye tracking + multiview video + view synthesis



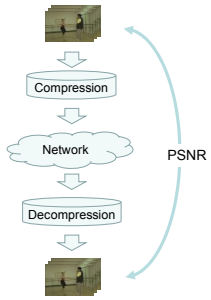
Core Idea

- At the encoding site, predict the viewer's future viewing position
  - Encode 2 nearby frames (Kurutepe 07)
  - Rate distortion optimization (Ramanathan 07)
- Encode each region of the views with different fidelity, *depending on the its use* (weight) on final (synthesized) image at the predicted viewing position.
- At the decoding site, decompress the multiview video, and render with the user's actual viewing position

Microsoft Research

Video Compression Error Criteria

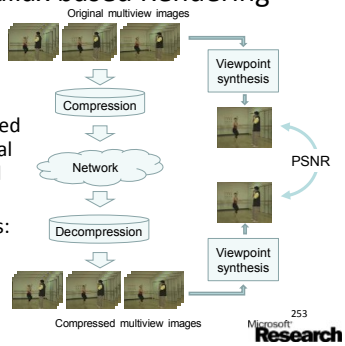
- Traditionally:
  - Input: Video
  - Output: Video
  - Measure PSNR between original frames and the decompressed frames
  - What the user sees: the decompressed frames



Microsoft Research

Multiview Video Compression with Motion Parallax based Rendering

- Input: Video
- Output: Video
- Measure PSNR between synthesized views of the original and decompressed videos
- What the user sees: the **synthesized** frames



Microsoft Research

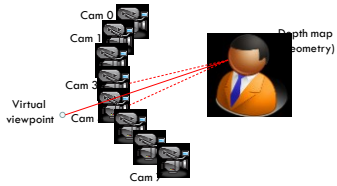
An Example Multiview Frame

254

Microsoft Research

Virtual View Synthesis

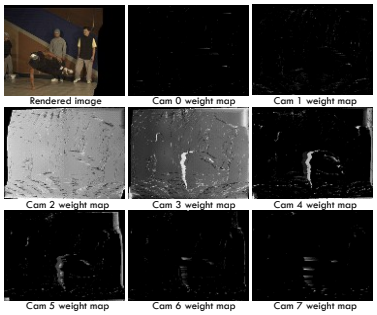
We project (each pixel of) the virtual view onto the depth map. Each point is then traced back to two nearest cameras for interpolation. Occlusion is explicitly computed. (Buehler 01)



255

Microsoft Research

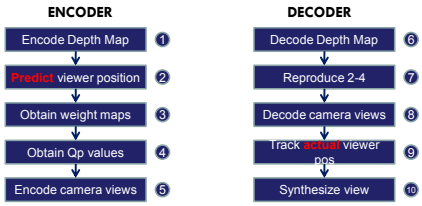
Weight Maps for Rendering



256

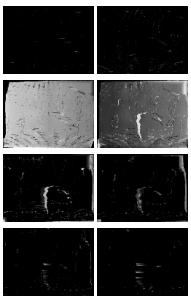
Microsoft Research

Overall Diagram

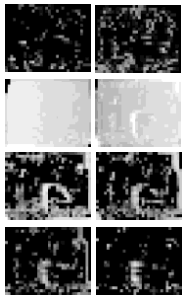


Microsoft Research

Weight Maps and Qp Maps



Weight maps



Qp maps

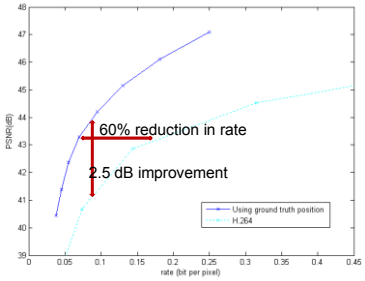
Microsoft Research

Cam 2 Cam 3 Cam 4

Microsoft Research

Microsoft Research

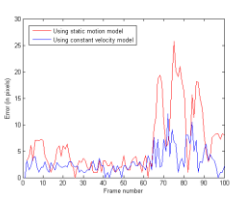
Result with Perfect Viewer Positions

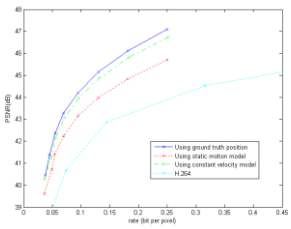


Microsoft Research

What About Prediction Errors?

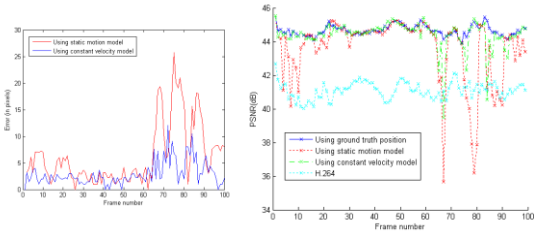
- Can cause significant performance drop





Microsoft Research

Frame-by-frame PSNR



Microsoft Research

Particle Filter (Condensation) based Tracking (Isard 98)

- A Posteriori Probability of the object status at time t:

$$p(\boldsymbol{x}_t|\boldsymbol{Z}_t) \propto p(\boldsymbol{z}_t|\boldsymbol{x}_t)p(\boldsymbol{x}_t|\boldsymbol{Z}_{t-1})$$

$\boldsymbol{x}_t$: object status; $\boldsymbol{Z}_t$: input tracking video stream

where:

$$p(\boldsymbol{x}_t|\boldsymbol{Z}_{t-1}) = \int_{\boldsymbol{x}_{t-1}} p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})p(\boldsymbol{x}_{t-1}|\boldsymbol{Z}_{t-1})$$

assuming object dynamics follows a temporal Markov chain.

Microsoft Research

Particle Filter (Condensation) based Tracking (Isard 98)

- A Posteriori Probability of the object status at time t:

$$p(\boldsymbol{x}_t|\boldsymbol{Z}_t) \propto p(\boldsymbol{z}_t|\boldsymbol{x}_t)p(\boldsymbol{x}_t|\boldsymbol{Z}_{t-1})$$

$\boldsymbol{x}_t$: object status; $\boldsymbol{Z}_t$: input tracking video stream

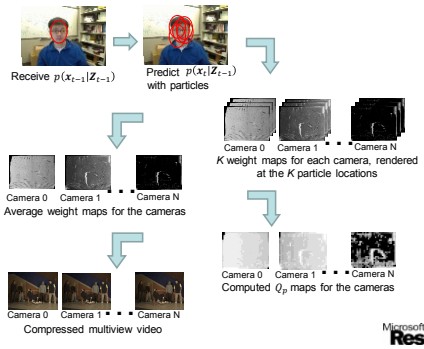
where: Best prediction at the encoder

$$p(\boldsymbol{x}_t|\boldsymbol{Z}_{t-1}) = \int_{\boldsymbol{x}_{t-1}} p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})p(\boldsymbol{x}_{t-1}|\boldsymbol{Z}_{t-1})$$

assuming object dynamics follows a temporal Markov chain.

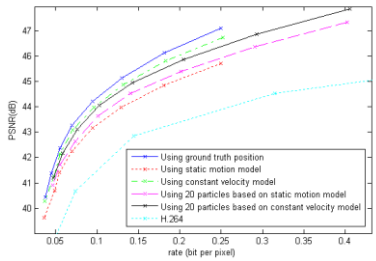
Microsoft Research

Joint Tracking and Compression



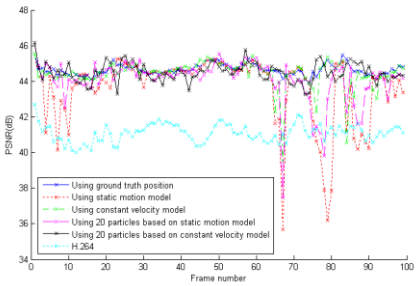
Microsoft Research

Results with Joint Tracking and Compression



Microsoft Research

Frame-by-frame PSNR



Microsoft Research

Frame-by-frame PSNR

Approach	Variance of PSNR (dB)
Using ground truth position	0.097
Using static motion model	3.819
Using constant velocity model	0.693
Using 20 particles based on static motion model	1.477
Using 20 particles based on constant velocity model	0.285
H.264	0.246

Microsoft Research

References

- C. Buehler, M. Bosse, L. McMillan, S. J. Gortler, and M. F. Cohen, "Unstructured lumigraph rendering," *ACM SIGGRAPH*, 2001
- E. Kurutepe, R. Civanlar and M. Tekalp, "Client-driven selective streaming of multiview video for interactive 3DTV", *IEEE Trans. On Circuits and Systems for Video Technology*, Vol. 17, No. 11, pp. 1558–1565, Nov. 2007.
- D. Florencio and C. Zhang, "Multiview video Compression and Streaming Based on Predicted Viewer Position," in *Proc. of ICASSP*, 2009.
- M. Isard and A. Blake, "CONDENSATION – conditional density propagation for visual tracking", *International Journal of Computer Vision,* vol. 29, no. 1, pp. 5–28, 1998.
- P. Ramanathan, M. Kalman and B. Girod, "Rate distortion optimized interactive light field streaming", *IEEE Trans. On Multimedia*, vol. 9, no. 4, pp. 813–825, June 2007.
- C. Zhang and D. Florencio, "Joint Tracking and Multiview Video Compression," in *Proc. of VCIP*, 2010.

Microsoft Research

Outline

- Introduction
- Sound source localization with compact mic arrays
- Spatial sound and multi-channel echo cancellation
- Real-time video processing
- Exploring depth sensors
- Conclusions and future work

Microsoft Research

Conclusions

- Compact mic array SSL and its further improvement with room models
- Audio/visual feature level fusion for speaker detection
- Headphone/loudspeaker audio spatialization
- Multi-channel echo cancellation with Kalman filtering
- Head pose/facial expression tracking
- Depth camera based skeleton tracking
- Learning based perceptual video quality enhancement
- Motion parallax in videoconferencing
- Joint tracking and multiview compression

Future Work

- Audio
  - Audio quality assessment
  - Personal HRTF measurement
  - Mic and speaker at unknown locations
  - Distributed mic arrays
  - …
- Video
  - More reliable/accurate tracking, low computational cost
  - Understand human emotions
  - Depth image processing
  - 3D view synthesis, hybrid virtual/real environments
  - …

Acknowledgement

- Our colleagues
  - Qin Cai, Wei-ge Chen, Ross Cutler, Philip Chou, Dinei Florencio, Rajesh Hedge, Zicheng Liu, Yong Rui, Jamie Shotton, Mike Sinclair, Jay Stokes, Jian Sun, Paul Viola, etc.
- Students
  - Demba Ba, Flávio Ribeiro, Myung-Suk Song, Zhaozheng Yin, Pei Yin, Aswin C. Sankaranarayanan, Qing Zhang, David Gallup, Mingxuan Sun, etc.
- Microsoft product groups
  - Microsoft Unified Communication Group, Microsoft Interactive Entertainment Division Kinect Team, etc.

Thank you!

Microsoft Research