

Tutorial

ACCV2018
2 – 6 December 2018 Perth Western Australia

Facial Micro-Expression Analysis – A Computer Vision Challenge

IV. Recognition

JOHN SEE Multimedia University, Malaysia

ANH CAT LE NGO TrustingSocial

SZE-TENG LIONG Feng Chia University, Taiwan



Outline

Recognition Pipeline

Pre-processing

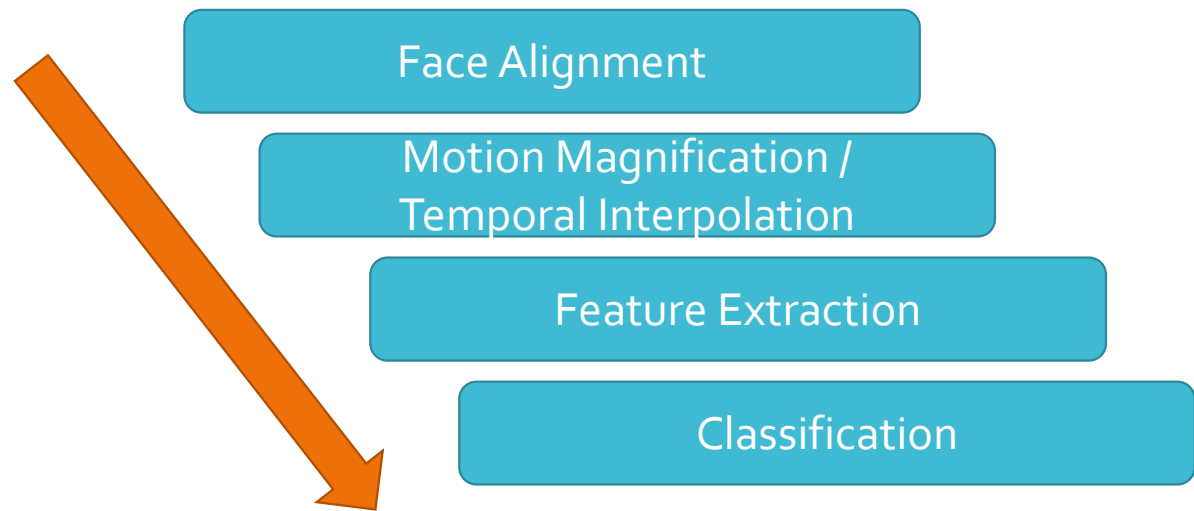
- Data Magnification (spatial)
- Data Interpolation (temporal)
- **Highlighted Work:** Micro-Expression Motion Magnification Global Lagrangian vs. Local Eulerian Approaches

Approaches

- LBP-based Methods
- Optical flow-based Methods
 - **Highlighted Work:** Less Is More: Micro-Expression Recognition from Video using Apex Frame
- Deep learning Methods
- Other Methods

ME Recognition Pipeline

Typically, a ME recognition process will follow these steps:



Pre-processing

Basic Pre-processing steps: Face Alignment, Face Registration, Region partitioning (not mandatory)

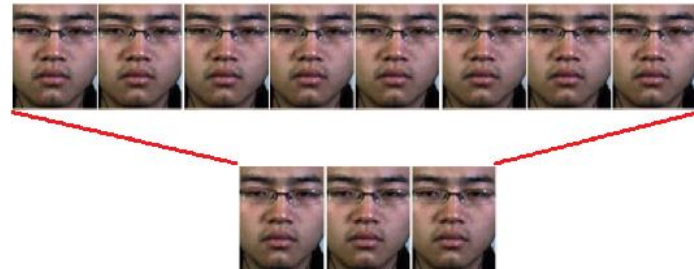
For **RECOGNITION**, 2 essential pre-processing steps:

- **Data Magnification:**

- Amplify or exaggerate facial information spatially → solves the subtleness in ME movements

- **Data Interpolation:**

- Interpolate or extrapolate facial information temporally → solves the unevenness of sample durations, and redundancy (or lack) of information



Motion Magnification

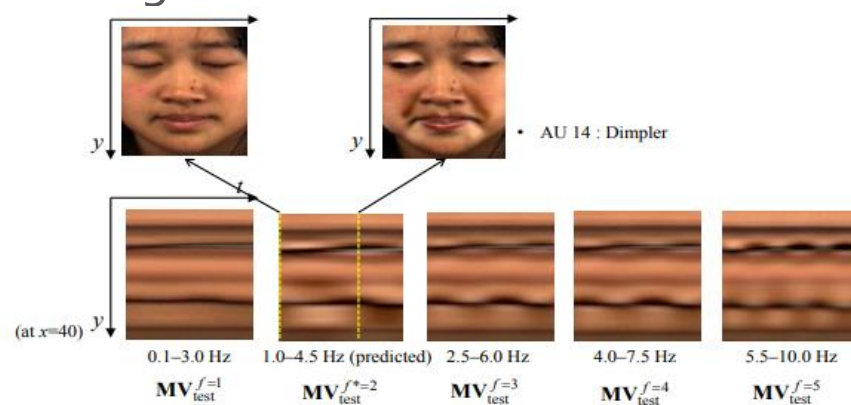
- “Subtleness”: Intensity levels of facial ME movements are very low → extremely difficult to discriminate ME types
- **Eulerian Motion Magnification** (Wu et al. SIGGRAPH 2012)
 - Different spatial frequency bands from decomposed video are band-passed at different spatial levels, and signals are amplified by a magnification factor



<https://people.csail.mit.edu/mrub/vidmag/>

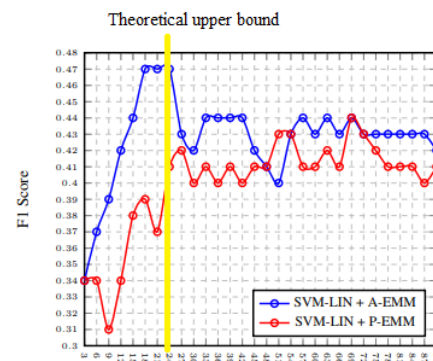
Motion Magnification in ME

- **Park et al. (2014)** – Adaptive selection of most discriminative frequency bands needed before magnification



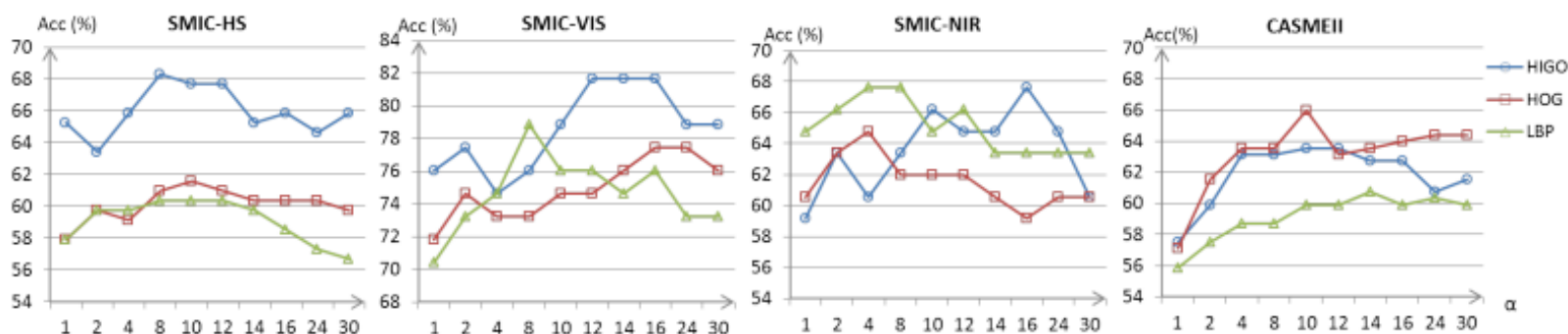
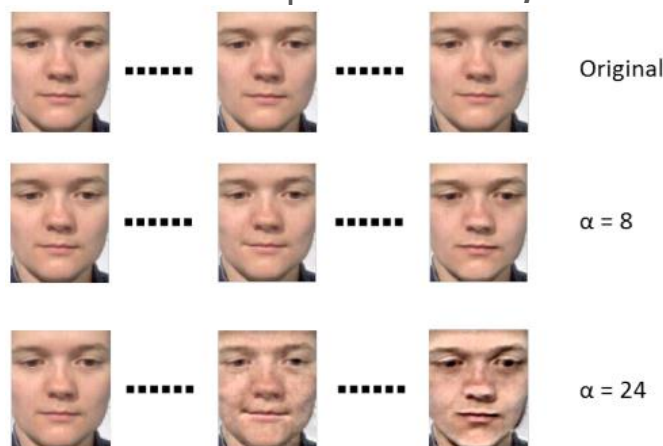
- **Le Ngo et al. (2016)** – Theoretical estimation of upper bounds of effective magnification factors
 - Empirical proof of Wu's proposed bounds w.r.t. spatial cut-off wavelength:

$$(1 + \alpha_{A-EMM}) * \delta(t) < \frac{\lambda_c}{8}$$



Motion Magnification in ME

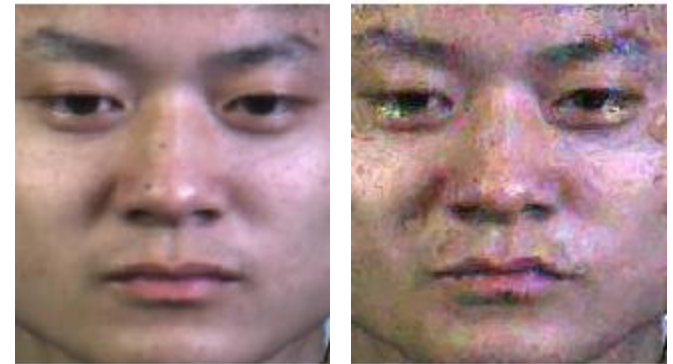
- **Li et al. (2017)** – Demonstrated that EVM can enlarge the difference between different ME categories (inter-class difference) → Recognition rate increases
 - Larger factors cause undesired amplified noise, which degrades performance



Micro-Expression Motion Magnification: Global Lagrangian vs. Local Eulerian Approaches

IEEE FG 2018

Anh Cat Le Ngo, Alan Johnston, Raphael C.W. Phan, **John See**



The University of
Nottingham



2 Perspectives of Motion Magnification

- **“Local” approach:**
 - Modifying intensities of video frames based on frame information
- **“Global” approach:**
 - Synthesizing magnified motion from statistical model of the whole video sequence

Motivation

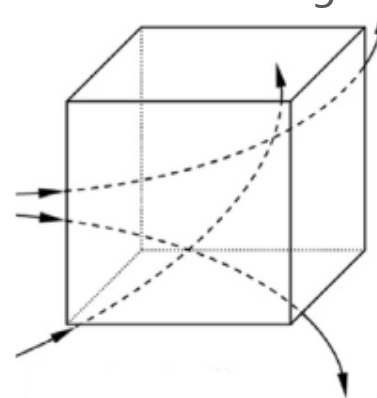
- ✓ Micro-movements can be magnified for better visualization and learning
- ✓ Useful pre-processing step for ME recognition

Local Eulerian Magnification

- **Eulerian dynamics:**

- Assumes existence of regular grid over motion fields. Motion is perceived as changes of properties over local spatial and temporal window
- In Eulerian perspective, motion is a local phenomenon unrelated to global transformation

$$I(x, y, t + 1) = I(x, y, t) + \sum_k \sigma_k \frac{\partial I(x, y, t)}{\partial t}$$



Eulerian Motion Magnification (AEMM)

- Image is decomposed into multiple levels k of dyadic scale space, then magnified with factor σ_k
- Drawback: Risks distorting edge-like features due to limited coverage through time

Lagrangian Dynamics

- **Lagrangian dynamics:**

- Requires estimating motion vectors of points over time
- Apply motion estimation on complete sequences prior to magnification

$$I_{t+1}(x, y, t + 1) = I_t(x + u, y + v, t)$$

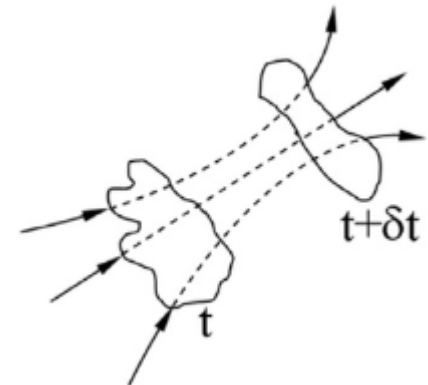
$$I_T = \text{warp}(I_R, \sigma(\mathbf{u}, \mathbf{v}))$$

warp: a synthesis operation from I_R to I_T

I_T : magnified image

I_R : reference image

σ : magnification of motion fields



Lagrangian dynamics: Convenient estimation of global displacement

- Given the displacement (u, v) in 2-D, a target frame can be synthesized by warping the reference frame that has been magnified by multiplication of displacement factor σ .

Global Lagrangian Motion Magnification (GLMM)

- Inspired by **Multi-Channel Gradient Model (McMG)**¹ proposed by Johnston et al:
 - Sharper caricatures of facial expressions² were produced by McMG instead of simple mapping of pixel values
 - Realistic synthesis of facial dynamics
- **Global Lagrangian Motion Magnification (GLMM)**
 - Simply use PCA to learn statistically dominant displacements for all frames
 - Truncation of less significant PCs can filter out insignificant movements

$$(\mathbf{u}, \mathbf{v}) \approx \sum_1^k C_k \text{PC}_k(\mathbf{U}, \mathbf{V})$$

k : principal component of PC_k of the global displacement (\mathbf{U}, \mathbf{V})
 C_k : coefficients of displacement vectors $(u(t), v(t))$

$$I_T = \text{warp}(I_R, \sigma \sum_1^k C_k \text{PC}_k(\mathbf{U}, \mathbf{V}))$$

¹ Johnston et al. (1999). **Robust velocity computation from a biologically motivated model of motion perception**. Proc. of the Royal Society of London B: Biological Sciences

² Nagle et al. (2012). **Techniques for mimicry and identity blending using morph space PCA**. ACCV

Visual Comparison

AEMM



$\sigma = 1$



$\sigma = 5$



$\sigma = 10$



$\sigma = 15$



$\sigma = 20$

GLMM (k=9)



$\sigma = 1$



$\sigma = 5$



$\sigma = 10$



$\sigma = 15$

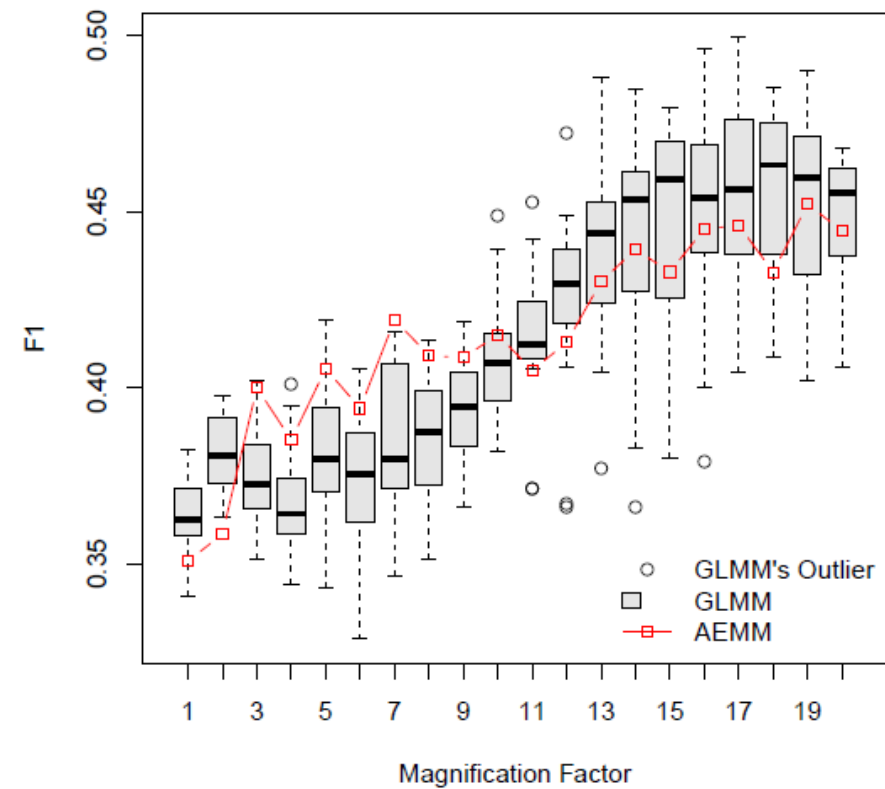


$\sigma = 20$

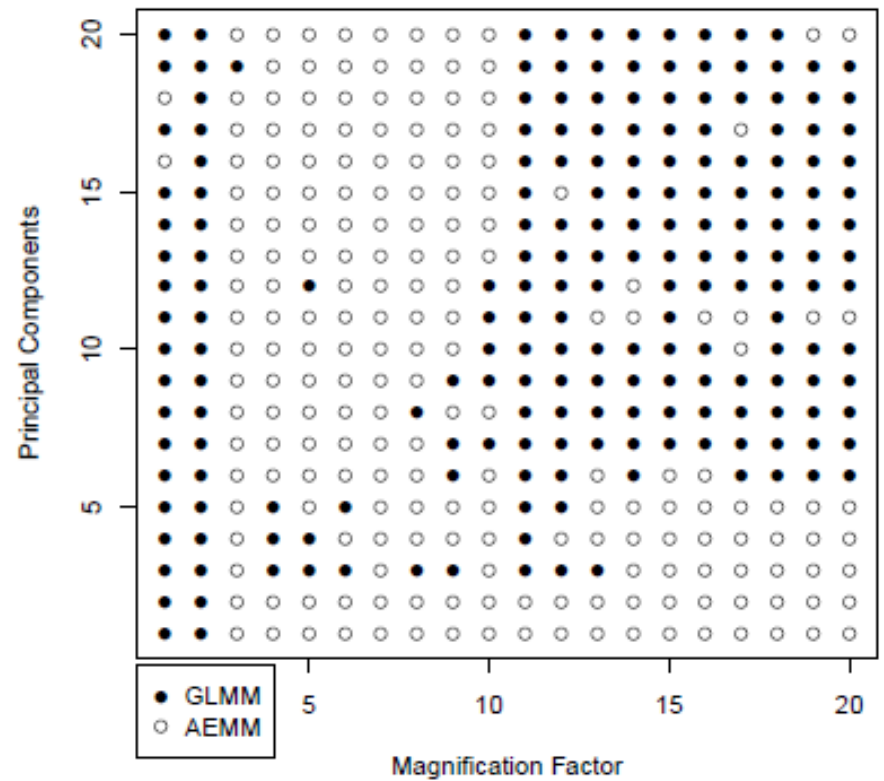
Experimental Setup

- All videos resized to 340x280 pixels, and converted to grayscale for standardization
- Dataset: CASME II, LOSO evaluation protocol
- Feature: LBP-TOP_{4,4,4,1,1,4} on block grid (rotation invariant version)
- Classifier: Linear SVM (C=10000)
- Parameter ranges considered:
 - Magnification factor, $\sigma = [1:20]$
 - # principal components (GLMM only), $k = [1:20]$

Results



Box Plot



Go-Chart

Analysis

Table II
Overall performance of GLMM vs AEMM

Measure	GLMM (PC, σ)		AEMM (σ)	
	Min (1 , 6)	Max (15 , 17)	Min (1)	Max (19)
F1	0.33	0.50	0.35	0.45
Recall	0.32	0.47	0.33	0.41
Precision	0.34	0.54	0.38	0.50

Table III
Performance comparison against state-of-the-art methods on CASME II


Measure	LBPTOP		DiSTLBP-IIP	Bi-WOOF
	Baseline [20]	GLMM	[5]	[12]
F1	0.39	0.50	0.62	0.61
Recall	0.38	0.47	0.60	0.60
Precision	0.41	0.54	0.65	0.63

- **Advantage:**


- Produces more significant amplification of ME movements
- Better recognition result with other choices fixed

- **Disadvantage:**

- Additional free parameter in the number of principal components
- GLMM requires more computations than AEMM



Large improvement over baseline features



Better features but with no magnification applied!

Future testing:
Evaluate these (and many other methods) with AEMM/GLMM

Verdict: Use GLMM More!

- **Reinforces** the benefits of motion magnification towards ME recognition performance
- **Offers** GLMM as an alternative for amplifying subtle changes in MEs
 - GLMM > AEMM provided parameters are tuned
- **Moving forward**
 - More rigorous testing on other settings (features, classifiers)
 - Direct formulation into feature extractors

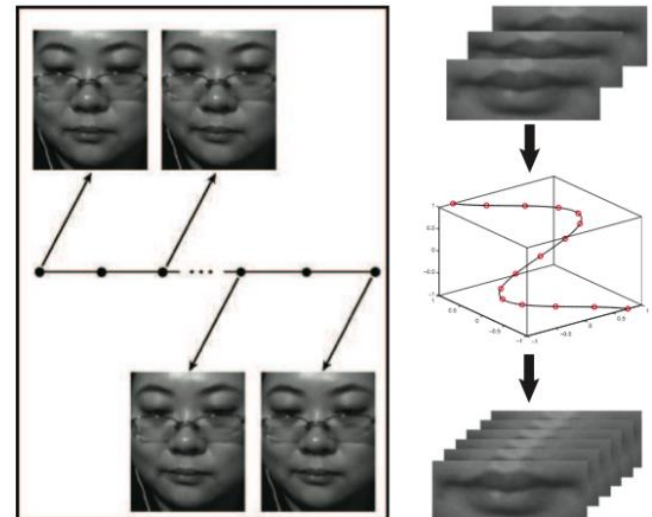
Data Interpolation

- “Redundancy” or “Brevity”: Uneven lengths of ME samples
 - Too short: Insufficient information
 - Too long: Redundant frames can produce poor representations
- **Temporal Interpolation Method (TIM)** (Zhou et al. CVPR2011)
 - Originally proposed for interpolating frames in lip-reading sequences

Basic Idea:

- Interpolate feature vectors to a manifold
- Create new feature vectors by sampling (at uniform intervals) from positions on manifold

Used in SMIC and CASME II baselines



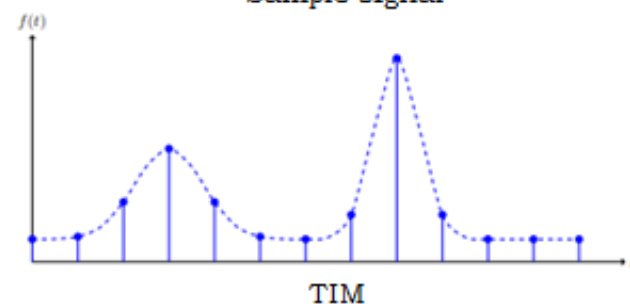
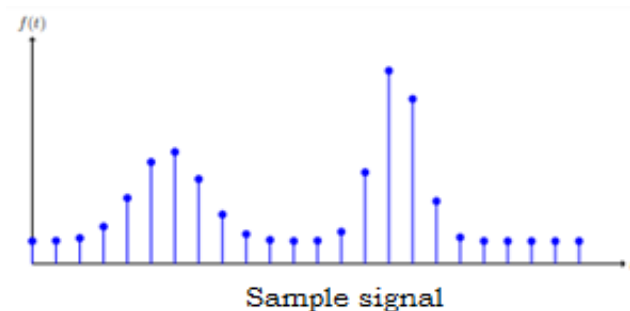
Dynamic selection: Reduce and compress

Interpolation/extrapolation is a “blanket” operation

- Does not consider intrinsic dynamics in each video
- Selection based on # frames does not generalize well to MEs exhibited by different people and emotion types

Intuition: Reduce-and-compress

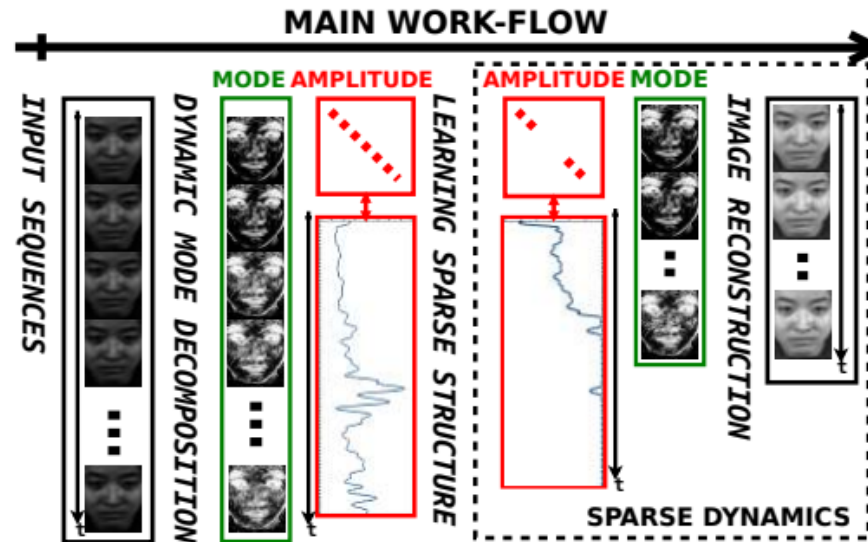
- In speech processing/lip reading, informative samples is more certain after trimming, TIM is acceptable → Interpolation can be done on the originally assumed manifold
- **What we want:** Find informative information based on sparse constraints, and make a reduced size selection (subset selection < number of frames)
- **What we need to make sure:** The informative stuff is preserved! (as well as we can)



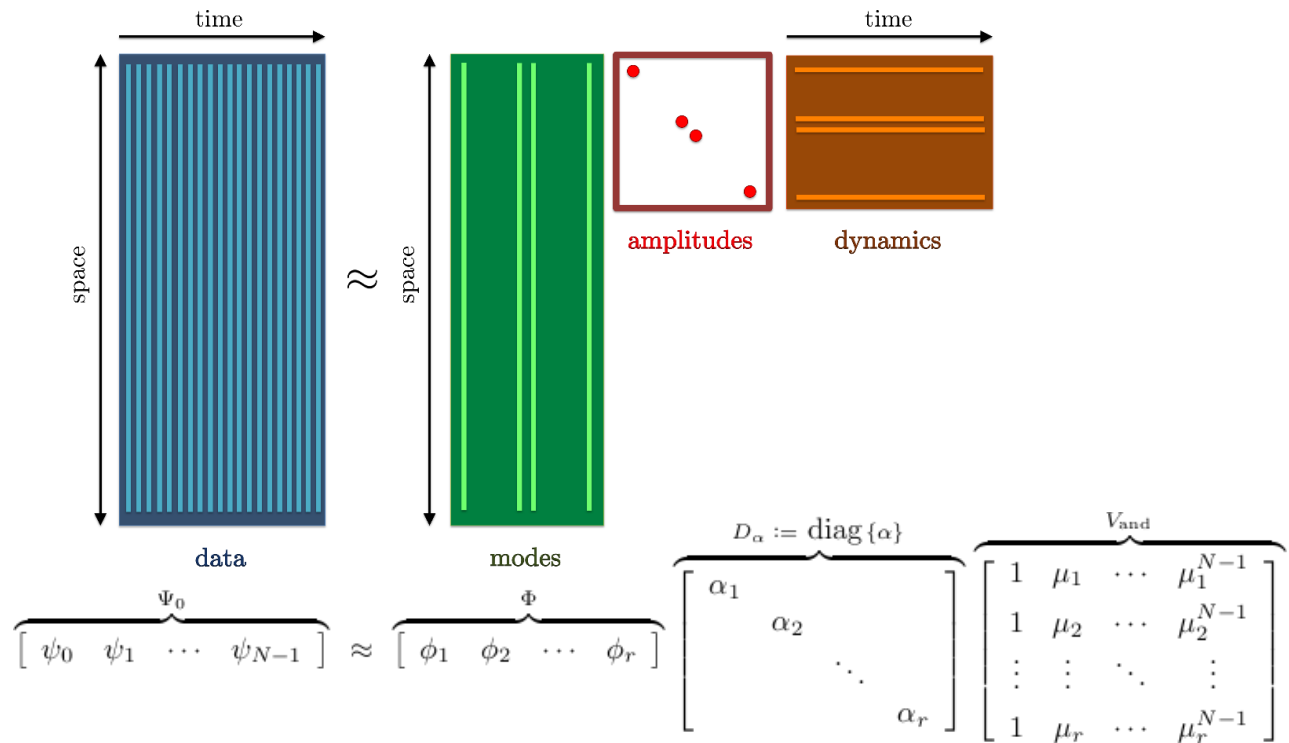
Sparsity-Promoting Dynamic Mode Decomposition (DMDSP)

Basic Idea of DMDSP:

- Decomposition by DMD
- Learn sparse structures (L1) to keep only modes that minimizes loss during reconstruction
- Reconstruct back shorter sequence using the modes



Sparsity-Promoting Dynamic Mode Decomposition (DMDSP)



DMDSP case

$$\arg \min_{\alpha} J(\alpha) + \gamma \sum_{i=1}^r |\alpha_i| \quad \Rightarrow \quad \arg \min_{\alpha} (|J(\alpha)|) \quad \text{s.t.} \quad E^T \alpha = 0$$

$$\alpha_{\text{dmdsp}} = \begin{bmatrix} I & 0 \end{bmatrix} \begin{bmatrix} P & E \\ E^T & 0 \end{bmatrix}^{-1} \begin{bmatrix} q \\ 0 \end{bmatrix}$$

DMDSP+LBP-TOP for ME: Results

	CASME II															SMIC								
	Others (O)			Disgust (D)			Happiness (H)			Surprise (S)			Repression (R)			Negative (N)			Positive (P)			Surprise (S)		
	F1	RR	PR	F1	RR	PR	F1	RR	PR	F1	RR	PR	F1	RR	PR	F1	RR	PR	F1	RR	PR	F1	RR	PR
SS	.56	.64	.50	.27	.23	.32	.58	.55	.62	.67	.52	.93	.39	.41	.38	.53	.48	.59	.60	.70	.53	.64	.62	.67
US	.52	.58	.47	.09	.07	.12	.36	.48	.29	.34	.24	.60	.32	.30	.35	.37	.36	.38	.40	.46	.35	.44	.37	.55
US*	.47	.52	.44	.20	.19	.20	.37	.44	.32	.12	.08	.22	.18	.15	.24	.46	.41	.52	.49	.53	.45	.48	.51	.46
RA	.49	.54	.44	.13	.12	.14	.25	.42	.18	.10	.06	.19	.32	.30	.34	.38	.36	.41	.36	.46	.30	.37	.29	.52
BL	.47	.53	.42	.25	.22	.27	.33	.31	.34	.42	.40	.43	.30	.26	.35	.39	.36	.42	.41	.49	.35	.39	.35	.45



	CASME II				SMIC			
	ACC	F1	RR	PR	ACC	F1	RR	PR
SS	.49	.51	.47	.55	.58	.60	.60	.60
US	.38	.35	.33	.37	.40	.41	.40	.43
US*	.33	.28	.27	.28	.48	.48	.49	.48
RA	.34	.29	.29	.29	.37	.39	.37	.41
BL	.38	.35	.34	.36	.40	.40	.40	.41

SS: Sparse Sampling (Proposed method),

US: Uniform Sampling w.r.t. % length

US*: Uniform Sampling w.r.t. fixed length (150 for CASME II, 10 for SMIC)

RA: Random Sampling w.r.t. % length

BL: Baseline (no changes to original sequence)

	CASME II				SMIC			
	ACC	F1	R	P	ACC	F1	R	P
Sparse Sampling	.49	.51	.47	.55	.58	.60	.60	.60
Huang et al. [10]	.59	.57	.51	.65	.57	.58	.58	.59
Oh et al. [11]	.46	.43	.35	.55	.34	.35	.35	.34
Liong et al. [12]	.42	.38	.36	.41	.53	.54	.55	.53
Wang et al. [33]	.46	.38	.32	.47	.38	.39	.40	.38
Le et al. [4]	.44	.33	.53	.29	.44	.47	.74	.40
Yan et al. [3]	.38	.35	.34	.36	N/A	N/A	N/A	N/A
Pfister et al. [6]	N/A	N/A	N/A	N/A	.40	.40	.40	.41

No fancy feature representation needed, just LBP-TOP!

SOTA when published. Now no longer best

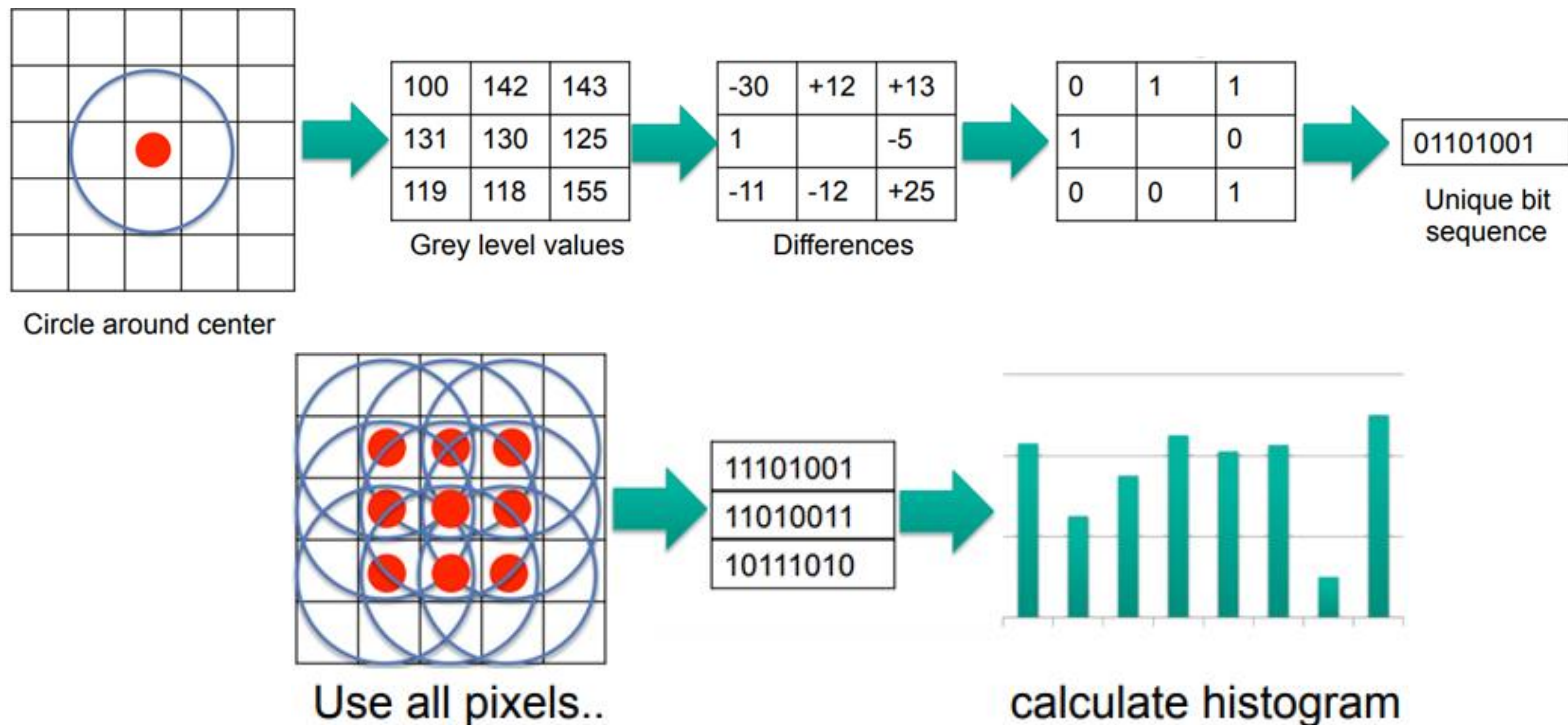


Feature Extraction Techniques

- LBP-TOP, LBP-based methods (texture)
- Optical Flow-based (motion)
- Gradient-based (shape)
- Wavelet representation
- Monogenic signal processing
- Deep representations

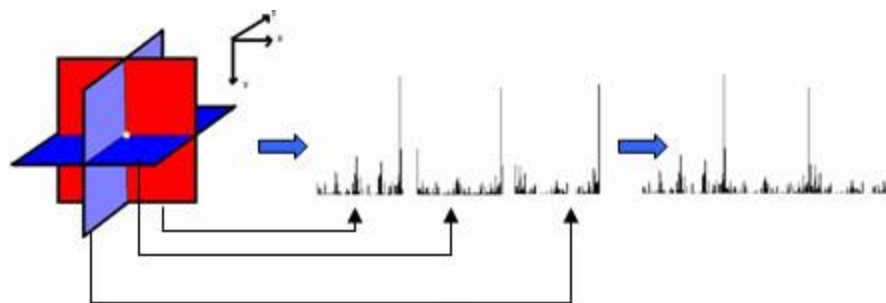
Local Binary Pattern (LBP)

- 2D texture descriptor → describes a particular local texture patch in very compact binary codes
- Popular and proven robust against image variations (rotation, translation, illumination)

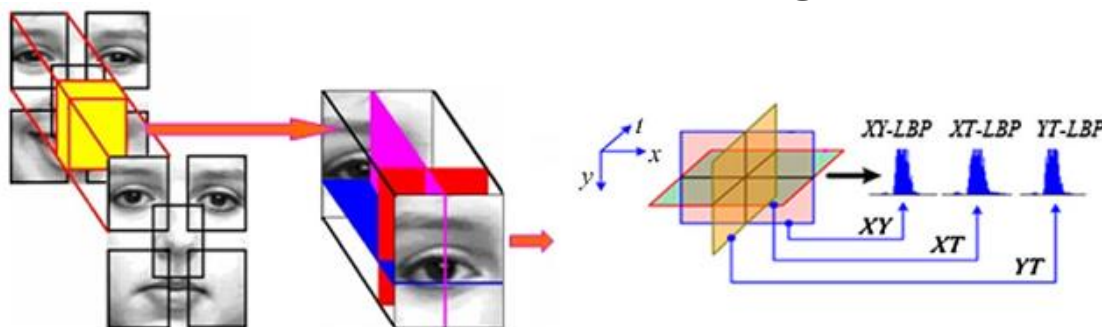


Local Binary Pattern (LBP) on Three Orthogonal Planes (TOP)

- LBP extended to temporal dimension (dynamic texture descriptor)
- Video is seen as a 3D volume
- Simple idea: Apply LBP to all 3 planes in volume (XY, XT, YT), concatenate histograms

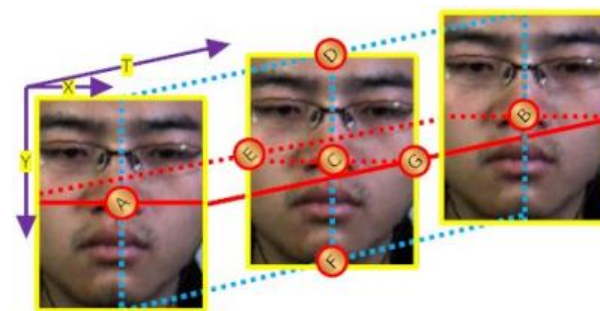
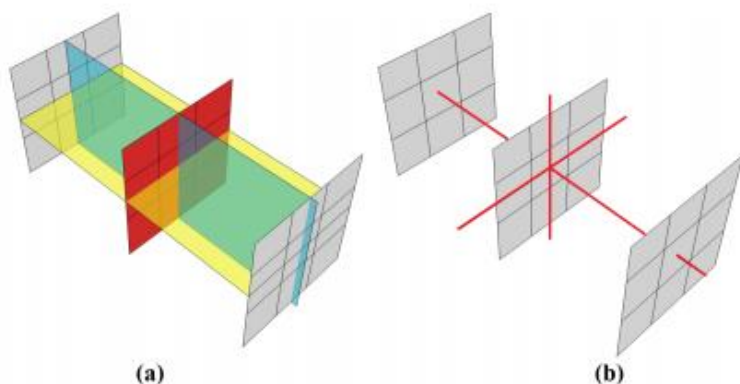


- Block-based LBP-TOP
 - Divide into blocks, each block extracts LBP-TOP histograms, concatenate again



Local Binary Pattern (LBP) on Six Intersection Points (SIP)

- Reduce 3 orthogonal planes to 6 distinct neighbour points (remove all overlapping points considered usually)



Feature extraction time: ~2.8x improvement

Feature dimension: ~2.4x reduction

4-neighbour points set $\{D, E, F, G\}$ for XY
 $\{E, A, G, B\}$ for XT
 $\{D, A, F, B\}$ for YT
 $XY \cap XT \cap YT = \{A, B, D, E, F, G\}$

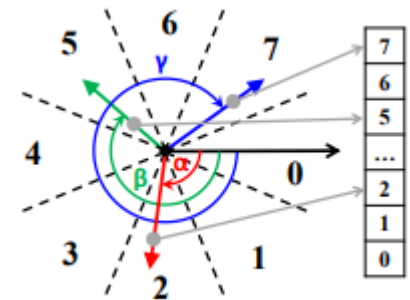
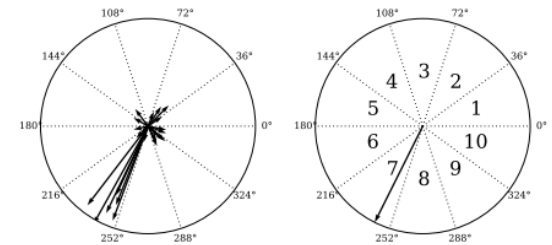
	CASMEII		SMIC	
	LBP-TOP (%)	LBP-SIP (%)	LBP-TOP (%)	LBP-SIP (%)
Linear	62.75	63.56	60.98	64.02
RBF	65.99	66.40	60.98	62.80

Other variants of LBP-TOP for ME

- **LBP-Mean of Orthogonal Planes (MOP)** (Wang et al., 2015)
- **Spatio-Temporal Completed Local Quantized Patterns (STCLQP)** (Huang et al., 2016)
 - Exploit more information: Sign, magnitude and orientation components
 - Codebook reduction
- **Spatio-temporal Local Random Binary Pattern (STRBP)** (Huang & Zhao, 2017)
- **Hot Wheel Patterns (HWP)** (Ben et al. 2017)
 - Encode discriminative features of macro- and micro-expressions
 - Coupled metric learning algorithm to model shared features

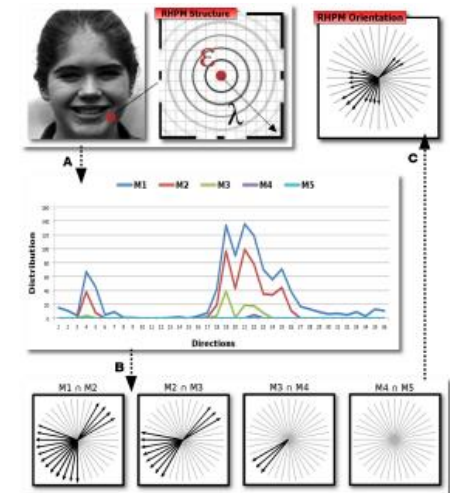
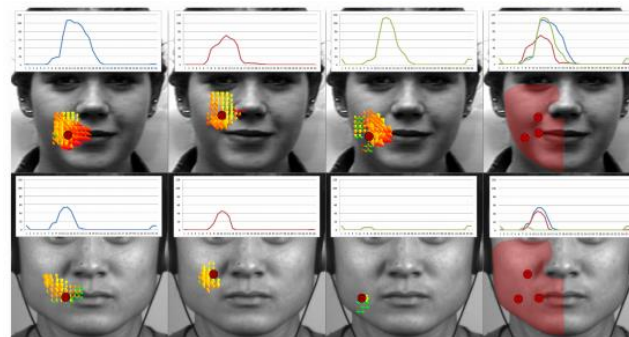
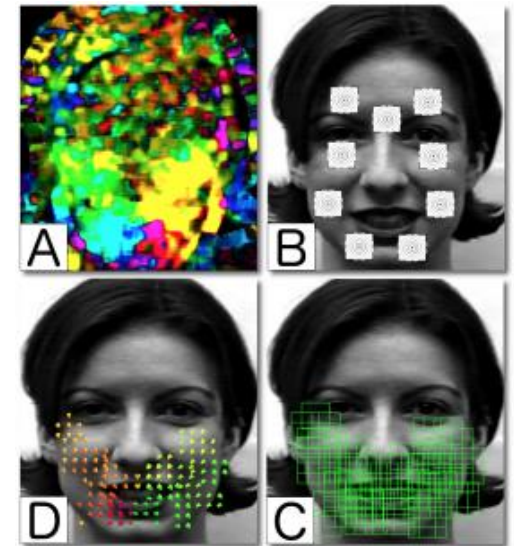
Selective towards principal directions of flow

- Two primary works seek to extract only principal directions of optical flow from ME sequences
 - **Facial Dynamic Map (FDM) (Xu et al., T-AC 2017)**
 - Divide each sequence into spatio-temporal cuboids in a chosen granularity
 - An optimal strategy computes the principal optical flow direction to be used as features
 - **Main Directional Mean Optical-flow (MDMO) (Liu et al, T-AC 2017)**
 - ROI-based normalized statistical feature based on the main direction of the optical flow in polar coordinates
 - 36 ROIs \rightarrow slim feature dimension of only $36 \times 2 = 72$



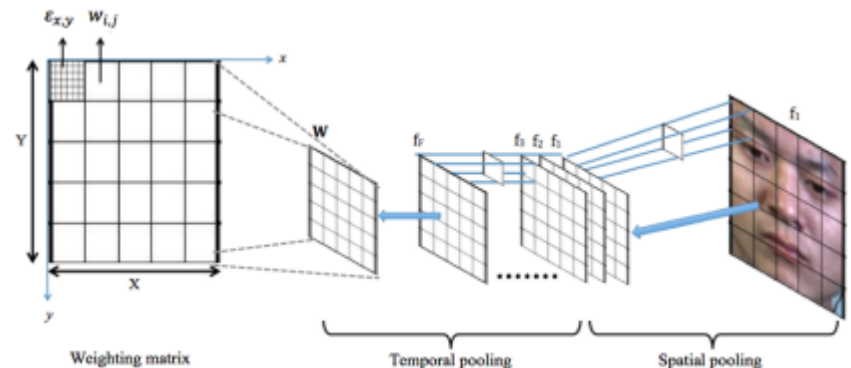
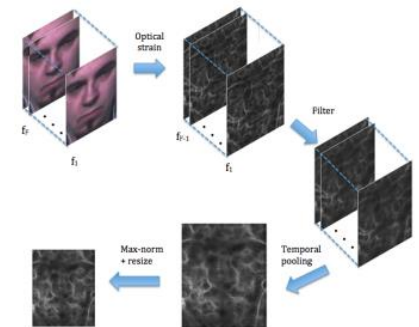
Selective towards regions of consistent flow

- Summary of idea by Allaert et al.
 - Dense Optical Flow (Farneback's) is used to capture local motions based on direction and magnitude constraints → known as Regions of High Probability of Movement or RHPM
 - Each RHPM analyse their neighbours' behaviours in order to estimate the propagation of motion in whole face
 - Filtered optical flow field is computed from each RHPM
 - Facial motion descriptors are constructed from the filtered optical flow field of 25 pre-designated ROIs



Optical Strain

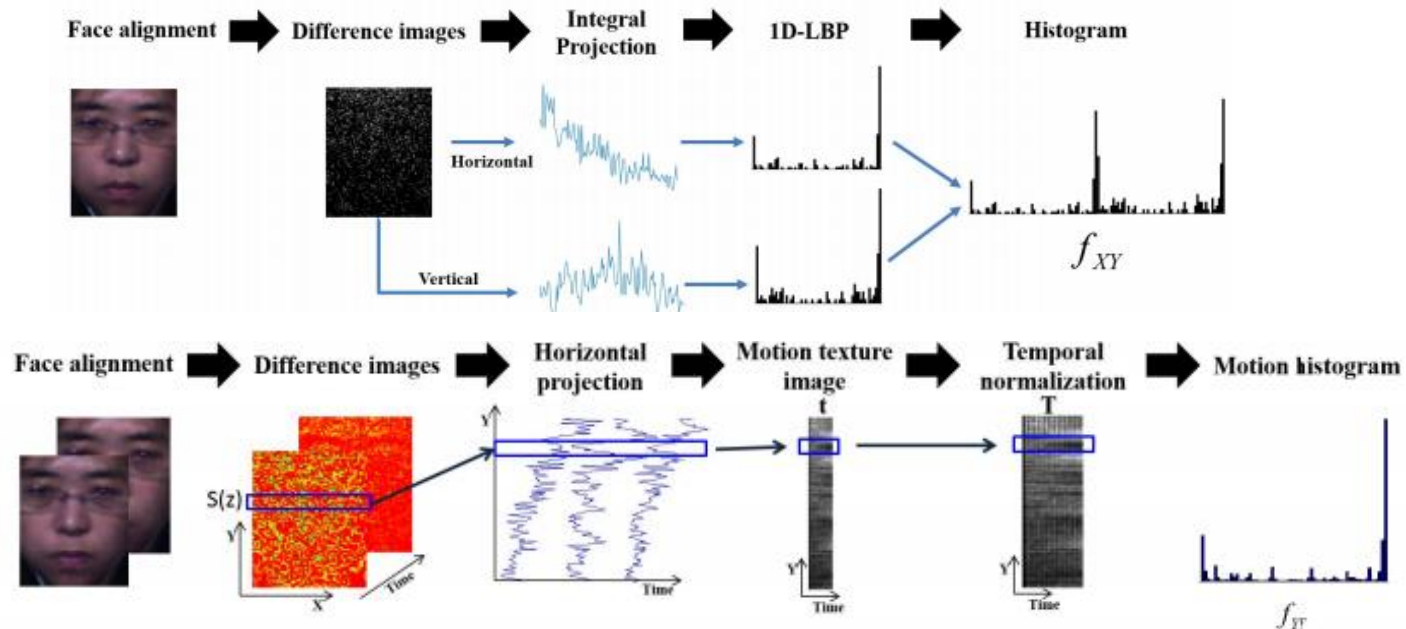
- Motivated by Shreve et al.'s original idea for ME spotting, the work of Liong revolves very much on how Optical Strain (OS) can be exploited for ME recognition
 - Transform OS magnitudes into features (Liong et al. 2014)
 - ➔ magnitudes are pooled temporally to form a single normalized OS map, resized to smaller matrix as feature
- OS-weighted LBP-TOP features (Liong et al., ACCV 2014)
 - ➔ allows regions that exhibit active ME motions to be given more significance, increasing discrimination between emotion types



Constructing histograms from flow

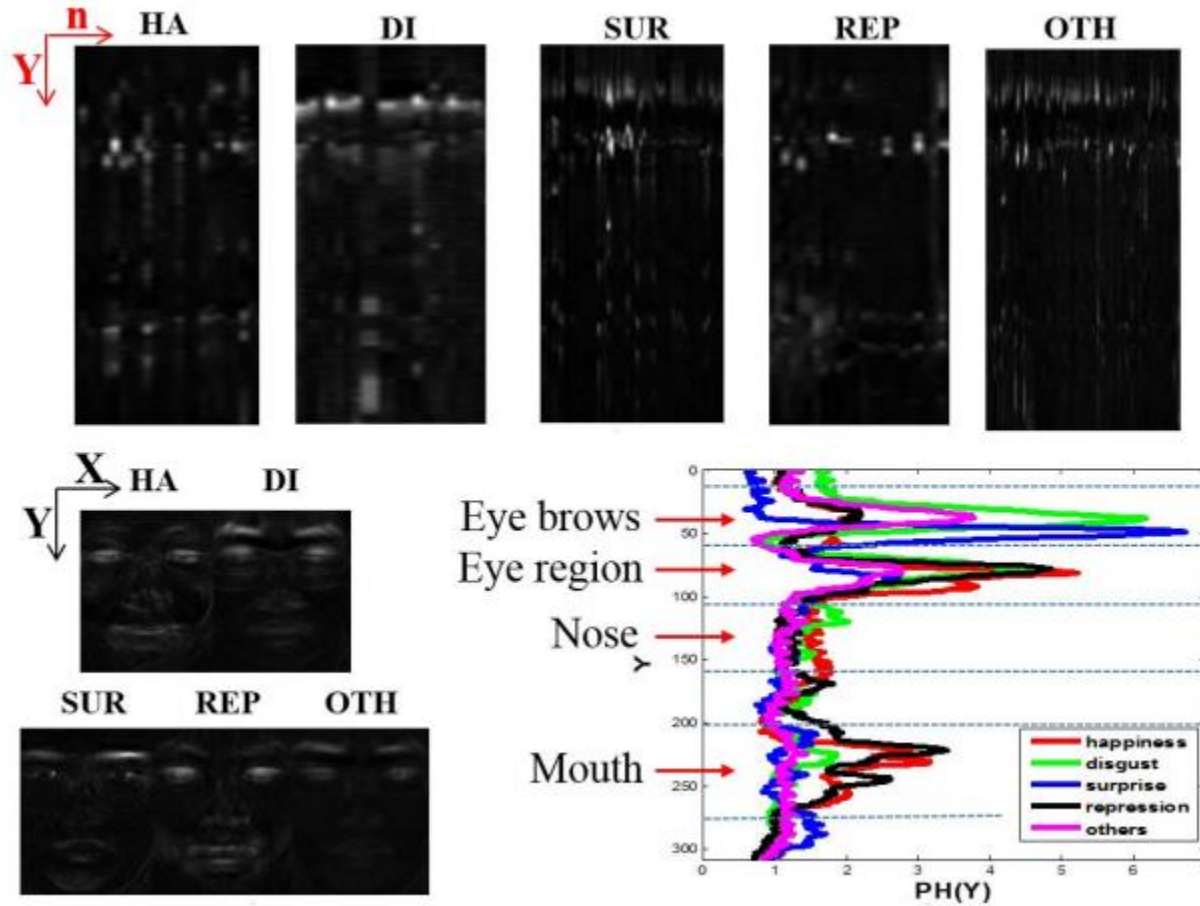
- **Zhang et al., 2017:** Region-by-region Aggregation of Histogram of Oriented Optical Flow (HOOOF) and LBP-TOP to construct rich local statistical features
 - Doing it with ROIs yield even better results than globally done
- **Happy & Routray, 2017:** Fuzzy histogram of optical flow orientations (FHOFO)
 - Assumption: MEs are so subtle that the induced magnitudes can be ignored.
 - Idea: "Fuzzify" the orientation angles to its surrounding bins as such that smooth histograms for motion vector are created

Integral projection



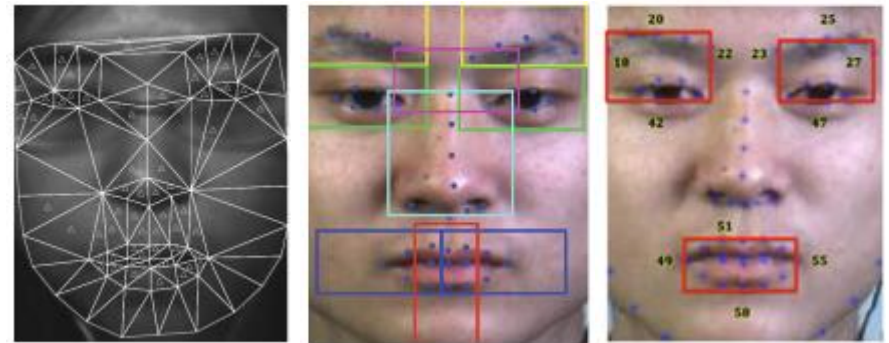
- **Huang et al., ICCV Workshops 2015:** Fuzzy histogram of optical flow orientations (FHOFO)
 - Integral projection based on difference images is used to obtain horizontal and vertical projections
 - Apply 1DLBP operators on both projections to obtain features

Integral projection



ROI-centric methods

- A number of works place priority in locating features at the most salient areas of the face that corresponds strongly to ME motions:
 - **Lu et al., ACCVW 2014**: Use Delaunay triangulation on facial landmark points to obtain 60 ROIs
 - **Zhang et al., MMM 2017**: Use the most representative 9 ROIs from 46 components decomposed from FACS
 - **Liong et al., JSPS 2018**: Use only 3 main ROIs as depicted by the eyes and mouth landmark boundaries



Other feature extractors

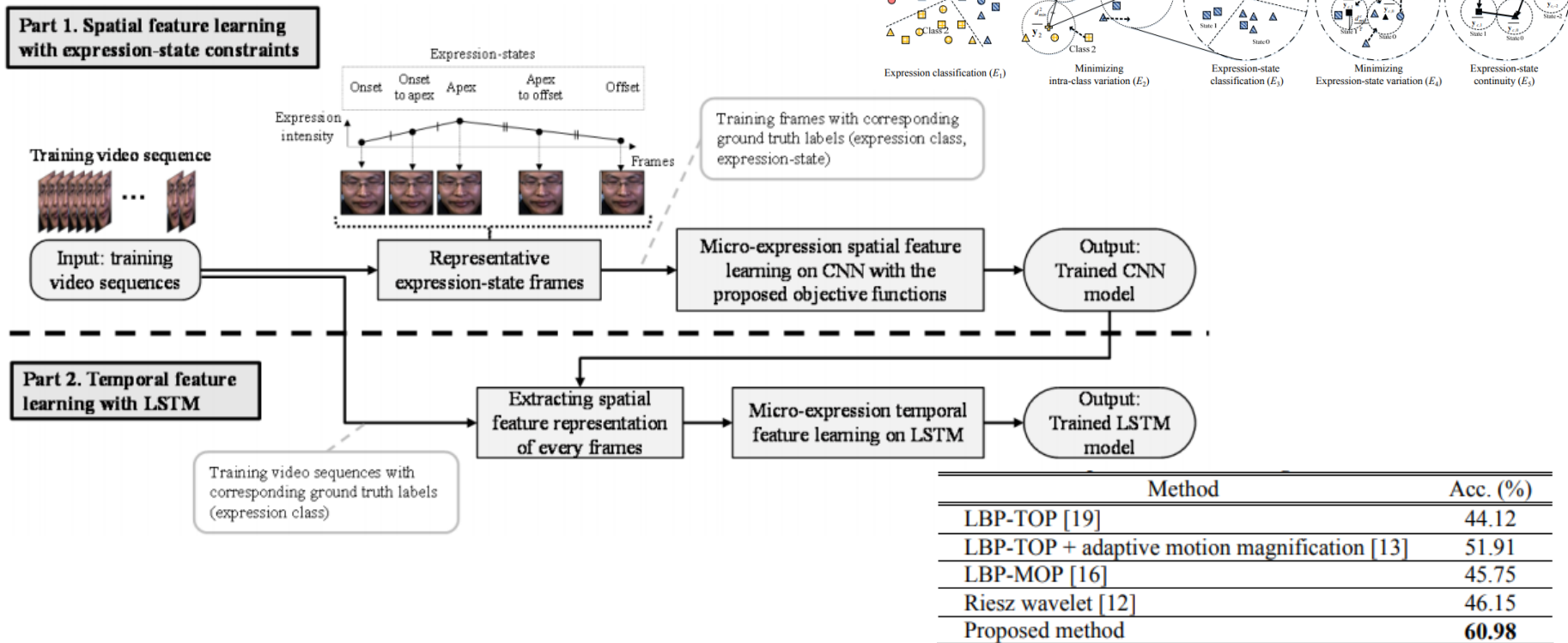
- Riesz wavelet representations
 - **Monogenic Riesz wavelet framework**, Oh et al., 2015
 - **Higher-order Riesz transform**, Oh et al., 2016
- Tensor space features
 - **Tensor Independent Color Space (TICS)**, Wang et al. 2015
 - **Sparse Tensor Canonical Correlation Analysis (STCCA)**, Wang et al., 2016
- Removing latent factors (pose, identity, race, gender)
 - **Robust PCA + Local spatio-temporal directional features**, Wang et al. 2014
 - **Multimodal Discriminant Analysis (MMDA)**, Lee et al. 2017

Deep Learning methods

- Deep Learning methods (needs no introduction here!) have been **slow in adoption** for ME recognition but **gaining some momentum** in recent years.
- **Key problems:**
 - Low number of samples (CASME II: 247, SAMM: 159) → Low in DL standards!
 - Databases have different number of classes (CASME II: 5, SMIC: 3, SAMM: 5, 6 or 7)
 - Existing architectures were built with large-scale natural “in-the-wild” images in mind (ImageNet, Places365, LFW)
- **Some hope:**
 - The closest models that we could find are those trained for face recognition and facial expression recognition.

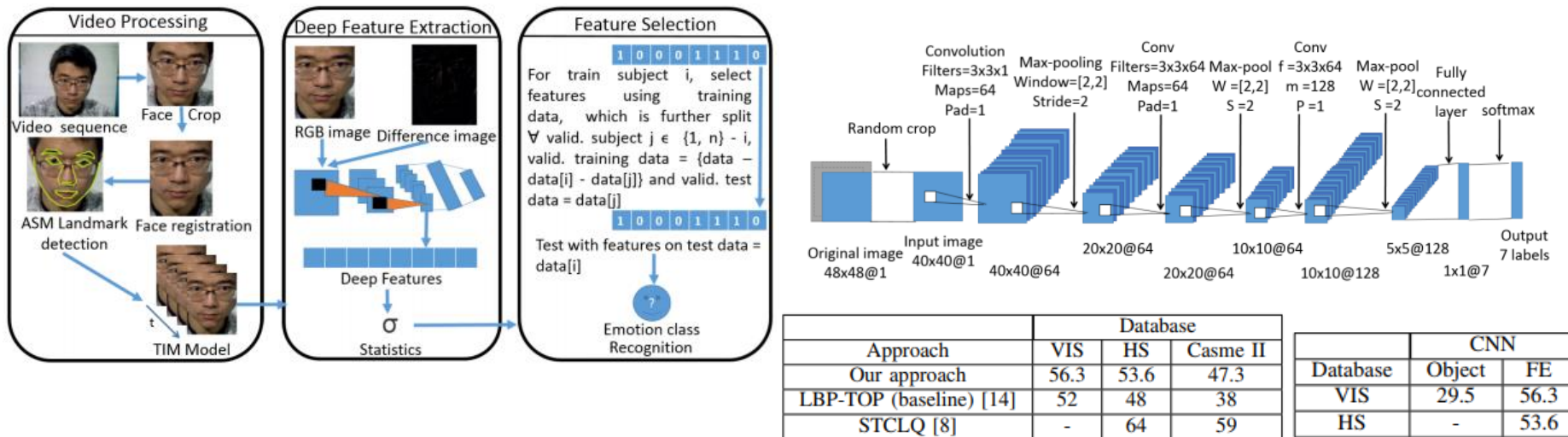
Deep Learning methods

- One of the earliest efforts – Kim et al. (MM 2016):
 - CNN with expression states +LSTM: 5-layer CNN for learning spatial features with expression-states, constrained by 5 objective terms connected to a 2-layer LSTM (512 units each)



Deep Learning methods

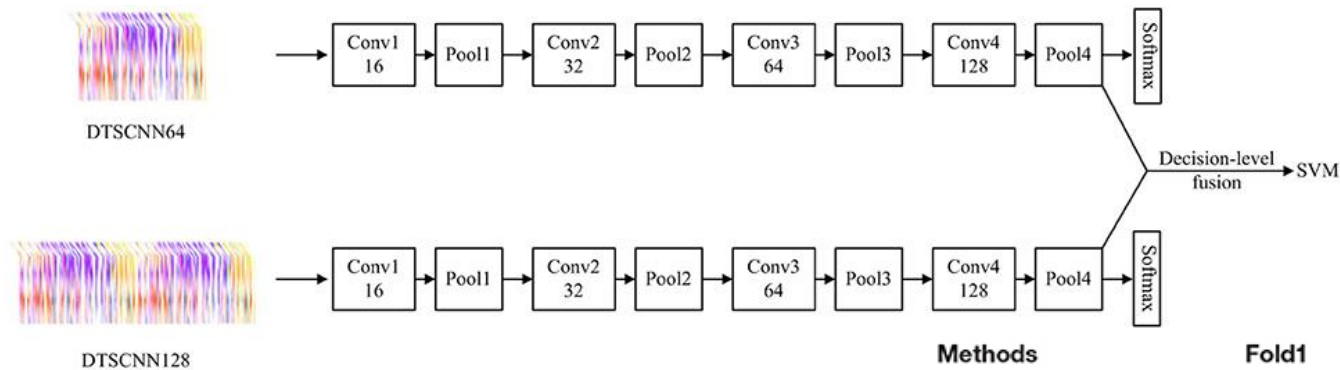
- Another early effort – Patel et al. (ICPR 2016):
 - Transfer learning from existing object and facial expression based CNN models



- Feature selection using evolutionary algorithm
 - ➔ Search for an optimal set of deep features so that it does not overfit training data and generalizes well for test data

Deep Learning methods

- **Dual Temporal Scale CNN – Patel et al. (ICPR 2016):**
 - 2-stream CNN → 64 channel & 128 channel, 5 layers each
 - CNN pre-trained on macro-expression datasets CK+ and SPOS



CASMEI/II	CASMEI	CASMEII
Negative (124)	Disgust (44), sadness (6), fear (2)	Disgust (63), sadness (7), fear (2)
Others (234)	Tense (69), repression (38), contempt (9)	Repression (27), Others (99)
Positive (41)	Happiness (9)	Happiness (32)
Surprise (45)	Surprise (20)	Surprise (25)

Methods	Fold1	Fold2	Fold3	Average
DTSCNN64 TIM64	65.45	65.45	65.45	65.45
DTSCNN128 TIM128	65.45	66.36	65.45	65.75
DTSCNN (fusion)	67.27	67.27	65.45	66.67

3-fold CV!

- Why “dual temporal scale”? CASME I is 6ofps, CASME II 200 fps
- Data selected from CASME I + II, 4 classes (Negative, Others, Positive, Surprise)
- Data augmentation strategy → Produces 20,000 video clips (500 clips / class)

Deep Learning methods

- 2 more DL methods to be discussed in Part 5 when we talk about Micro-Expression Grand Challenge

Classification

- A large majority of works use the standard **SVM** classifier (linear kernel) to classify the extracted features
- Three other notable classifiers (**k-NN**, **Random Forest**, **MKL**) are also used in a few works but very rare (!):
 - Observations: RF and MKL tends to overfit to much of the features used, while k-NN performs quite poorly due to infeasibility for sparse high-dimensional data
- Several works tried dealing with the sparseness by proposing:
 - **Relaxed K-SVD** (Zheng et al., 2016)
 - **Sparse representation classifier (SRC)** (Zheng, 2017)
 - **Kernelized GSL** (Zong et al, 2018)
 - **Extreme Learning Machine (ELM)** (Adegun & Vadapalli, 2016)
- Deep learning methods mainly rely on the **softmax layer** to classify, since they can be trained end-to-end with feature learning

Evaluation Protocol & Performance Metrics

- **Leave-One-Subject-Out (LOSO) cross-validation:** ME datasets are collected from different subjects → The subjects form groups that can be “held-out” to avoid identity bias.
 - First discussed and analysed in-depth by Le Ngo et al. (2014)
 - Some early papers reported LOVO (leave-one-video-out), but primarily almost everyone uses LOSO now 😊
- **Performance Metrics**
 - Typically many works still report the Accuracy metric, which tends to be bias in ME datasets which are naturally imbalanced
 - We advocate the use of F1-score (can be either micro-averaged or macro-averaged) to provide a better reflection of performance

$$F1\text{-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

ME Recognition State-of-the-art

LOSO							
Li et al., 2013	–	LBP-TOP	SVM	–	48.78	–	–
Liong et al., 2016a	–	OSF + OS and weighted LBP-TOP	SVM	–	52.44	–	–
Liong et al., 2014a	–	OS	SVM	–	53.56	–	–
Liong et al., 2014b	–	OS weighted LBP-TOP	SVM	42.00	53.66	0.38	0.54
La Ngo et al., 2014	–	STM	Adaboost	43.78	44.34	0.3337	0.4731
Wang et al., 2015b	–	LBP-MCP	SVM	44.13	50.61	–	–
Xu et al., 2017	–	Facial Dynamics Map	SVM	45.93	54.88	0.4053	0.538
Oh et al., 2016	–	Monogenic + LBP-TOP	SVM	–	–	0.41	0.44
Oh et al., 2015	–	Riesz wavelet + LBP-TOP	SVM	–	–	0.43	–
Liong et al., 2018	ROIs	LBP-TOP	SVM	46.00	54.00	0.32	0.52
Wang et al., 2014b	–	LBP-SIP	SVM	46.56	44.51	0.448	0.4492
La Ngo et al., 2016a	A-EMM	LBP-TOP	SVM	–	–	0.51	–
La Ngo et al., 2016b	DMDSP	LBP-TOP	SVM	49.00	58.00	0.51	0.60
Park et al., 2015	Adaptive MM	LBP-TOP	SVM	51.91	–	–	–
Happy and Routray, 2017	–	HFOFO	SVM	56.64	51.83	0.5248	0.5243
Liong et al., 2018b	–	Bi-WOOF	SVM	–	–	0.56	0.53
Huang et al., 2016	–	STCLQP	SVM	58.39	64.02	0.5836	0.6381
Huang et al., 2015	–	STLBP-IP	SVM	59.51	57.93	0.57*	0.58*
Liong et al., 2016c	–	Bi-WOOF (apex frame)	SVM	–	–	0.61	0.62
He et al., 2017	–	MMFL	SVM	59.81	63.15	–	–
Kim et al., 2016	–	CNN + LSTM	Softmax	60.98	–	–	–
Liong and Wong, 2017	–	Bi-WOOF + Phase	SVM	62.55	68.29	0.65	0.67
Zhang et al., 2016	–	LBP-TOP	RK-SVD	63.25	–	–	–
Zong et al., 2018a	–	Hierarchical STLBP-IP	KGSL	63.83	60.78	0.6110	0.6126
Huang and Zhao, 2017	TIM	STREP	SVM	64.37	60.98	–	–
Huang et al., 2017	–	Discriminative STLBP-IP	SVM	64.78	63.41	–	–
Allaert et al., 2017	–	OF Maps	SVM	65.35	–	–	–
Li et al., 2017	TIM+EVM	HGO	SVM	67.21	68.29	–	–
Zhang, 2017 ††	–	2DSGR	SRC	–	71.19	–	–
Li et al., 2016 †	–	MEMO	SVM	67.37	80.00	–	–
Deaton et al., 2017 †	–	HCOF	SVM	76.60	–	0.55	–
LOVO							
Wang et al., 2015a ††	TIM	LBP-TOP on TICS	SVM	62.30	–	–	–
Yan et al., 2014a	–	LBP-TOP	SVM	63.41	–	–	–
Wang et al., 2014a	TIM	DLSTD	SVM	63.41	68.29	–	–
Happy and Routray, 2017	–	HFOFO	SVM	64.06	56.10	0.6025	0.5536
Liong et al., 2014b	–	OS weighted LBP-TOP	SVM	65.59	–	–	–
Wang et al., 2015b	–	LBP-MCP	SVM	66.80	60.98	–	–
Wang et al., 2014b	–	LBP-SIP	SVM	67.21	–	–	–
Fing et al., 2016	–	LBP-TOP	GSLSR	67.89	70.12	–	–
Park et al., 2015	Adaptive MM	LBP-TOP	SVM	69.63	–	–	–
Wang et al., 2017	EVM	LBP-TOP	SVM	75.30	–	–	–
Li et al., 2017	TIM+EVM	HGO	SVM	78.14	75.00	–	–

Did not follow
standard protocol

ME Recognition State-of-the-art

SOTA methods	Group	Accuracy (CASME II)
He et al. (2017)	SYSU	59.81
Kim et al. (2016)	KAIST	60.98
Liong et al. (2017)	MMU	62.55
Zong et al. (2018)	SEU	63.83
Allaert et al. (2017)	Lille	65.35
Li et al. (2017)	Oulu	67.21

← CNN+LSTM

SOTA

- **Quo Vadis DL?**

- Can DL methods provide the leap forward?
- Can DL methods be assisted through other means (e.g. more data) to achieve SOTA?
- Can DL architectures be better designed (shallower? wider?) to accommodate

ME Recognition State-of-the-art

SOTA methods	Group	Accuracy (CASME II)
He et al. (2017)	SYSU	59.81
Kim et al. (2016)	KAIST	60.98
Liong et al. (2017)	MMU	62.55
Zong et al. (2018)	SEU	63.83
Allaert et al. (2017)	Lille	65.35
Li et al. (2017)	Oulu	67.21
?	?	Close to 70



SOTA

- **Work underway for new DL methods**
 - “Shallower” deep neural network
 - Good choice of input (grayscale is not sufficiently discriminative)
 - Multiple stream learning

Less Is More: Micro-Expression Recognition from Video using Apex Frame

Signal Processing: Image Communication, 2018

Sze-Teng Liong, John See, KokSheik Wong, Raphael C.W. Phan





Prima facie

I. The apex frame is the **most important** frame in the micro-expression clip

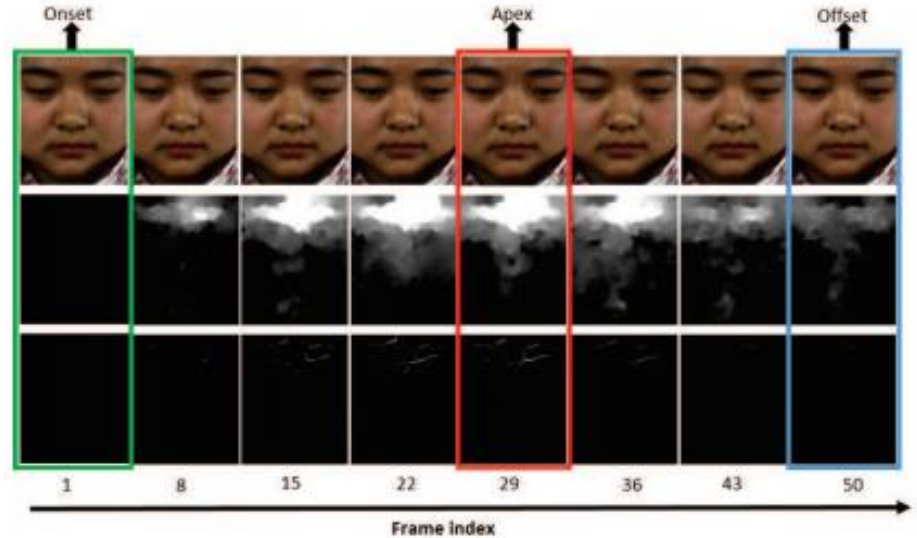
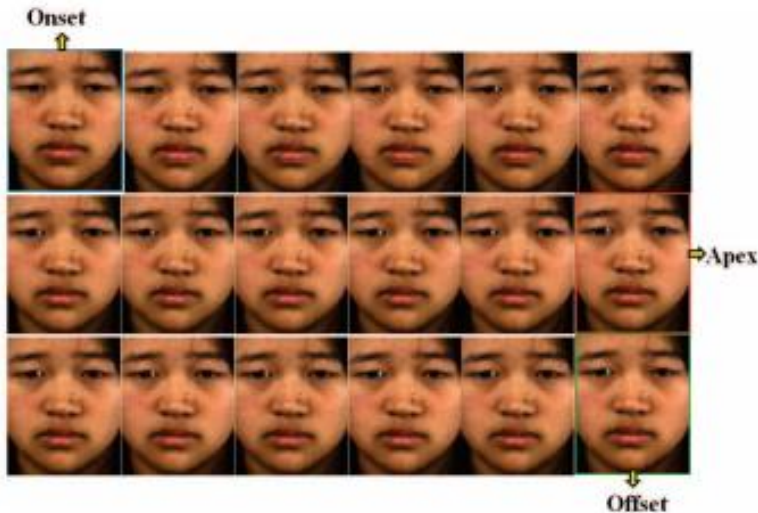
- ✓ Ekman: Emotions are characterised by the change in facial contraction.
- ✓ Exposito: Visual information (video) conveys poor emotional information, due to cognitive overload.

II. The apex frame is **sufficient** for micro-expression recognition

- ✓ “Less is more”? Could too much data clouding the ability to create good feature representations?
- ✓ If performance with one frame is as good as using a full sequence, computation cost can be saved.

The apex should then contain the strongest change in facial movements, and we can also reduce redundancy

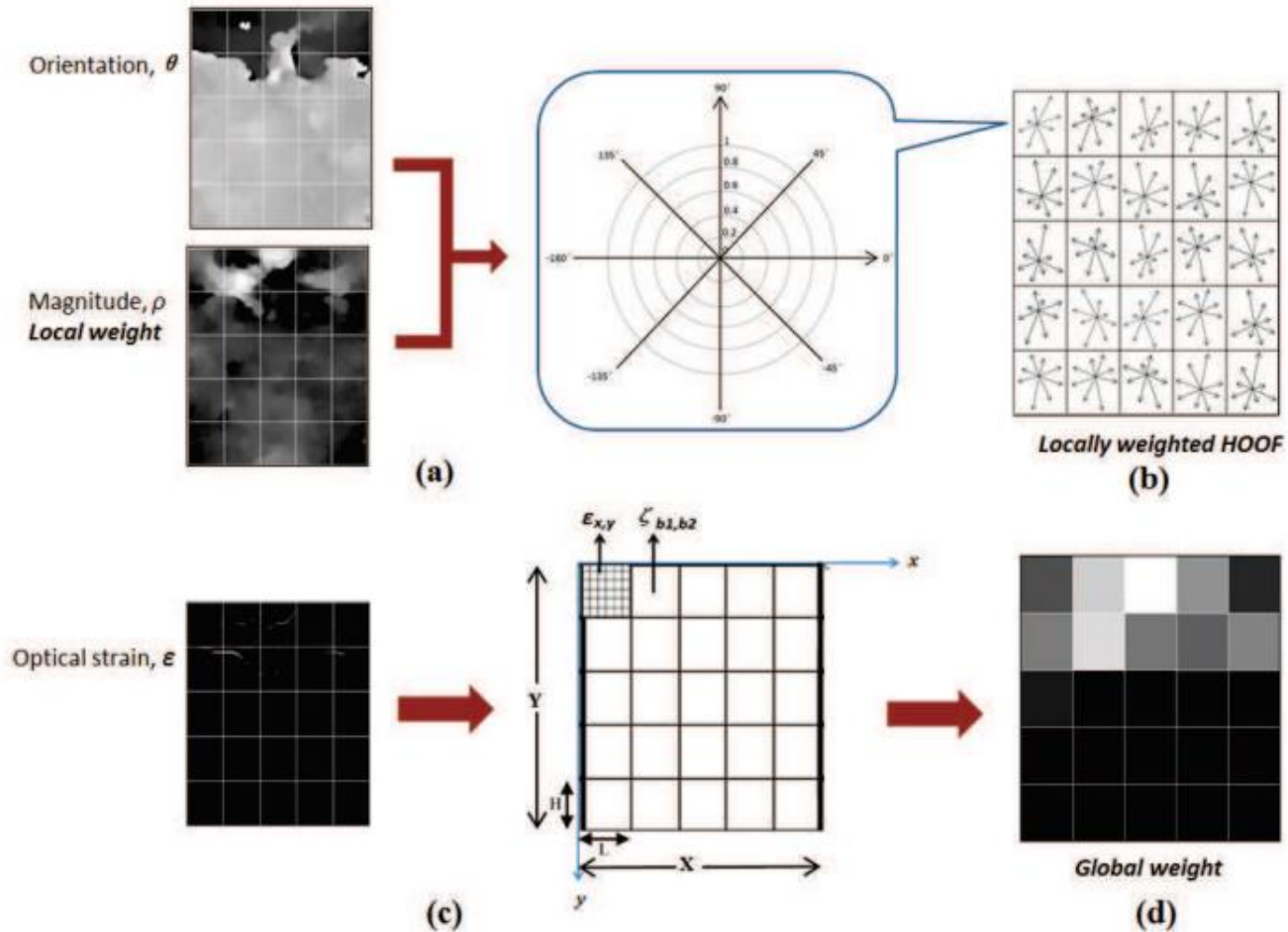
What is there at the apex?



- **Apex:** The frame where the AU reaches the peak or the point of highest intensity of facial motion.
- Optical Flow and Optical Strain shows significant magnitude at the apex.
- Datasets that do not provide the apices (SMICs) required spotting apex¹ in advance. CASME II apex can be directly used.

¹ Liong et al. (2015). **Automatic apex frame spotting in micro-expression database**. ACPR

Bi-Weighted Oriented Optical Flow (Bi-WOOF)



$$\zeta_{b_1, b_2} = \frac{1}{HL} \sum_{y=(b_2-1)H+1}^{b_2H} \sum_{x=(b_1-1)L+1}^{b_1L} \epsilon_{x,y},$$

Optical Flow & Optical Strain



(a) p



(b) q



(c) θ



(d) ρ



(e) ε

Optical Flow estimation

Horizontal and vertical flow $\vec{p} = [p = \frac{dx}{dt}, q = \frac{dy}{dt}]^T$

Magnitude & orientation (Euclidean \rightarrow Polar coordinates of the flow vector)

$$\rho_{x,y} = \sqrt{p_{x,y}^2 + q_{x,y}^2}$$

$$\theta_{x,y} = \tan^{-1} \frac{q_{x,y}}{p_{x,y}}$$

Optical Strain calculation

Approximating deformation intensity: Strain tensor

$$\begin{aligned} \varepsilon &= \frac{1}{2} [\nabla \mathbf{u} + (\nabla \mathbf{u})^T] \\ &= \begin{bmatrix} \varepsilon_{xx} = \frac{\partial u}{\partial x} & \varepsilon_{xy} = \frac{1}{2} \left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} \right) \\ \varepsilon_{yx} = \frac{1}{2} \left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y} \right) & \varepsilon_{yy} = \frac{\partial v}{\partial y} \end{bmatrix} \end{aligned}$$

$$|\varepsilon_{x,y}| = \sqrt{\varepsilon_{xx}^2 + \varepsilon_{yy}^2 + \varepsilon_{xy}^2 + \varepsilon_{yx}^2}$$

Experimental Results & Benchmarking



Methods		CASME II	SMIC-HS	SMIC-VIS	SMIC-NIR
Sequence-based	1 LBP-TOP [9, 14]	.39	.39	.39	.40
	2 OSF [24]	-	.45	-	-
	3 STM [51]	.33	.47	-	-
	4 OSW [25]	.38	.54	-	-
	5 LBP-SIP [21]	.40	.55	-	-
	6 MRW [26]	.43	.35	-	-
	7 STLBP-IP [22]	.57	.58	-	-
	8 OSF+OSW [52]	.29	.53	-	-
	9 FDM [30]	.30	.54	.60	.60
	10 Sparse Sampling [29]	.51	.60	-	-
	11 STCLQP [23]	.58	.64	-	-
	12 MDMO [28]	.44	-	-	-
	13 Bi-WOOF	.56	.53	.62	.57
Apex-based	14 LBP (random & onset)	.38	.40	.48	.51
	15 LBP (apex & onset)	.41	.45	.49	.54
	16 HOOF (random & onset)	.41	.40	.51	.50
	17 HOOF (apex & onset)	.43	.48	.49	.47
	18 Bi-WOOF (random & onset)	.50	.46	.56	.50
	19 Bi-WOOF (apex & onset)	.61	.62	.58	.58

(a) Baseline

	DIS	HAP	OTH	SUR	REP
DIS	.20	.11	.66	.02	.02
HAP	.09	.47	.25	0	.19
OTH	.21	.12	.58	.08	0
SUR	.12	.36	.20	.32	0
REP	.07	.33	.26	.04	.30

(b) Bi-WOOF (apex & onset)

	DIS	HAP	OTH	SUR	REP
DIS	.49	.07	.44	0	0
HAP	.03	.59	.28	.03	.06
OTH	.21	.09	.62	.01	.06
SUR	.04	.12	.08	.76	0
REP	.07	.19	.22	0	.52

CASME II		SMIC-HS	
Bin	F-measure	Accuracy	F-measure
1	.39	46.09	.46
2	.61	57.20	.50
3	.59	55.56	.49
4	.54	51.03	.58
5	.60	58.02	.53
6	.58	54.32	.54
7	.57	54.32	.50
8	.61	58.85	.62
9	.59	56.38	.49
10	.61	59.67	.59

Ablating the weights

(a) SMIC-HS

		Local		
Weights		None	Flow	Strain
Global	None	.44	.42	.43
	Flow	.51	.52	.50
	Strain	.54	.62	.59

(b) CASME II

		Local		
Weights		None	Flow	Strain
Global	None	.43	.52	.49
	Flow	.53	.58	.56
	Strain	.59	.61	.59

How do the Bi-WOOF weights affect the outcome of recognition?

Crucial for Strain information to weigh the contribution of blocks globally

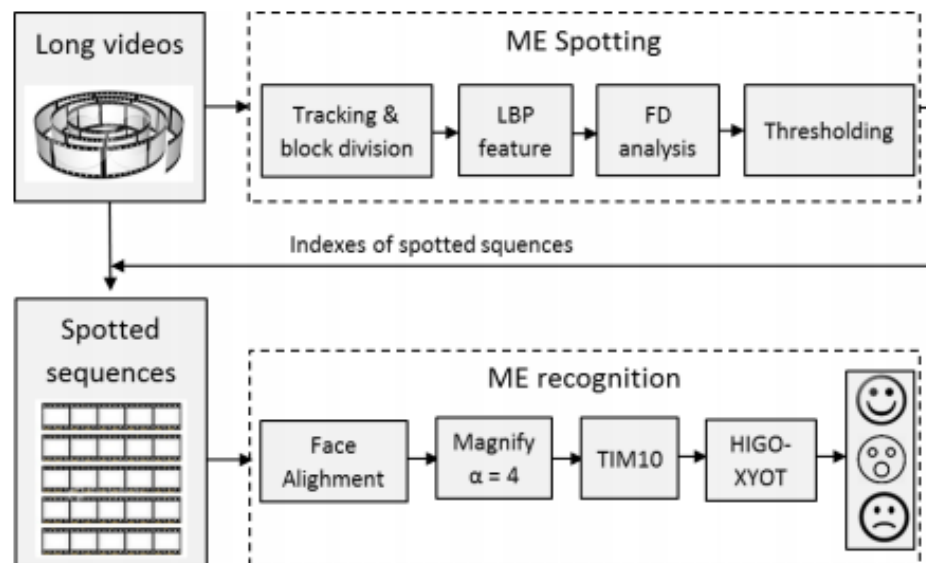
Locally, Flow magnitudes are good as weights to the Flow orientation

No weights, not good!

Computational cost savings of ~33 times

Towards Reading Hidden Emotions: A Comparative Study of Spontaneous ME Spotting and Recognition Methods

Li et al. (2017, T-AC)



Spotting TPR = 74.86%

"Spot-then-recognize" accuracy = 56.67%
using correctly spotted ME sequences

Overall system performance = $74.86 \times 56.67 = 42.42\%$

Towards Reading Hidden Emotions: A Comparative Study of Spontaneous ME Spotting and Recognition Methods

Li et al. (2017, T-AC)

	SMIC-HS	SMIC-VIS	SMIC-NIR	CASMEII
LBP	57.93%	70.42%	64.79%	55.87%
LBP+Mag	60.37%	78.87%	67.61%	60.73%
HOG	57.93%	71.83%	63.38%	57.49%
HOG+Mag	61.59%	77.46%	64.79%	63.97%
HIGO	65.24%	76.06%	59.15%	57.09%
HIGO+Mag	68.29%	81.69%	67.61%	67.21%
HIGO+Mag*	75.00%*	83.10%*	71.83%*	78.14%*
Li [18]	48.8%	52.1%	38.0%	N/A
Yan [20]	N/A	N/A	N/A	63.41%*
Wang [39]	71.34%*	N/A	N/A	65.45%*
Wang [57]	64.02%*	N/A	N/A	67.21%*
Wang [58]	N/A	N/A	N/A	62.3%
Liong [59]	53.56%	N/A	N/A	N/A
Liong [60]	50.00%	N/A	N/A	66.40%*

** results achieved using leave-one-sample-out cross validation.*

Best Recognition Accuracy (with hand-labelled ME sequences)

= **67.21%** (CASMEII) → **SOTA**

Towards Reading Hidden Emotions: A Comparative Study of Spontaneous ME Spotting and Recognition Methods

Li et al. (2017, T-AC)

Benchmarking via Human Test

- 15 subjects (avg. age 28.5 years, 10 male, 5 females)
- Definition of emotions explained, ME clips from SMIC-VIS were shown, subjects asked to select their answers after watching them
- Mean accuracy = **72.11%** (SMIC-VIS accuracy using proposed method = **81.69%**)

Insights:

- A very first attempt at a combined spotting and recognition pipeline
- Limitations: Problems in spotting (fixed spotting intervals, non-ME movements) hamper recognition capability

End of Part 4

Questions?