

Unsupervised audio-visual model adaptation for person re-identification with wearable cameras

Alessio Brutti¹

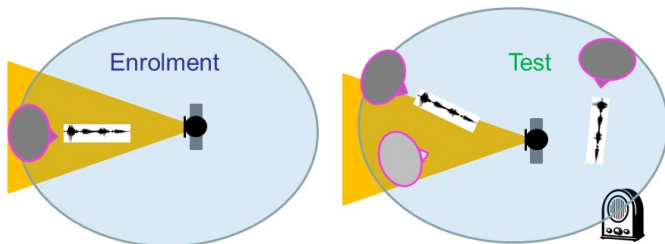
¹Fondazione Bruno Kessler, ICT-irst, Trento, Italy

Smart Cities and Communities Seminars
Trento, April 12, 2018

Application Scenario

Goal

Given an audio-visual device, with co-located sensors
⇒ recognize the identity of the person being seen or heard (or both)



$$\hat{y}_t = \arg \max_{s \in \mathcal{S}} p(y_t = s | \mathbf{x}_t)$$

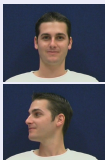
Wearable Cameras

BigGo

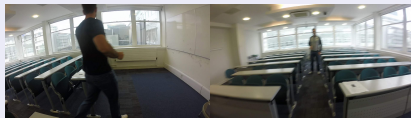


The person identification task on ego-centric data captured with wearable cameras is far more complex

VidTImiT



Wearable Cameras



Wearable Cameras: Challenges

Video

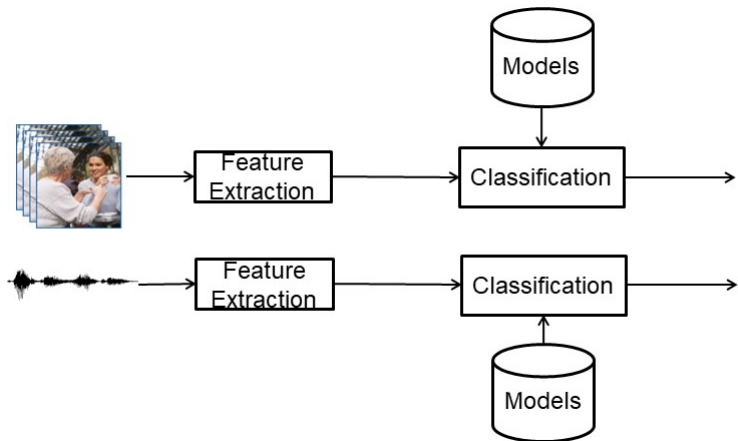
- variety of poses, distortions, ranges, light conditions;
- rapid changes;
- blur due to moving camera;
- person outside FoV or partially visible;
- small amount of training/adaptation material.

Audio

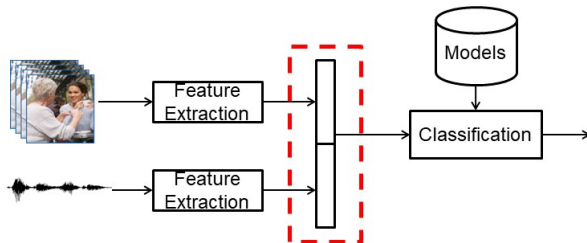
- variety of background noises and acoustic conditions;
- interfering sources;
- noise due to microphone movements;
- conversational speech;
- small amount of training/adaptation material.

Advanced state-of-the-art features and models for auditory-visual person ID are often not applicable

Audio-Visual person ID: overview



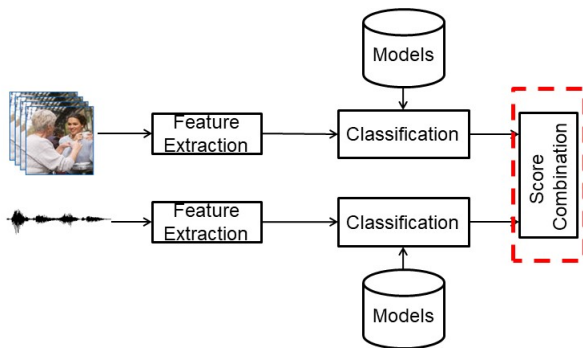
Audio-Visual person ID: overview



Early Fusion

- Concatenation of feature vectors (McCowan et al. 2003) (Smaragdis and Casey 2003)
- Often followed by dimensionality reduction [e.g., PCA (Ngiam et al. 2011)]
- Common in AV speech recognition (Petridis et al. 2017)

Audio-Visual person ID: overview



Late Fusion

Audio and visual info processed individually and independently

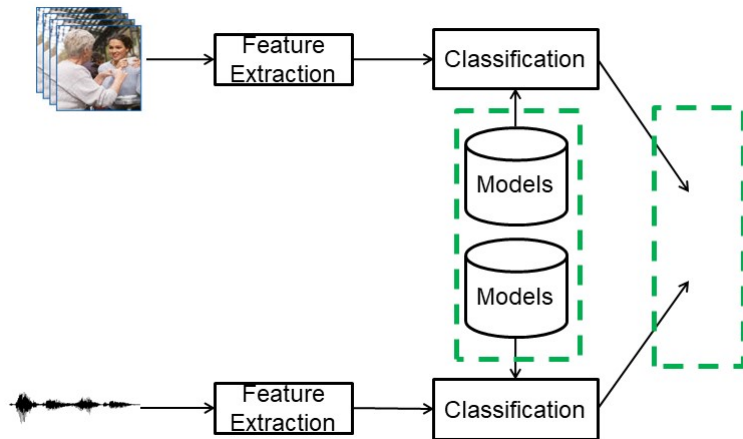
score or decision fusion

(Fox et al. 2007) (Erzin et al. 2005)

score classification

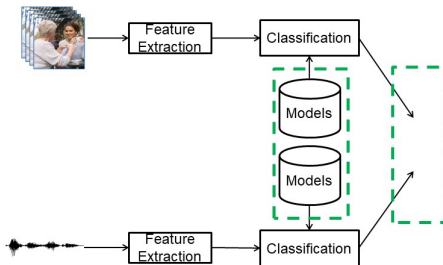
MLP (Sanderson et al. 2003)
RNN (Gandhi et al. 2016)

Proposed approach: model adaptation



Proposed approach: model adaptation

Besides late fusion, we explore multi-modality to adapt/improve models



- appearance changes
- varying conditions
- small training samples

- it doesn't replace late fusion
- **it helps also when one modality is no longer available**

Model adaptation: related works

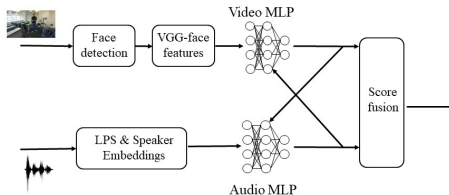
Mono-modal off-line approaches

- unsupervised domain adaptation for speaker verification (using I-vectors) (Garcia-Romero et al. 2014)
- domain adaptation for person re-identification (Ma et al. 2015)

Cross-modal off-line

- Co-EM: combine weak labels from multiple views (Bickel and Scheffer 2005)
- Co-Adaptation for gesture and speech recognition (Christoudias et al. 2006)

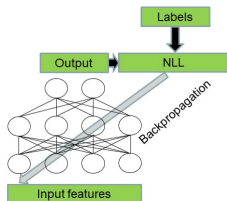
Unsupervised Cross-Modal Model Adaptation



Adaptive time variant models, controlled by the other modality

$$\Theta_t^i \leftarrow f(\Theta_{t-1}^i, \mathbf{x}_t^i, c^j(s, t))$$

Unsupervised Cross-Modal Model Adaptation

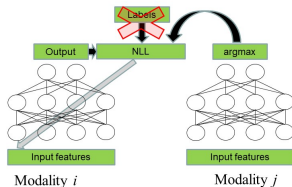


Supervised DNN adaptation

re-training using small learning rates^a

$$\mathcal{L}(\Theta|\mathbf{x}) = -\sum_{t=1}^T \log(p(\tilde{y}_t|\mathbf{x}_t)) + \lambda\|w\|^2$$

^aYu et al. 2013.

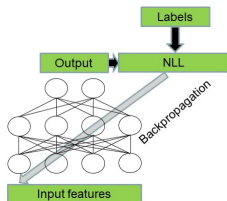


Unsupervised DNN adaptation

labels from the other modality:

$$\mathcal{L}(\Theta^i|\mathbf{x}^i) = -\sum_{t=1}^T \log(p(\hat{y}_t^j|\mathbf{x}_t^i)) + \lambda\|w\|^2$$

Unsupervised Cross-Modal Model Adaptation

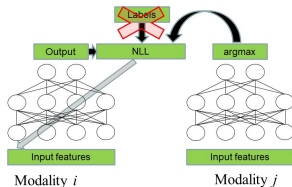


Supervised DNN adaptation

re-training using small learning rates^a

$$\mathcal{L}(\Theta|\mathbf{x}) = -\sum_{t=1}^T \log(p(\tilde{y}_t|\mathbf{x}_t)) + \lambda\|w\|^2$$

^aYu et al. 2013.



Unsupervised DNN adaptation

labels from the other modality:

$$\mathcal{L}(\Theta^i|\mathbf{x}^i) = -\sum_{t=1}^T \log(p(\hat{y}_t^j|\mathbf{x}_t^i)) + \lambda\|w\|^2$$

Risk of divergence due to erroneous labels or in highly mismatched conditions

KLD regularization

The target distribution is a linear combination between the output of the original network and the delta target distribution^{ab}

$$\tilde{p}(y_t^i = s | \mathbf{x}_t^i) = (1 - \rho) \hat{p}(y_t^i = s | \mathbf{x}_t^i) + \rho \bar{p}(y_t^i = s | \mathbf{x}_t^i)$$

where:

- $\hat{p}(y_t^i = s | \mathbf{x}_t^i)$ is the delta target distribution (the unsupervised labels)
- $\bar{p}(y_t^i = s | \mathbf{x}_t^i)$ is the output distribution of the original network

^aYu et al. 2013.

^bFalavigna et al. 2016.

The KLD-regularized loss is:

$$\tilde{\mathcal{L}}(\Theta_t^i | \mathbf{x}_t^i) = (1 - \rho) \mathcal{L}(\Theta_t^i | \mathbf{x}_t^i) - \rho \sum_{s=1}^S \bar{p}(y_t^i = s | \mathbf{x}_t^i) \log(p(y_t^i = s | \mathbf{x}_t^i))$$

Adaptive KLD regularization

$$\tilde{\mathcal{L}}(\Theta_t^i | \mathbf{x}_t^i) = (1 - \rho) \mathcal{L}(\Theta_t^i | \mathbf{x}_t^i) - \rho \sum_{s=1}^S \bar{p}(y_t^i = s | \mathbf{x}_t^i) \log(p(y_t^i = s | \mathbf{x}_t^i))$$

The parameter ρ controls the amount of regularization:

- $\rho = 1$: no adaptation, the network is constrained towards the original distribution;
- $\rho = 0$: re-training without regularization

Adaptive regularization

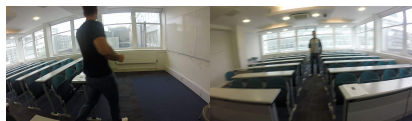
time varying ρ related to the output of the other network

$$\rho_t^i = f(c_t^j)$$

QM-GoPro Dataset

1 person speaking for approximately 1 minute to a person wearing a chest mounted GoPro camera

- 4 different conditions
- 13 subjects
- video:
1920x1080 @ 25 fr/s
- audio: 48kHz, 16 bits
- 5 second long clips (125 images)



Indoor (C1)

Indoor (C2)



Outdoor quiet
(C3)

Outdoor noisy
(C4)

QM-GoPro Dataset

1 person speaking for approximately 1 minute to a person wearing a chest mounted GoPro camera

- 4 different conditions
- 13 subjects
- video:
1920x1080 @ 25 fr/s
- audio: 48kHz, 16 bits
- 5 second long clips (125 images)



QM-GoPro Dataset

1 person speaking for approximately 1 minute to a person wearing a chest mounted GoPro camera

- 4 different conditions
- 13 subjects
- video:
1920x1080 @ 25 fr/s
- audio: 48kHz, 16 bits
- 5 second long clips (125 images)



Experimental set up

- Training: first 3 segments for each person in each recording condition
- Test segments are sorted in random order; results are averaged over 20 sequences

Evaluation

- multi-class classification: recognition accuracy
- Performance evaluated using:
 - ▶ *Matched* models: models are trained in the same condition as test
 - ▶ *Mismatched1* models: models trained in C1 (indoor 1)
 - ▶ *Mismatched2* models: models trained in C3 (outdoor quiet)
- algorithms:
 - ▶ “baseline”: no adaptation
 - ▶ “cross”: cross-modal labels without KLD
 - ▶ “KLD”: proposed approach

Implementation details

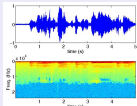
Video

- Face Detector^a(**Wu-2015**)
- 2048-dim deep-features from VGG-face^b
- average pooling for segment classification
- $\rho_t^v = c_t^a$

^a<https://github.com/tornadomeet/mxnet-face>

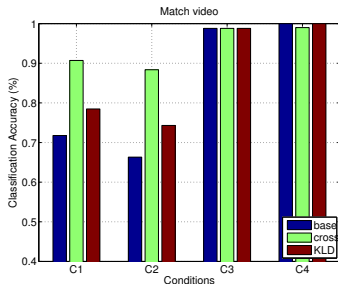
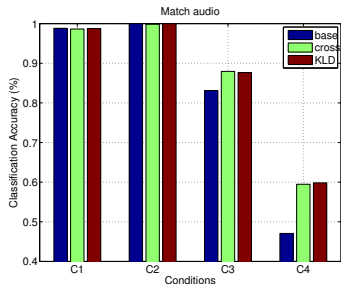
^b<https://github.com/rcmalli/keras-vggface>

Audio

- Log-Power-Spectrum
 $\log(\| \text{FFT}([s(t), \dots, s(t+T)]) \|)$
- The figure contains two subplots. The top subplot is a waveform plot showing amplitude over time (0 to 5 seconds). The bottom subplot is a spectrogram plot showing frequency (0 to 2 kHz) over time (0 to 5 seconds). Both plots have a y-axis scale of 10^2.
- DNN for frame-base speaker classification trained on (part of) WSJ
- 1024-dim speaker embeddings
- average pooling for segment classification
- $\rho_t^a = c_t^v$

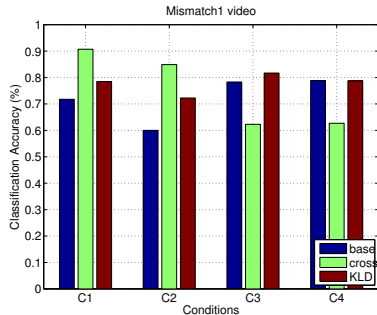
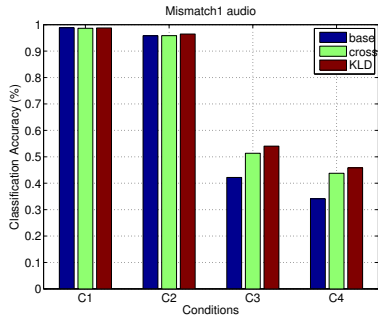
Experimental results

Results in *matched conditions*



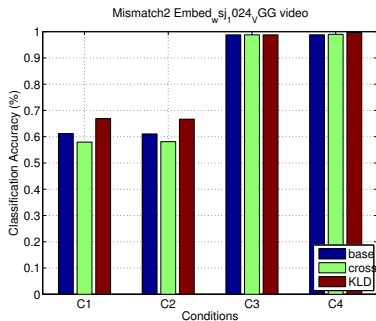
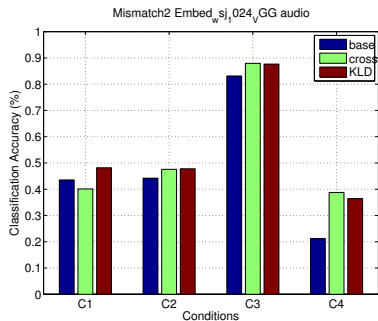
Experimental results (2)

Results in *mismatched1* conditions (C1)



Experimental results (3)

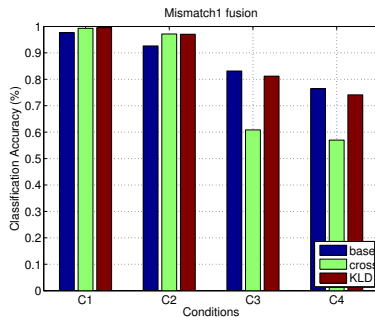
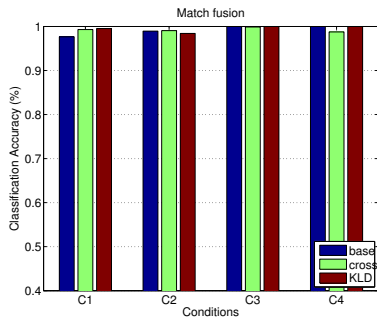
Results in *mismatched2* conditions (C3)



Experimental results (4)

Score Fusion

Sum rule: $c^{av}(s, t) = \gamma_t c^a(s, t) + (1 - \gamma_t) c^v(s, t)$
 γ_t depends on score variances



Take away

- Person re-identification for wearable cameras brings specific challenges
- Cross-modal adaptation helps coping with rapidly varying conditions
- KLD regularization is a promising and transparent tool to avoid over-fitting and model divergence

Open issues/ Limitations

- Deriving an optimal regularization parameter (the current is far from being optimal)
- The dataset is very small (hard to generalize and performance rapidly saturate)

Other efforts on audio-video processing

3D person tracking with co-located sensors (Qian et al. 2018)

- co-located camera and microphone array do not allow 3D inference
→ no triangulation for depth estimation
- single camera → targets outside FoV or occluded
- smart use of multi-modality:
face detection + generative visual likelihood +
video-driven audio processing for depth estimation



Thank you for your attention

Questions?

Modality	Adapt		Matched				Mismatched			
			C1	C2	C3	C4	C1	C2	C3	C4
audio	cross	mean	90.71	90.00	91.63	70.94	90.71	78.05	55.36	32.00
		std	2.55	1.91	2.19	3.21	2.55	2.94	4.18	4.86
	KLD	mean	92.53	92.47	92.41	73.47	92.53	80.68	56.02	29.71
		std	1.99	1.53	1.87	3.20	1.99	1.68	1.68	2.49
video	cross	mean	87.06	87.47	85.66	63.59	87.06	75.95	50.48	32.71
		std	2.75	2.18	2.61	3.13	2.75	2.53	3.26	4.15
	KLD	mean	74.12	80.68	93.98	81.82	74.12	62.53	36.33	11.59
		std	2.38	2.88	1.35	3.01	2.38	1.94	2.00	1.28