Learning Based Multimedia Quality Assessment

LI Qiaohong

(Supervisor: Prof Weisi Lin)

(Co-Supervisor: Prof Daniel THALMANN)

Oct. 29, 2013

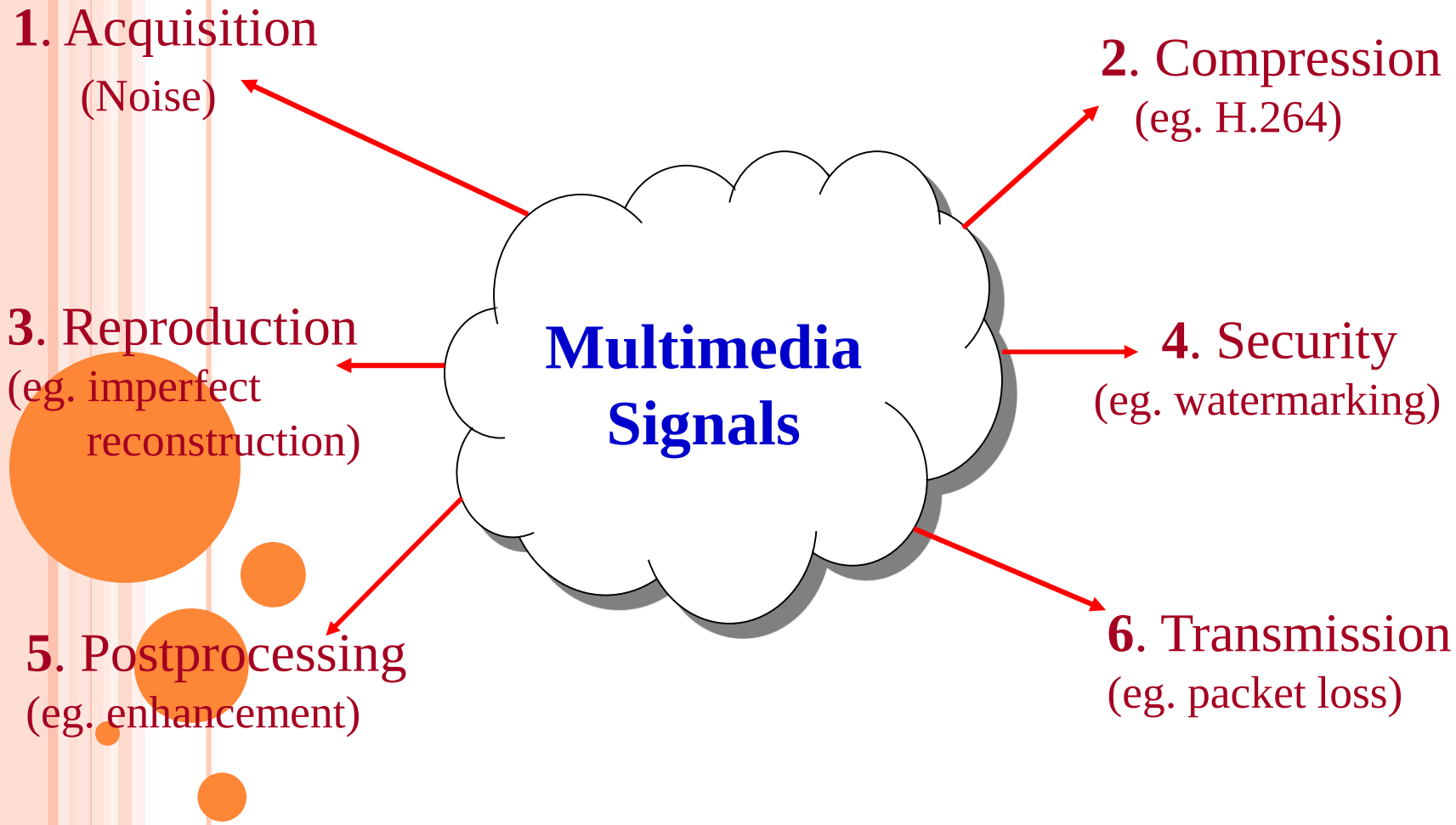
NANYANG TECHNOLOGICAL UNIVERSITY

School of Computer Engineering

Outline

- Motivation
- Review
- Application of deep learning in Paralinguistic Recognition
- Application of deep learning in QA problem
- Conclusion and future work

Motivation


1. Acquisition
(Noise)
2. Compression
(eg. H.264)
3. Reproduction
(eg. imperfect
reconstruction)
Multimedia
Signals
4. Security
(eg. watermarking)
5. Postprocessing
(eg. enhancement)
6. Transmission
(eg. packet loss)

Motivation

- **Applications**: In optimizing resource allocation
  - signal acquisition, enhancement, watermarking, compression, transmission, reconstruction, authentication, medical imaging… (**applications in virtually all multimedia communication systems with regards to QoS and QoE!!!)**

- Subjective assessment suffers from **drawbacks**
  - time-consuming, laborious and expensive; requires many human subjects and repeated viewing/listening sessions
  - Not feasible for on-line signal manipulations (such as encoding, transmission, relaying, etc.)
  - depends upon viewers' physical conditions, emotional states, personal experience etc

Approaches for Visual Quality Assessment

- Full-Reference VQA
  - **Vision Modeling approach (Bottom up)**

  Attempt to **model** the **Human Visual System** (HVS), try to incorporate aspects of human vision, Involve expensive computations

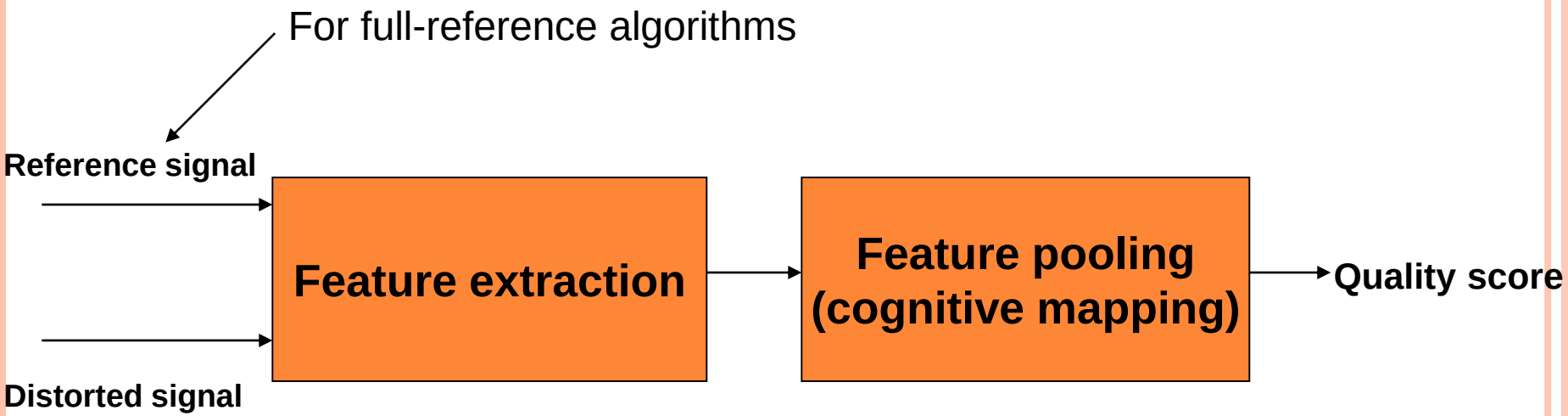
  Eg. Visual Difference Predictor (VDP), DCTune, Sarnoff JND, PSNR-HVS-M…
  - **Engineering approach (Top down)**

  Based on extraction and analysis of features (exploit statistical properties)

  Focus is on **image content** and **distortion analysis** rather than fundamental vision modeling

  Eg. Structural Similarity Index Measure (SSIM), Visual Information Fidelity (VIF), Edge based metrics etc.
- No-Reference VQA
- Reduced-Reference VQA

General Framework



Stage I

Exploits signal processing techniques

Stage II

Based on machine learning

No-Reference VQA

- Before 2010
- Distortion-specific
- JPEG/JPEG2000/Blur

- After 2010
- Distortion-agnostic(NSS, machine learning)

- Recently
- General

General NR IQA method

Machine Learning has been substantially used in NR-IQA problem

- ***Unsupervised machine learning to extract features for machine learning algorithms***

  Eg. Sparse dictionary learning+SVM

- ***Hand-craft features for machine learning algorithms***

  Eg. Garbor filters for feature extraction, Natural Scene Statistics features, wavelet features

- ***Neural Network methods***

  Eg. Combine the feature extraction and regression model training together.

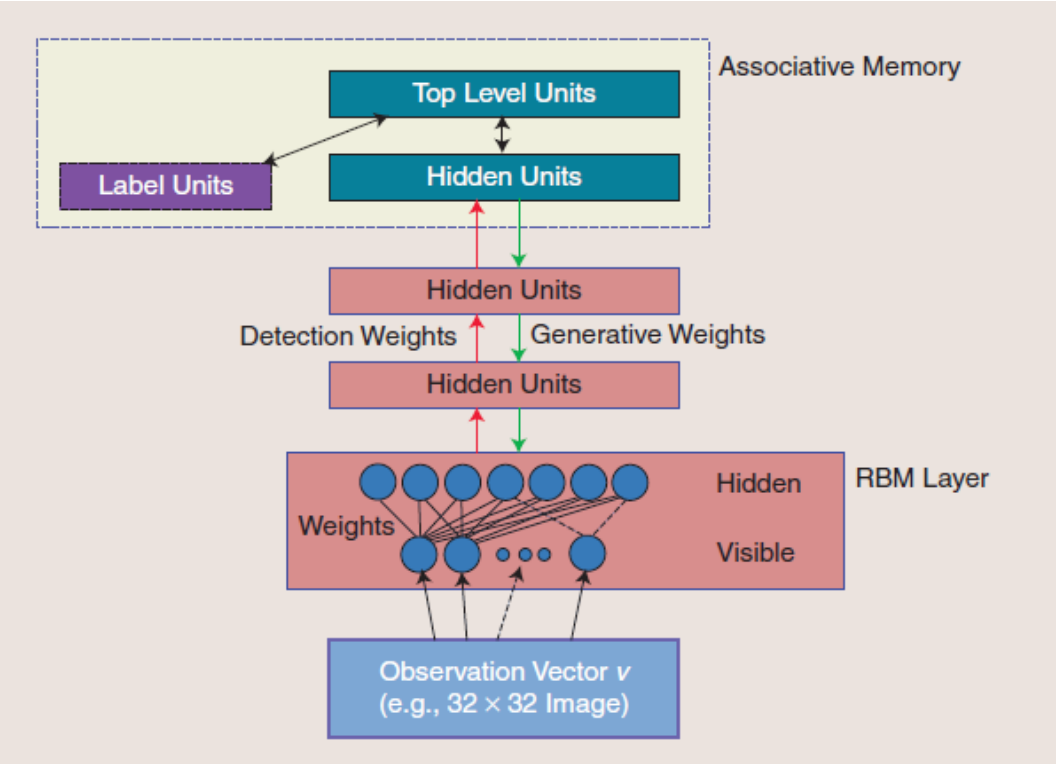
  - ***multivariate GGD model to fit the feature***

Deep Machine Learning

- ***Definition***

  A class of machine learning techniques that exploit many layers of non-linear information processing for supervised or unsupervised feature extraction and transformation, and for pattern analysis and classification.

  - ***Illustration of Deep Belief Network***



Application in Paralinguistic Recognition

Paralinguistic

Speech signal, as one of the most important media of interpersonal communication and human-machine interaction, not only conveys the textual information, also includes many supra-segmental properties that modulate and enhance its meaning

The Autism Sub-Challenge is based upon the "Child Pathological Speech Database

	UAR_{devel} (Typicality)	UAR_{devel} (Diagnosis)
Baseline(SVM)	92.8	52.4
Auto features+SVM	92.8	**55.4**
MLP	91.8	51.1
RBM+NN	91.6	**53.9**
ClassRBM+NN	**92.9**	**57.6**
ClassRBM	91.8	**53.4**

Future Work

- Using Auto-encoder for feature extraction and SVR for regression for Speech Quality Assessment
- Using DBN for pretraining DNN for Speech Quality Assessment
- The same architecture for VQA
- Joint audiovisual quality assessment:
  - humans perceive 'overall' multimedia quality and not separate assessment
  - Possible approaches include one-stage and two-stage fusion (OSF/TSF)
    - OSF: both audio and speech features pooled in one stage
    - TSF: first pool audio, then video features and the two scores into an overall score

Thank you!

Questions?