Video to the rescue

Girija Chetty & Michael Wagner

girija.chetty@canberra.edu.au

University of Canberra, Australia
Motivation

• Next generation HCIs need to be personalized.
• Personalization features of HCI will increase the rich interaction experience for day-to-day civilian applications:
  – Travel planning and arrangements
  – Mobile banking, online shopping, access control
• Security/identity management needed for personalized HCIs.
• Identity management not based on traditional pins/passwords.
Current problems

- Two fold:
  - HCIs not personalized or secure.
  - Security/Identity management systems not quite suitable for civilian HCIs.

- Technical/deployment problems of HCIs limit systems’ ability to perform satisfactorily in real-world settings under adverse conditions.

- Efficient identity management solutions based on intrusive biometrics are available for high-security applications operating in controlled settings.

- More acceptable biometric traits needed for HCI operating scenarios.
Biometrics for personalized/secure HCIs

• Recent EU report– proposed enormous biometric diffusion by 2015
  

• Example everyday scenarios
  
  – biometric access control at child care centres which unobtrusively scans parents when they ring the doorbell.
  
  – Over-65 bus pass holders with facial template on the smart cards which needs to be renewed every year
  
  – Computer games
  
  – Video rentals
  
  – Gas/cooking appliances (kitchen) -avoid accidents with kids.
  
  – **US demand to grow 10+% per year through 2009**
  
  – The US market for biometric and other electronic access control products and systems was $4.4 billion in 2004 and will increase more than ten percent per year through 2009 to $7.2 billion

[Link to US market report](http://www.the-infoshop.com/study/fd33172-biometric.html)
Video/Visual HCIs for personalization and security

• Video-Rates high in terms of User acceptance and usability.

• Video is an inherently multimodal signal
  – with Voice & Face info (static and dynamic)
  – Video provides abundant data - better training
  – Video allows face tracking- track facial expressions and facial/acoustic emotions
  – Video provides temporal continuity

• However not all properties of Video exploited so far in building Visual/Video HCI systems
But video biometric trait more vulnerable to forgery

- Current video based identity verification techniques more vulnerable to forgery
- Forgery scenarios
  - Pre-recorded audio
  - Still photo
  - Pre-recorded video
  - Animated video from a still photo
- Vulnerable to environmental degradations
  - Acoustic noise effects
  - Visual artefacts-illumination/pose variations and compression artefacts
Vulnerability of biometric person authentication to fraudulent attacks

Solution - Multimodal fusion and Liveness Verification

Spoof attacks

Fake biometric

Replay attacks

Modify templates

Intercept The channel

Override templates

Override matcher

Override Final decision

YES/NO

Synthesize Feature vector

Override feature extractor

Replay old data

Replay

Final decision
Video based Biometric Security framework

- Inherent multimodality, data abundance and temporal continuity in video

- Visual manifestation of speech (speaking faces) contains person specific information.

- A multi-level security framework based on
  - Multimodal Fusion of face-voice biometric information from and
  - Liveness detection/verification by different type of fusion and feature-extraction techniques

Multi-level security
- Level 1 security: static attacks
- Level 2 - video attacks
- Level 3 - synthetic speaking face attacks

Human Computer Communication Laboratory, School of ISE, BLIS Division
VidTimit, AVOZES, and UCBN databases used for Evaluation

Human Computer Communication Laboratory, School of ISE, BLIS Division
Methodology

• Identity Verification – Late fusion of audio-visual features
Methodology

• Liveness Verification – feature fusion of audio-visual features
Level 1 Security-Bi-Modal Feature Fusion (BMF)

- Verify liveness by feature-level fusion of acoustic features + dynamic lip features from lip region
Level 2 Security: Cross-Modal Fusion (CMF)

- Detect liveness of face-voice biometric data by extracting audio-visual synchrony based on:
  - Latent Semantic Analysis (LSA) features:
  - Canonical Correlation Analysis (CCA) features:

Visual PCA vector

Acoustic MFCC vector

Singular Value Decomposition/
Maximizing mutual
Information

LSA/
CCA Features
Level 3 Security: 3D Multi-modal Fusion (3MF)

- Detect Liveness by extracting depth information from faces:
- 3D Face models
  - 3D Shape + texture + acoustic features
- Pre-processing/enhancement
  TPS warp (shape features)
  & SRT (texture features)
Impostor and Replay attack scenario modeling

• Traditional identity verification experiments using Client/Impostor trials, with audio-visual GMM speaker models.

• Extended the protocol for testing replay attacks by synthetic replay attacks
  – Still photo + pre-recorded audio
  – Synthesized video from still photo

• Extending the protocol – speaking face synthesis
  – Direct encoding of Context Information.
  – Fusion of different acoustic features
  – LDA on Acoustic fusion vector
  – Use of HMM for mapping AV information

Human Computer Communication Laboratory, School of ISE, BLIS Division
Feature extraction

• Audio Features: predictive of orofacial motion.

• Three types of features:
  – prosodic group comprising (F0) and (E);
  – the segmental group comprising (LPC) and (LSF);
  – perceptual group comprising MFCC and PCBF
Visual Features

- Four orofacial parameters: $MW(=V1)$, $MH(=V2)$, $CH(=V3)$ and $LL(=V4)$
HMM: facial parameter prediction

• HMM was first trained on the joint A/V space, by combining the audio and visual features into one joint observation vector.
• Audio HMM is extracted from trained HMM.
• Then, the audio input & HMM corresponding to each state is used to analytically derive the visual estimate (“inverted” Baum Welch re-estimation algorithm)

\[ \hat{v}_t = \arg \max P (a_t, v_t, q_t | \lambda) \]
FAP animation

• To verify the accuracy of the tracking method and the resulting predictions, we reconstruct the faces from facial animation parameters (FAP) using an MPEG4 face animation engine :Xface toolkit.
Results: BMF approach
CMF approach

Human Computer Communication Laboratory, School of ISE, BLIS Division
3MF approach

Human Computer Communication Laboratory, School of ISE, BLIS Division
Conclusions

• Visual/Video based HCIs allow more personalized user experience.
• video-based multilevel security framework for personalization
  – eliminates impostor attacks completely.
  – detects fraudulent replay attacks.
• Completely software oriented solution.
• An extendable framework that can address all futuristic impostor forgery scenarios.