Correlating Video Quality Metrics to User Experience: an Event-based Approach

- Master's Thesis Defense -

Yongfeng Huang

Supervisor: Prof. James Won-Ki Hong

Division of ITCE, POSTECH, Korea

xgjonathan@postech.ac.kr

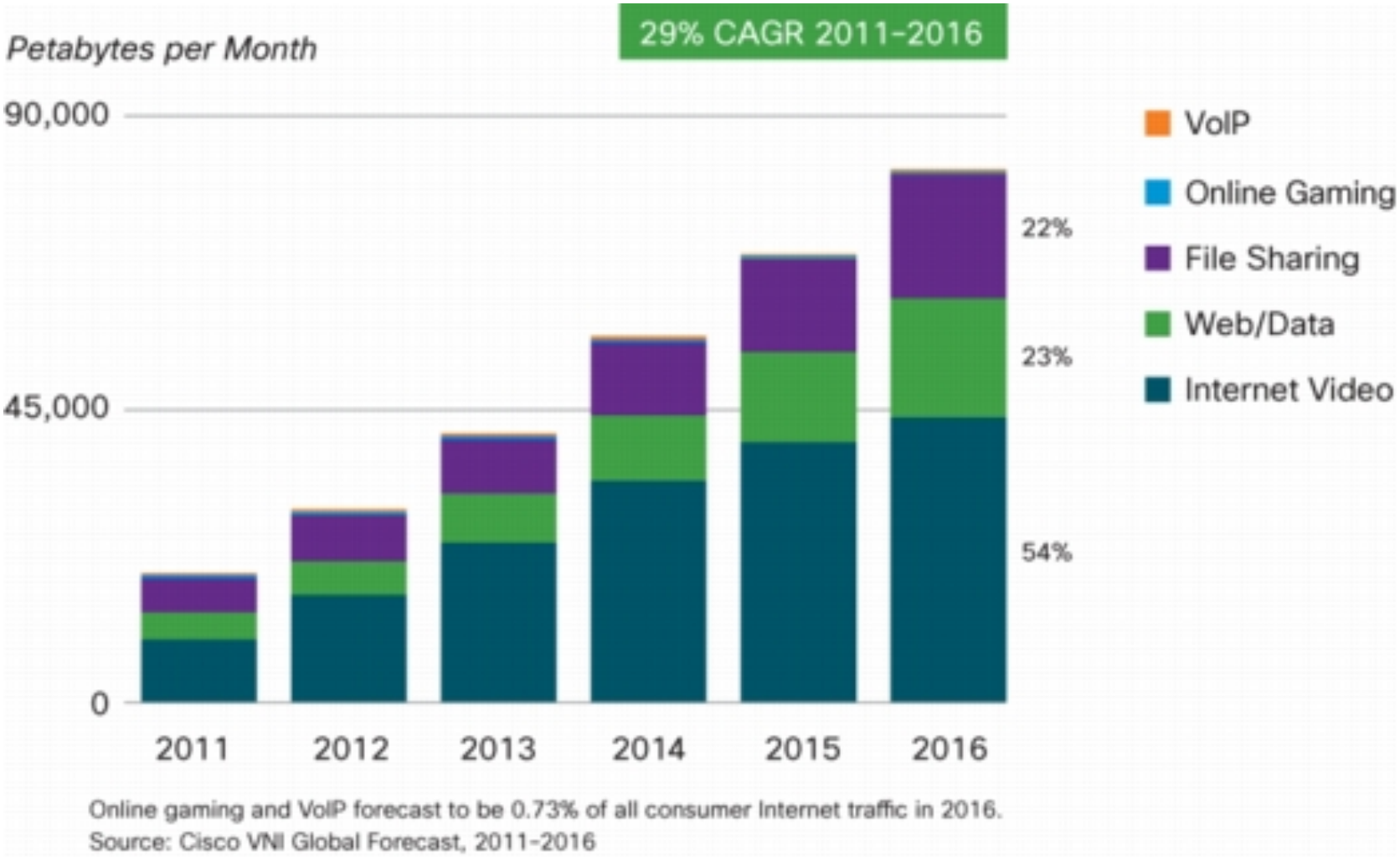
2012.06.22

Contents

- Introduction & Motivation
- Background & Related Work
- Event-based user experience assessment
  - Classification of defect events
  - Event to experience correlation
- Validation
- Conclusion & Future Work

Internet Traffic Trend

❖ Multimedia will dominate IP traffic



http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/VNI_Hyperconnectivity_WP.html

The Management Challenge

❖ Management of multimedia services

- For content providers, ensuring service quality is important (*service differentiator to attract and retain customers*)
- To be specific, network service providers need to **assess**, **manage**, and **improve** user perceptual experience

❖ User experience of video quality

- Compared to quality of service (QoS), quality of experience (QoE) considers human perception
- Human vision system (HVS) is very complicated to model, i.e. user's preference for video content, user's interest of frame region, etc.

Current Limitations

❖ User feedback is impractical in reality

- Well-controlled user experiment with a wide variety of test elements is not easy to design or administer
- Users cannot be bothered with giving truthful feedback each time

❖ Limitation of current objective measurements

- Temporal factors are missed, i.e. the peak signal-to-noise ratio (PSNR)
- Content-related factors are missed
- Events play a crucial role in human experience (*not reflected*) [J. M. Zacks et al., 2010]

Objectives

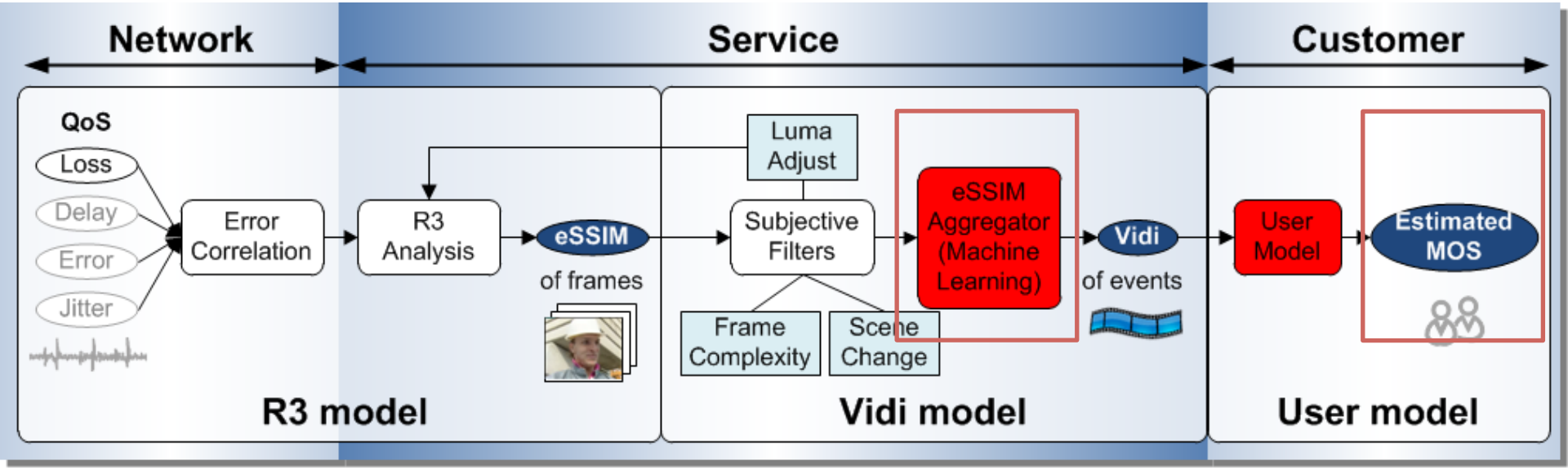
- Propose an approach for correlating video quality metrics to user experience
  - Accurate and efficient
- Consider event-based human experience
  - How to identify/segment defect events in videos
  - How to classify defect events
  - How to correlate events to user experience

Contributions

- An event-based approach for correlating video quality metrics to user experience
  - The first (to the best of my knowledge) to consider that human experience is event-based
  - Event classification: accurate and efficient classification of defect events (eSSIM Aggregator)
  - Event-to-experience correlations: correlate different event types to user experience by constructing event-specific user models.

❖ VIDAR: video quality analyzer in real-time [A. Kwon et al., 2012]

- Correlate network performance to user experience
- Structure Similarity (SSIM) – frame quality metric [Z. wang et al., 2004]



Machine Learning (ML)

Mean opinion score (MOS)

Related Work

❖ Statistical approach

- Rely on one or two salient perception characteristics
- Mean square error (MSE) [A. Bhat et al., 2009]
- PSNR [O. Nemethova et al., 2006]
- Content factors are missed

❖ Machine learning approach

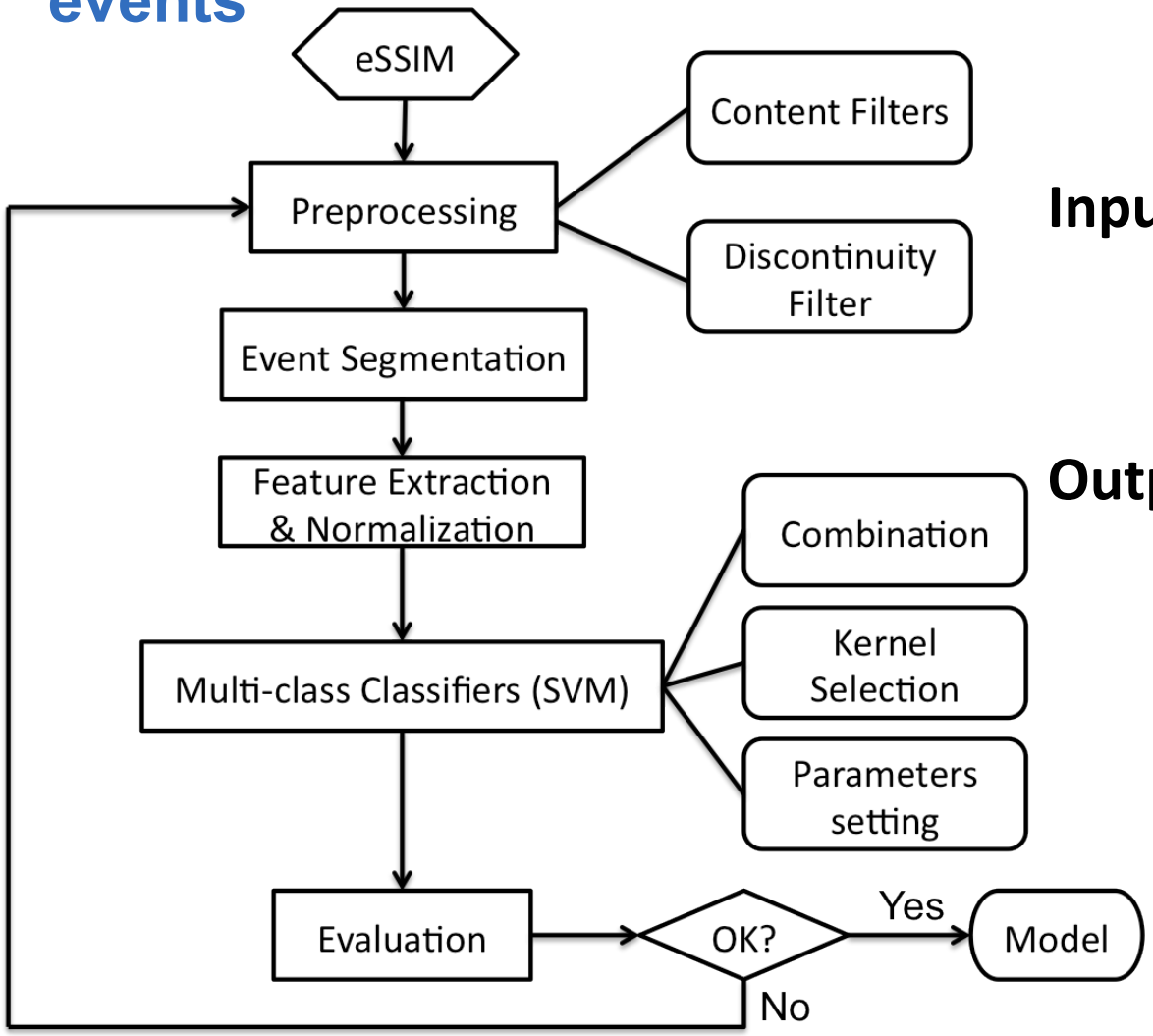
- Decision tree [V. Menkovski et al., 2009]
  - Data instances: video spatial information, video temporal information, frame rate and bit rate
  - Result: “YES” or “NO”

Contents

- Introduction & Motivation
- Background & Related Work
- Event-based user experience assessment
  - Classification of defect events
  - Event to experience correlation
- Validation
- Conclusion & Future Work

Process Flow of eSSIM Aggregator

- **eSSIM aggregator for detecting and classifying defect events**



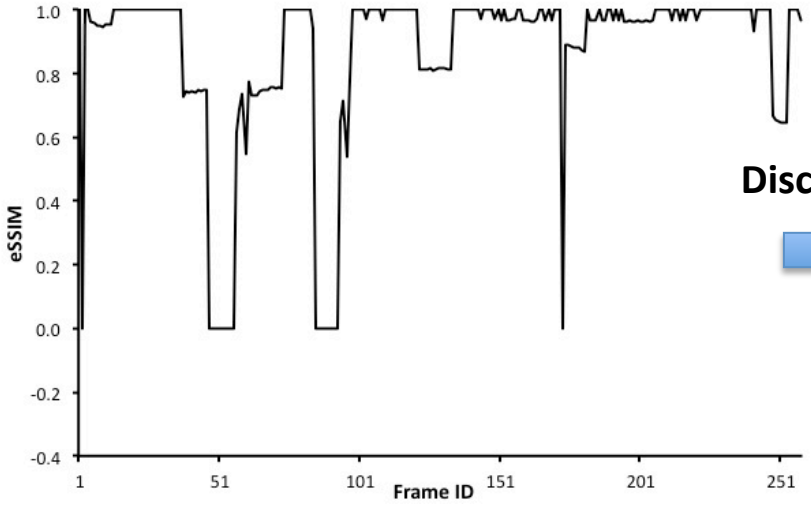
Input: eSSIM of frames

Output: defect events

Preprocessing (Discontinuity Filter)

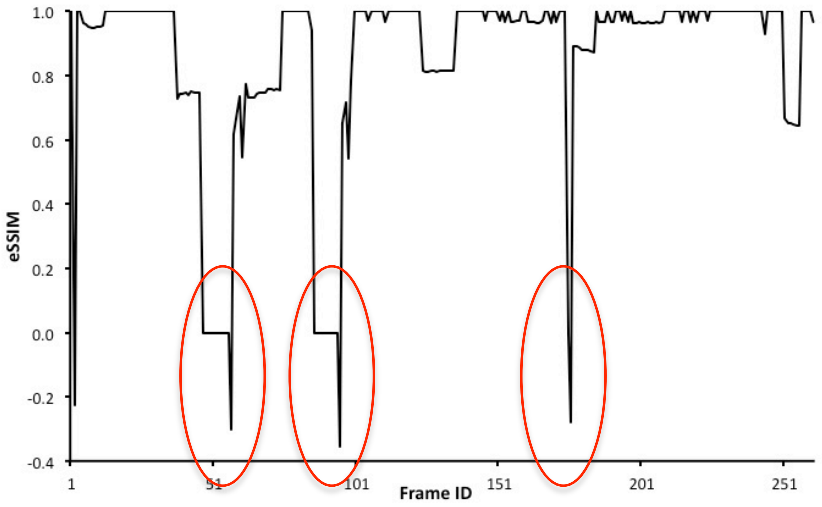
- Meaning of eSSIM values
- Show the intensity of discontinuity in the video

eSSIM of frames



Discontinuity filter

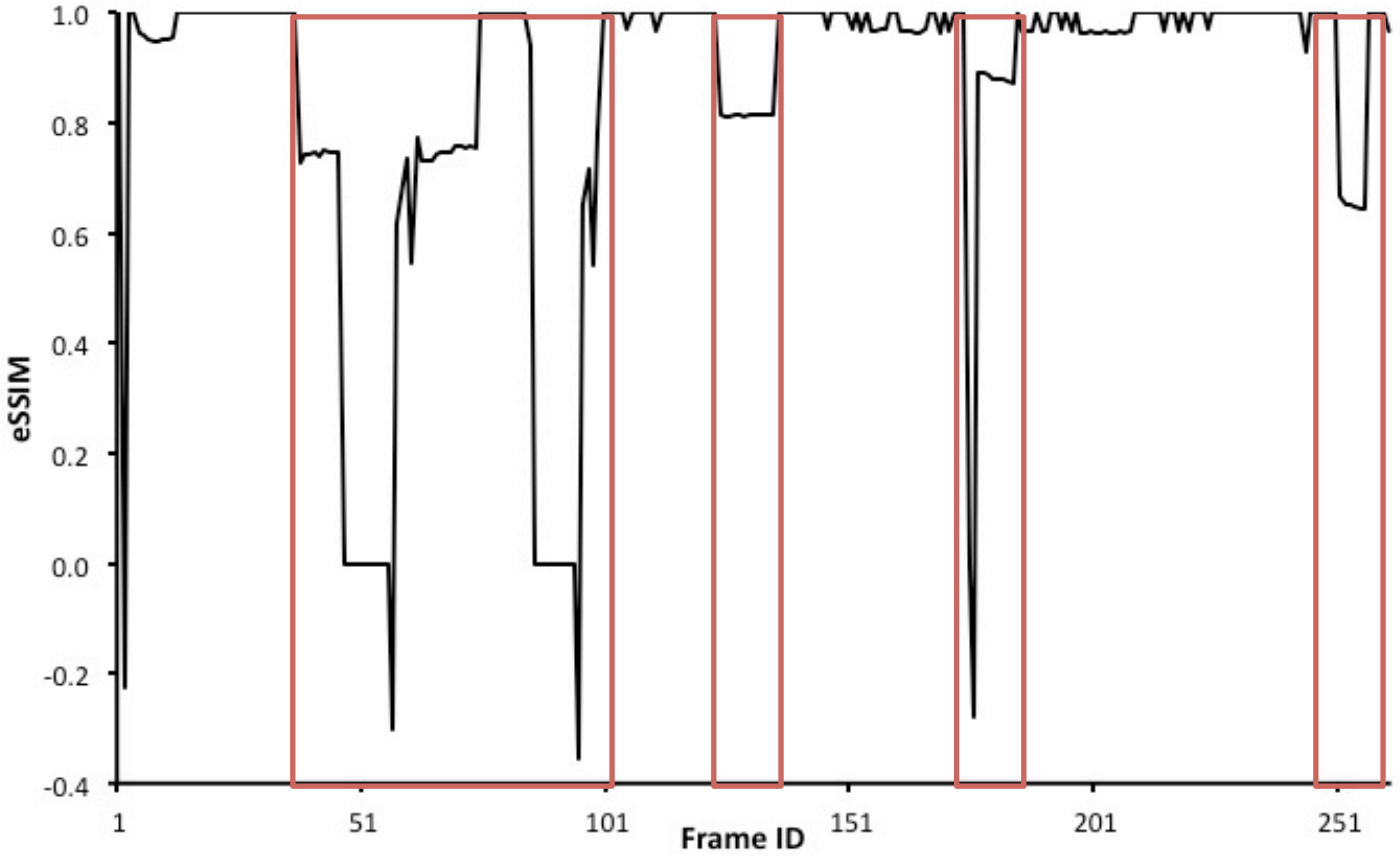
eSSIM of frames with discontinuity mark



Foreman, GOP = 12, 2% GE loss model

❖ A set of rules to segment defect events

- i.e. the first 10 frames of each video are ignored



Foreman, GOP = 12, 2% GE loss model

Feature Extraction and Normalization

❖ Purposes of extracting features

- To distinguish different types of defect events
- Reduce the dimension and input of ML classifiers

❖ Features extracted

- ***Mean (μ)*** **and** ***standard deviation (δ)***
- ***Minimum***: the minimum e***SSIM*** value of an event
- ***Defective ratio***:

$$ratio = \frac{N_{eSSIM<0.95}}{n}.$$

- Severity of dropped and duplicated frames:

$$severity = \frac{N_{eSSIM\leq 0.0}}{n}.$$

- Skewness: measure of the asymmetric of the probability distribution of event data
- Kurtosis: measure of whether the distribution is peaked or flat, relative to a normal distribution

❖ Feature normalization

- Middle-range normalization: [-1, 1]
- Mean-std normalization: mean = 0, std = 1

Multi-class Classifiers - Types of Defect Events

- Distortion: a series of frames containing perceivable distortions
- Glitch: a series of frames, where distortion is short and slightly perceivable
- Freezing: a series of duplicated frames
- Video samples of defect events

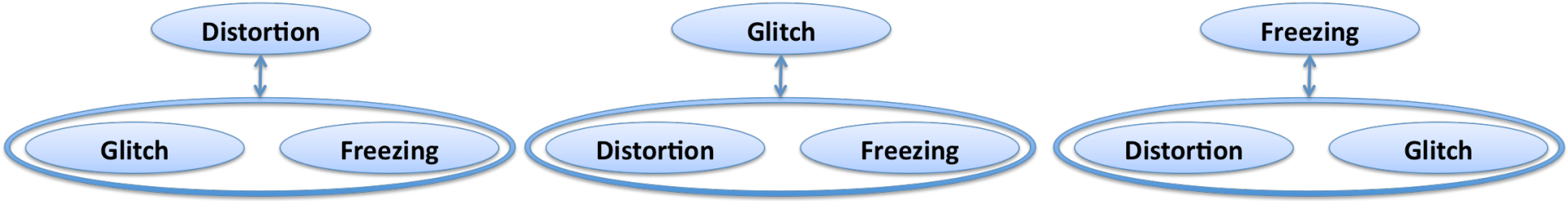
Multi-class Classifiers

❖ Support vector machine (SVM)

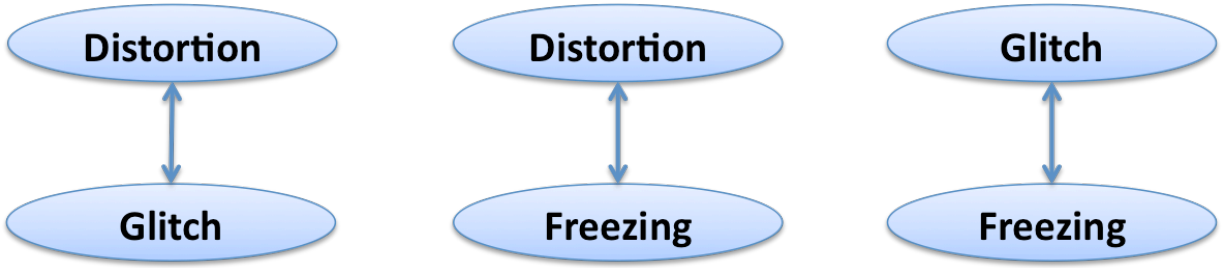
- Multi-dimension and continuous features
- Nonlinear situations

❖ Multi-class classification with SVM

- Combination of binary SVM classifiers
  - One-versus-all:



  - One-versus-one:



Correlating event metrics to User Experience

- **Correlational models for predicting user MOS from event metrics**

- **Intensity features for Distortion**

  {*μ*, *ratio*, *frame complexity*, *motion speed*}

- **Intensity features for Glitch**

  {*μ*, *ratio*, *frame complexity*, *motion speed*}

- **Intensity features for Freezing**

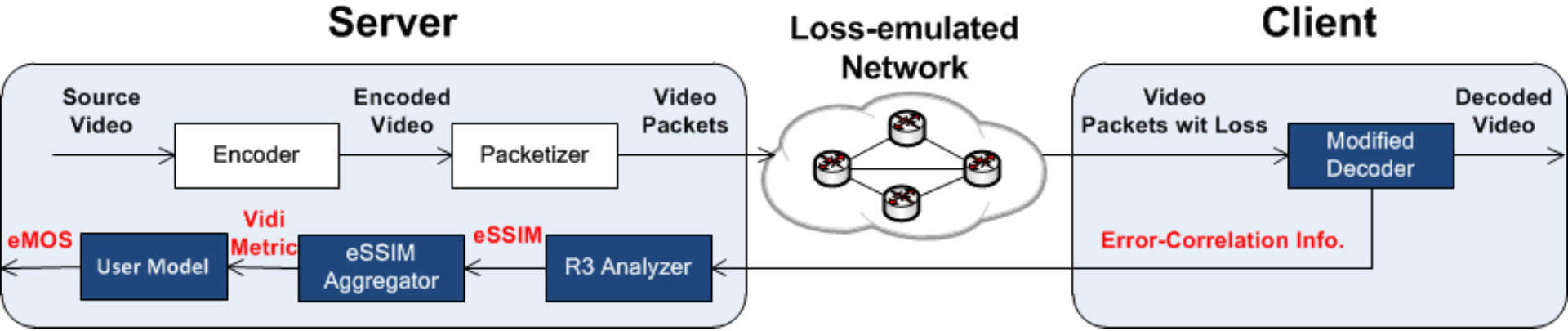
  {*the time duration*, *discontinuity intensity*}

- **Artificial neural network (ANN)**
  - **No readily discernable patterns statistically**
  - **ANN shows good accuracy**

Contents

- Introduction & Motivation
- Background & Related Work
- Event-based user experience assessment
  - Classification of defect events
  - Event to experience correlation
- Validation
- Conclusion & Future Work

❖ Experiment setup



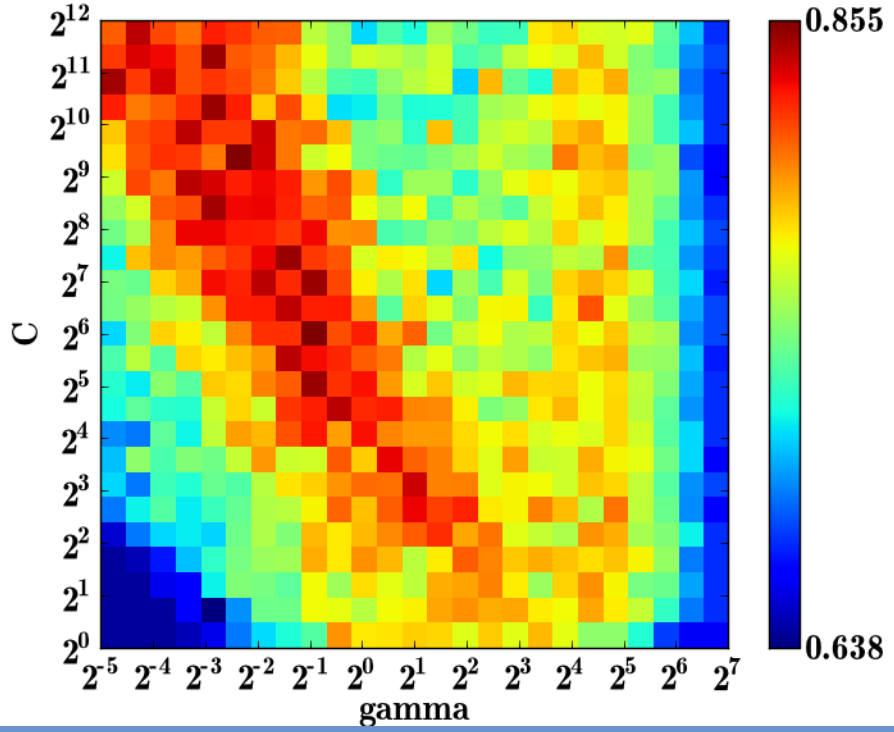
Name	Number of Scene Cuts	Scene Moving Speed	Object Moving Speed	Average Frame Complexity	Content
Bus	0	Middle	Fast	27.95	Bus
Container	0	No	Slow	16.18	Ship with containers
Flower	0	Middle	Slow	22.14	Flower and house
Football	0	Fast	Fast	11.96	Football players
Foreman	0	Slow	Slow	18.73	Portrait and construction
Mother &Daughter	0	No	Slow	12.60	Portrait
Stefan	0	Slow	Middle	23.62	Tennis player
Inception	6	Fast	Fast	18.00	Portrait, construction and etc.

Type	Number
Distortion	**151**
Glitch	**39**
Freezing	**48**
Total	**238**

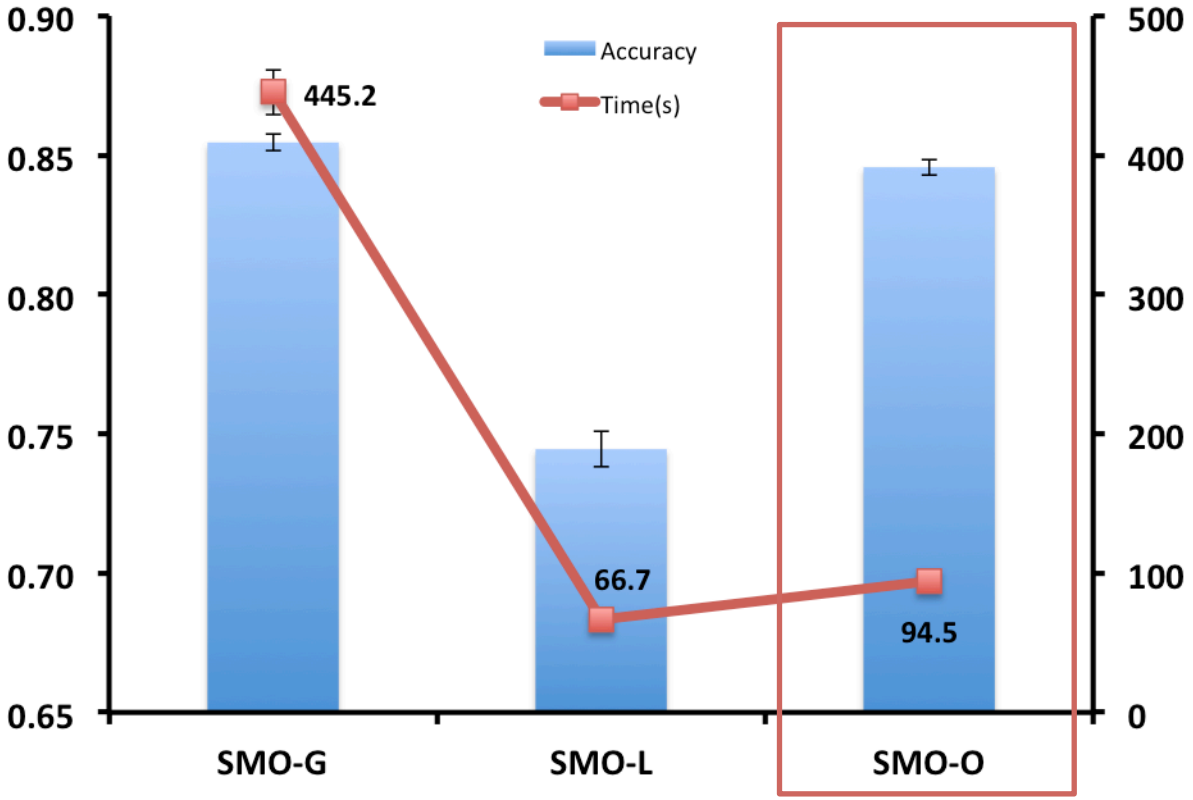
Kernel Selection and Parameter Setting (SVM)

- **Sequential minimal optimization (SMO) (J. Plat et al., 1998)**
  - An algorithm for efficiently solving the optimization problem that arises during t training of SVMs
- **Kernel selection**
  - Linear kernel
  - Radial basis function (RBF) kernel
- **Parameter setting (RBF)**
  - Grid search
  - Optimized method (S. Keerth et al., 2003)

1. SMO-L: SMO with Linear kernel
2. SMO-G: SMO with RBF kernel, and Grid search for parameters
3. SMO-O: SMO with RBF kernel, and Optimized method for parameters



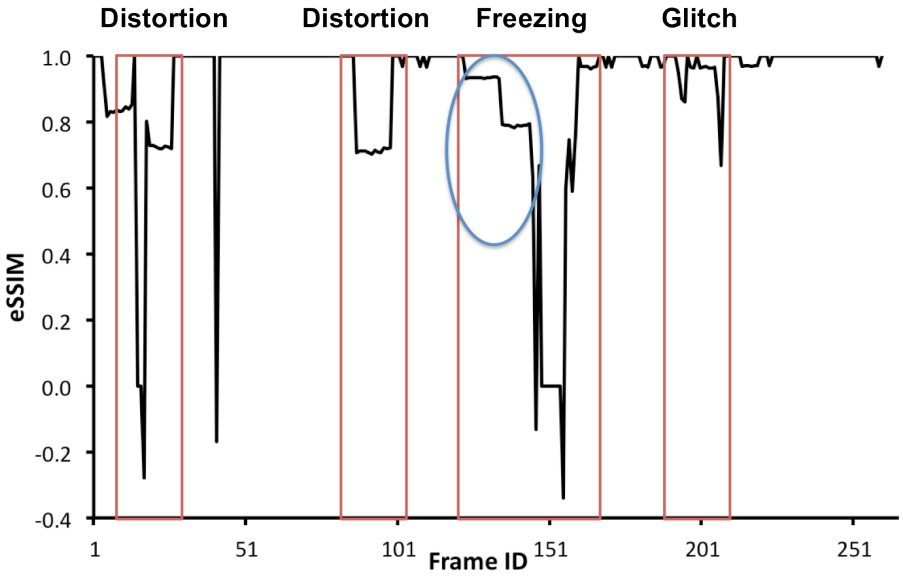
❖ Efficiency of multiclass classification



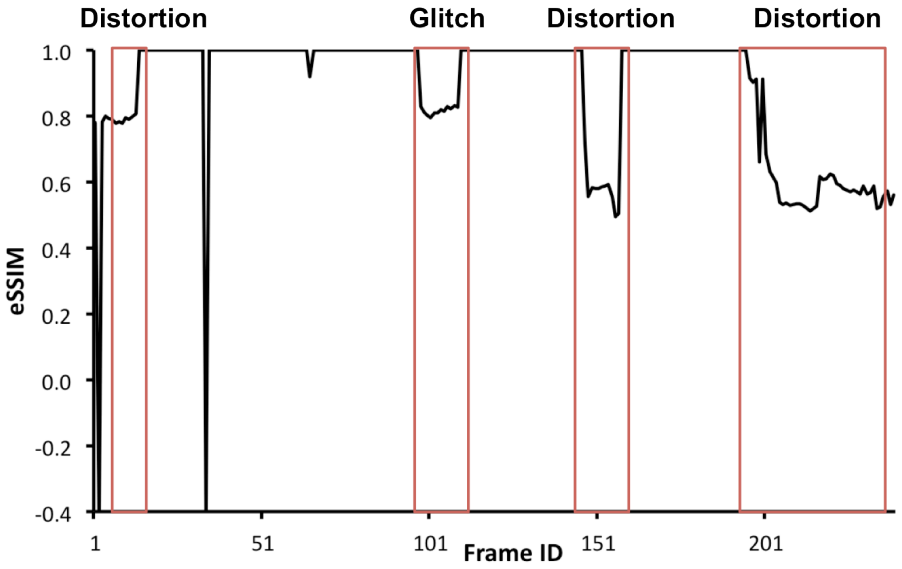
Experiment Result – eSSIM Aggregator (2/3)

❖ Classification result of each binary classifier

	Distortion vs. *Glitch*	*Distortion* vs. *Freezing*	*Glitch* vs. *Freezing*
Accuracy	92.04%±0.24%	85.38%±0.32%	97.18%±0.21%



Foreman, GOP = 12, 2% GE loss model, Frame complexity = 18.73



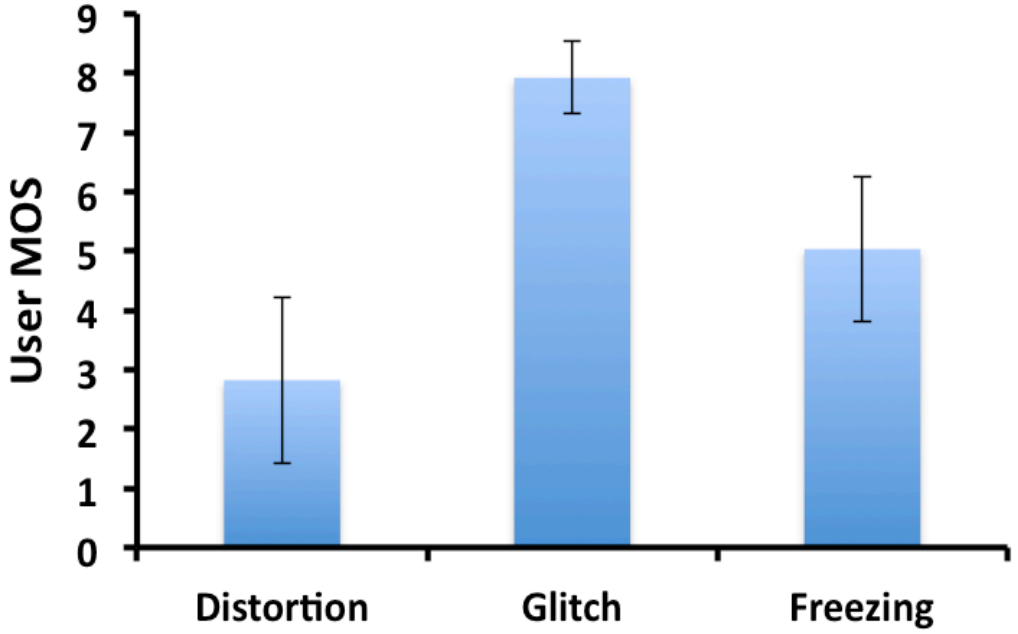
Flower, GOP = 12, 2% uniform loss model, Frame complexity = 22.14

Experiment Result – User Model (1/2)

❖ Subjective testing of user model

MOS	1, 2	3, 4	5, 6	7, 8	9, 10
Quality	Bad	Poor	Fair	Good	Excellent
Impairment	Very annoying	Annoying	Slightly annoying	Perceptible but not annoying	Imperceptible

❖ Average user MOS of defect events



MOS range of Distortion: 1, 2, 3, 4, 5

MOS range of Glitch: 6, 7, 8, 9

MOS range of Freezing: 3, 4, 5, 6, 7

❖ Classification result of user MOS on Distortion

Classified MOS

User MOS

	1	2	3	4	5
1	1	10	4	0	0
2	11	11	11	0	1
3	2	13	21	2	1
4	0	1	5	7	0
5	0	3	11	1	0

(a) {*μ*, *ratio*}

Classified MOS

User MOS

	1	2	3	4	5
1	3	10	2	0	0
2	4	17	11	1	1
3	1	11	19	6	2
4	0	2	6	4	1
5	0	3	8	4	0

(b) {*μ*, *ratio*, *frame complexity*}

Classified MOS

User MOS

	1	2	3	4	5
1	3	8	4	0	0
2	5	17	11	0	1
3	0	10	21	5	3
4	0	2	4	5	2
5	0	2	8	3	2

(c) {*μ*, *ratio*, *frame complexity*, *motion speed*}

	Exact	Drift ±1
(a)	34%	79%
(b)	37%	83%
(c)	41%	83%

Conclusion

❖ Summary

- An event-based model to correlate video quality metrics to user experience, using ML
- Event classification and event-to-experience correlations

❖ Contribution

- The first (to the best of my knowledge) to consider that human experience is event-based
- Event classification (eSSIM aggregator)
  - Provides an accurate classification of defect events with low training time
- Relating to user experience
  - Shows a strong correlation between key features (specific to each event type) and user MOS
- A good method for problem decomposition

Future Work

❖ User MOS estimation for entire video session

- Current user MOS estimation is at the event level
- A number of psychological theories will be used for modeling the complicated human perception system

❖ Reference model of user perception

- Find common response among users to generate reference model
- Identify key user-specific features that specialize reference model to individual user (or groups of like-minded users) MOS


Thank You!

Reference

1) V. Menkovski, A. Oredope, A. Liotta, and A. Cuadra, "Predicting quality of experience in multimedia streaming," In Proceedings of the 7th International Conference on Advances in Mobile Computing and Multimedia, pp. 52-59, 2009.
2) S. Kanumuri, P. C. Cosman, A. R. Reibman and V. A. Vaishampayan, "Modeling packet-loss visibility in MPEG-2 Video," IEEE Transactions on Mulitmedia, vol. 8, pp. 341-355, 2006.
3) M. Narwaria, W. Lin, and A. Liu, "Low-complexity video quality assessment using temporal quality variations," IEEE Transactions on Multimedia, vol. 14, pp. 525-535, 2012.
4) A. Kwon, J. Xiao, S.-S. Seo, J. W.-K. Hong, and R. Boutaba, "The impact of network performance on perceived video quality and user experience in H.264/AVC," in IEEE/IFIP Network Operations and Management Symposium (NOMS), mini-conference, April 2012.
5) A. R. Reibman and D. Poole, "Predicting packet-loss visibility using scene characteristics," in Packet Video 2007, pp. 308-317, November 2007.
6) G. W. Cermak, "Videoconferencing service quality as a function of bandwidth, latency, and packet loss," Verizon Laboratories, T1A1.3/2003-026, 2003.
7) J. M. Zacks, "How we organize our experience into events," American Psychological Association, 2010
8) Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," Signal Process.: Image Commun., vol. 19, no. 2, pp. 121-132, Feb. 2004

Related Work – Machine Learning Techniques

❖ Performance of ML

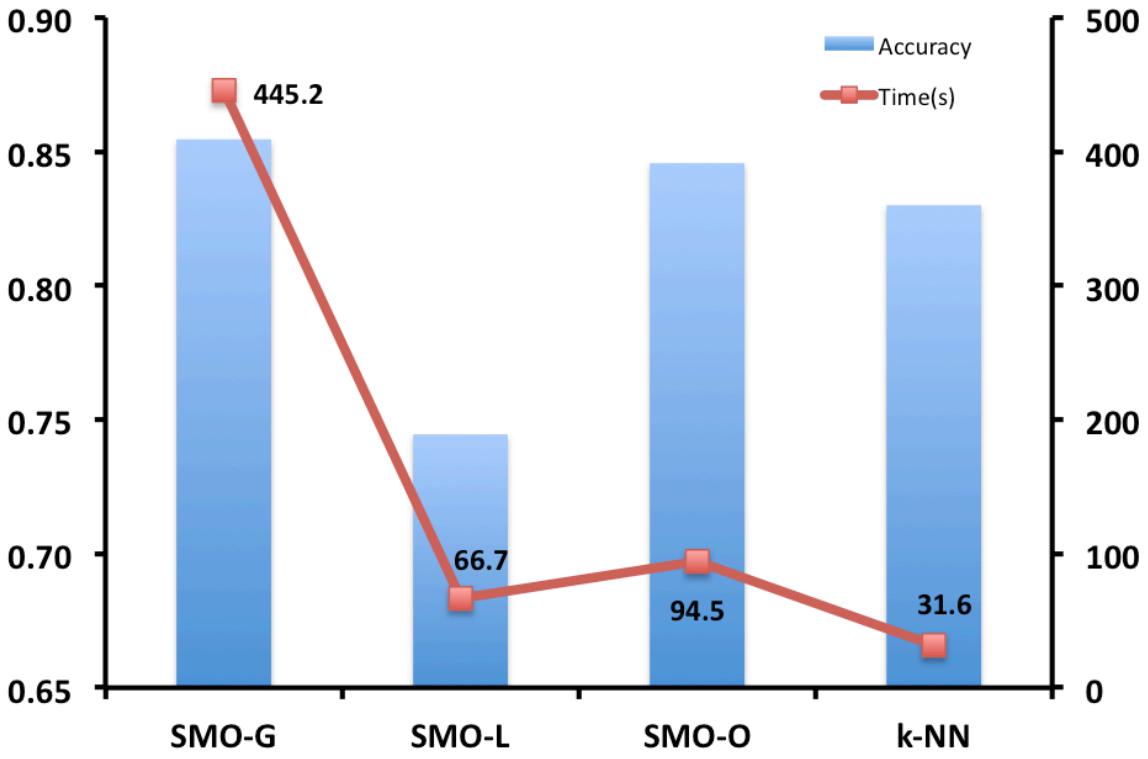
- **Accuracy**: A strong correlation between the input parameters (i.e., **features**) and the output result
- **Efficiency**: The number of features should be small and indicative

❖ Comparison of supervised ML algorithms

- Logic-based algorithms, i.e. decision tree
  - Do not perform well with numerical features
- Perceptron-based algorithms, i.e. artificial neural network
  - Inefficient with the presence of irrelevant features
- Statistical learning algorithms, i.e. k-nearest neighbor (k-NN)
  - Very sensitive to irrelevant features
- Support vector machine (SVM)
  - Performs well with multi-dimension and continuous features, as well as when a nonlinear relationship exists among the input and output features

❖ Event segmentation

- **Rationales:**
  - Human experience is event-based
  - The length of event can be scalable and the boundary between events can be coarse
- **Rules:**
  - The minimum length of a defect event is 10
  - The maximum length of a defective event is 100, about 3 seconds
  - The minimum length of boundary between two events is 10
  - When *eSSIM* >= 0.95, distortion is too small to be perceived. Therefore, we regard these frames as normal ones with *eSSIM* = 1
  - The first 10 frames of each video are ignored, because they can be counted as frames after a scene change


0.90
0.85
0.80
0.75
0.70
0.65
500
400
300
200
100
0
Accuracy
Time(s)
445.2
66.7
94.5
31.6
SMO-G
SMO-L
SMO-O
k-NN

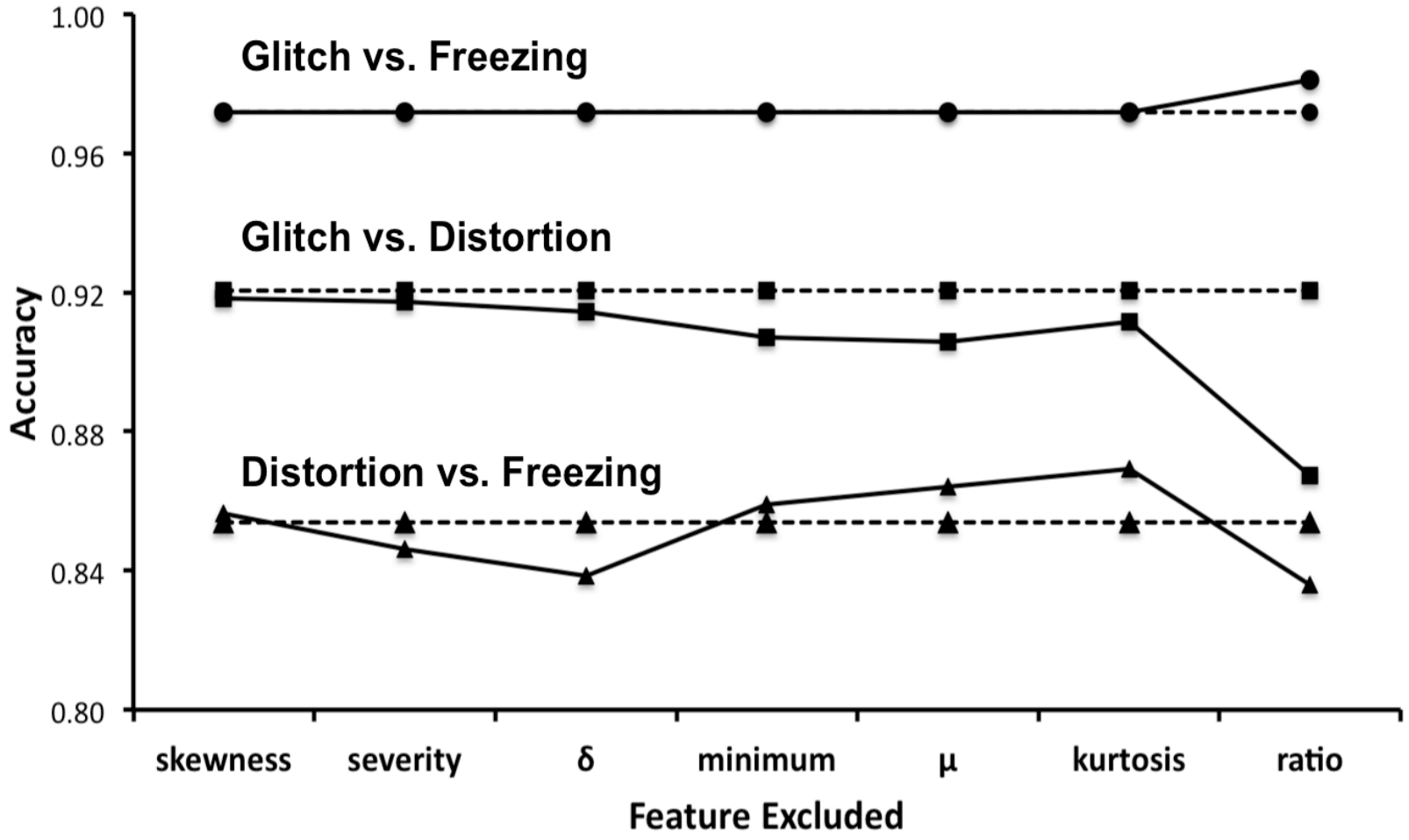
Background (2/2)

- **H.264/AVC**
  - Use spatial and temporal redundancy for compression
- **Measuring user experience**
  - Mean opinion score (MOS)
- **Machine learning (ML) techniques**
  - ML has been widely used in literature to establish correlation among metrics that are:
    - complex to mathematical model or statistically correlate
    - Containing many inter-dependent parameters (difficult to isolate and deduce from ground data)

Experiment Result – eSSIM Aggregator (3/3)

❖ Feature sensitivity

Dashed line: accuracy without feature excluded
Solid line: accuracy with feature excluded



❖ Preprocessing for eSSIM raw data

- Content filters (A. Kwon et al., 2012): luminance, frame complexity, scene change and motion
- Discontinuity filter
  - Frames dropped
  - Frames duplicated

$$eSSIM_{disc} = ssim(frame_b, frame_a) - 1,$$

- $eSSIM_{disc}$ has a range of [-1, 0]
- Points need to mention
  - The process of discontinuity filter is done at the server side
  - Scene change remedy: the first few frames following a scene change are not perceived (A. R. Reibman et al., 2007)

❖ Classification result of user MOS on Glitch and Freezing

Classified MOS

User MOS	6	7	8	9
6	0	0	1	0
7	0	1	4	1
8	0	0	13	0
9	0	1	3	2

(a) Glitch

16/26, 23/26

Classified MOS

User MOS	3	4	5	6	7
3	1	1	0	0	0
4	0	3	4	0	1
5	0	2	8	2	0
6	0	0	0	2	3
7	0	0	0	2	2

(b) Freezing

16/31, 30/31

User Experiment

- A single stimulus procedure
- A random playlist

eSSIM Aggregator (9/)

❖ Features extracted

- μ and δ: $\mu = \frac{\sum_1^n eSSIM_i}{n}$. $\delta = \sqrt{\frac{\sum_1^n eSSIM_i^2}{n} - \mu^2}$.
- *Minimum*: the minimum e*SSIM* value of an event
- *Defective ratio*: $ratio = \frac{N_{eSSIM<0.95}}{n}$.
- Severity of dropped and duplicated frames: $severity = \frac{N_{eSSIM\leq 0.0}}{n}$
- Skewness:
  - Measure of the asymmetric of the probability distribution of event data
  - [-1.0, 0.0, 0.1, 0.5, 0.8, 0.9, 0.95, 0.98, 1.0]
- Kurtosis: measure of whether the data are peaked or flat, relative to a normal distribution

❖ Preprocessing for features

- Middle-range normalization: [-1, 1]
- Mean-std normalization: mean = 0, std = 1

Objective metrics

❖ SSIM

- Full reference metric
- human visual perception is highly adapted for extracting structural information from a scene
- Structural information is the idea that the pixels have strong inter-dependencies especially when they are spatially close

❖ PSNR

- computed by averaging the squared intensity differences of distorted and reference image pixels
- Inconsistent with human eye perception

❖ Statistical approach

- **A. Bhat et al., 2009**

$$MOS_p = 1 - k(MSE),$$

  - **Where MSE is the mean squared error, and k is derived from the spatial edge strength.**
  - **Do not consider distortion caused by varying network conditions**

- **O. Nemethova et al., 2006**

$$\widehat{MOS_{PSNR}}[n] = a \cdot PSNR[n] + b,$$

  - **Where n is frame index, a and b are derived from the relationship between PSNR and MOS**
  - **Frame-level metric, full-reference**

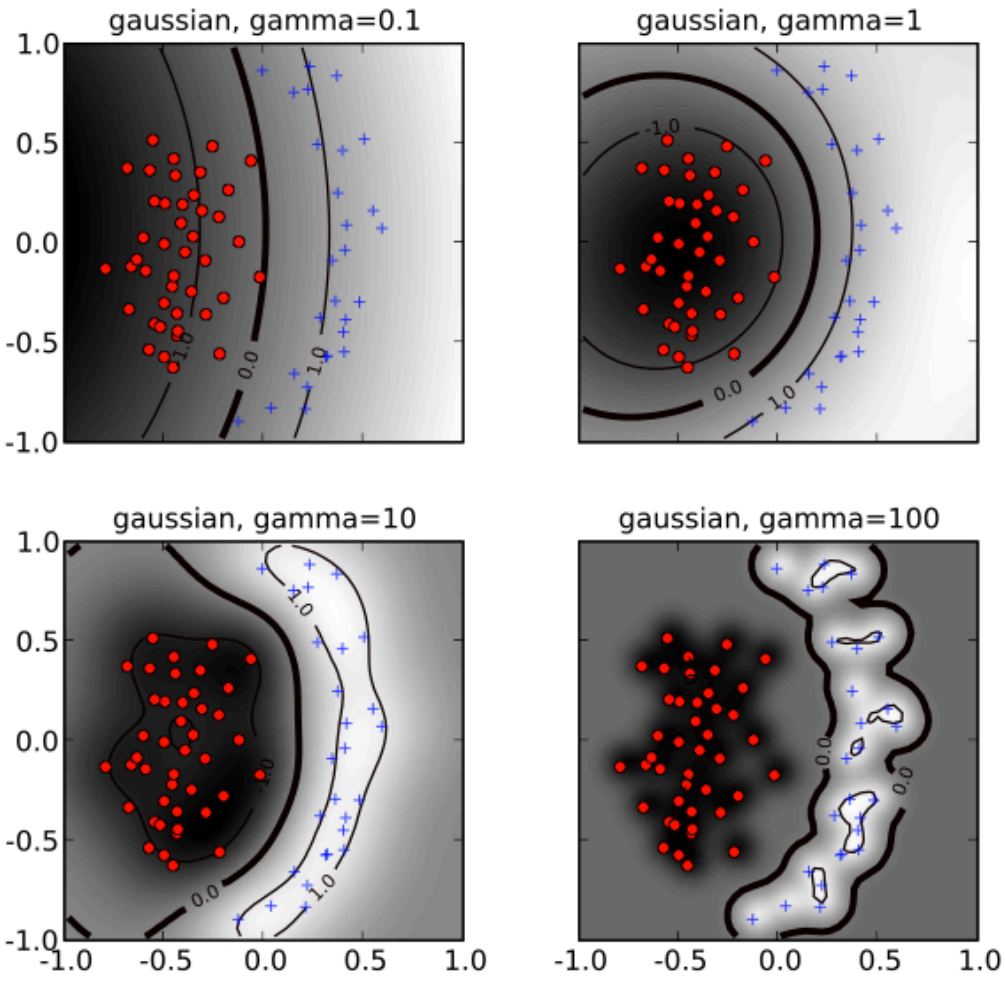
❖ Multiclass classification with SVM

- **Kernel selection**
  - **Radial basis function (RBF)**

$$k(\vec{x}, \vec{x_i}) = exp(-\gamma \|\vec{x} - \vec{x_i}\|^2),$$

- **Combination of binary SVM classifiers**
  - **One-versus-all: *M***
  - **One-versus-one: *M*(*M*-1)/2**

- **Parameters setting:**
  - ***C* and *γ***



A. Ben-Hur, "A user's guide to support vector machines,"