Tutorial

ACCV 2020

Recent Advances and Challenges in Facial Micro-Expression Analysis

ME Recognition

John See Multimedia University, Malaysia

Moi-Hoon Yap Manchester Metropolitan University, UK

Su-Jing Wang Chinese Academy of Sciences, China

Jingting Li Chinese Academy of Sciences, China

Sze-Teng Liong Feng Chia University, Taiwan

Outline

Recognition Pipeline

Pre-processing

- Data Magnification (spatial)
- Data Interpolation (temporal)

Feature Extraction / Learning

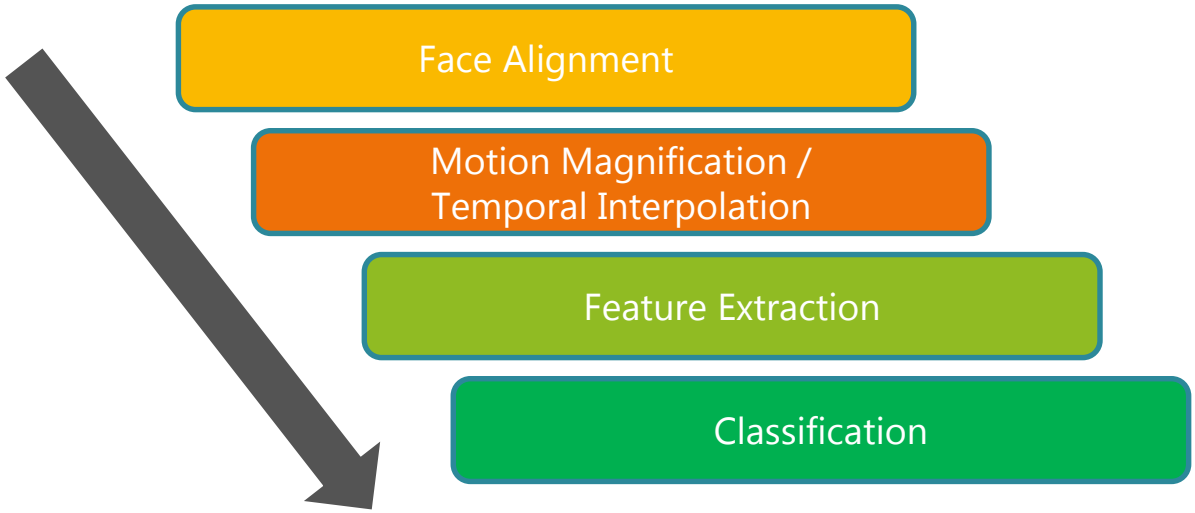
- LBP-based Methods
- Optical flow-based Methods
  - **Highlighted Work:** Less Is More: Micro-Expression Recognition from Video using Apex Frame
- Deep learning Methods … based on
  - Input
  - Depth
  - Data Domain

Spot-then-Recognize Approach

Spotting "in the wild"

ME Recognition Pipeline

Typically, a ME recognition process will follow these steps:

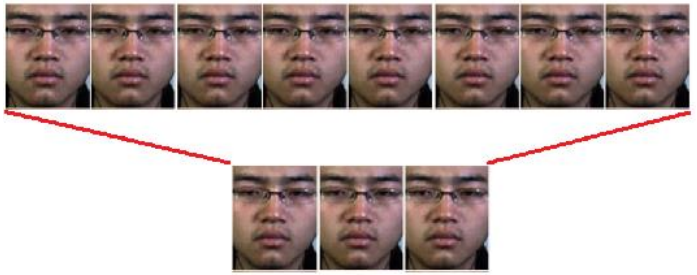


Pre-processing

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



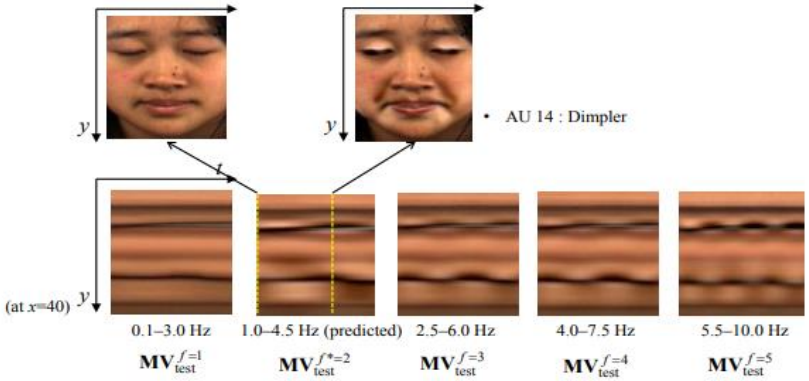
Motion Magnification

- "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/

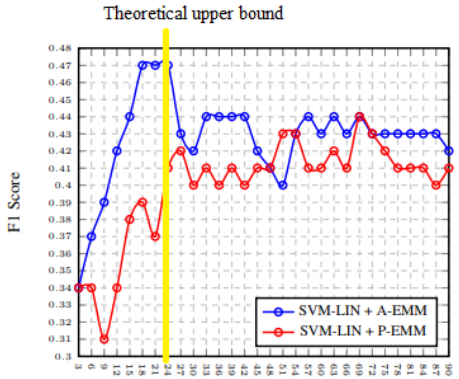
Motion Magnification in ME

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



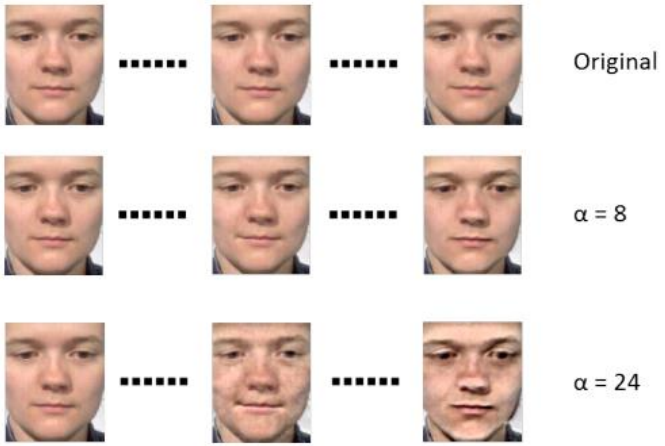
- **Le Ngo et al. (2016)** – Theoretical estimation of upper bounds of effective magnification factors
  - Empirical proof of Wu's proposed bounds w.r.t. spatial cut-off wavelength:

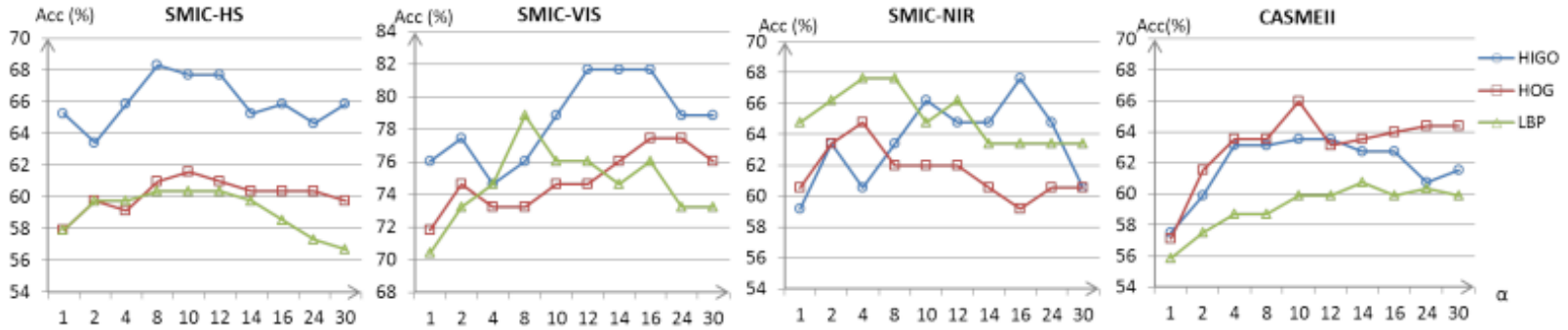
$$(1+\alpha_{A-EMM}) * \delta(t) < \frac{\lambda_c}{8}$$



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

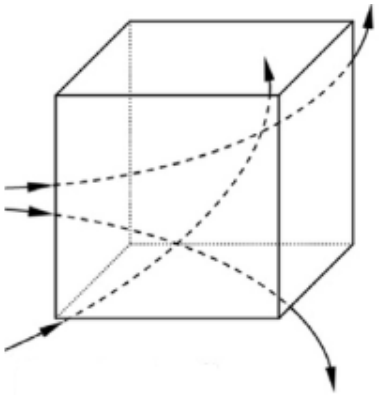




Motion Magnification in ME: "Local" vs. "Global"

- **Le Ngo et al. (2018)**
  - **"Local" Eulerian model** ➔ Modifying intensities of video frames based on frame information

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

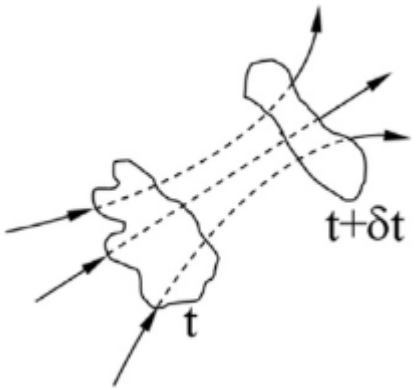


  - **"Global" Lagrangian model** ➔ Synthesizing magnified motion from statistical model of the whole video sequence

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

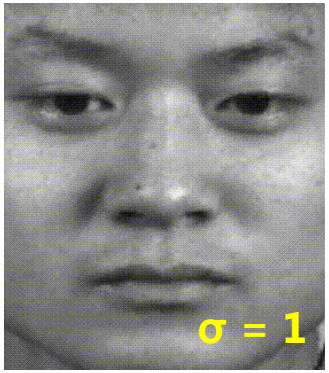
$$I_T = \text{warp}(I_R, \sigma(\mathbf{u}, \mathbf{v}))$$

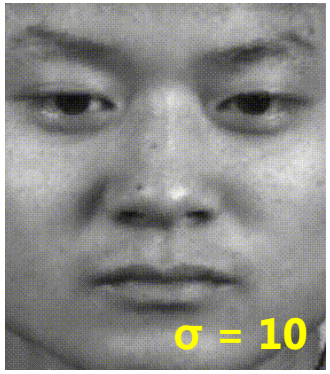
warp: a synthesis operation from I_R to I_T
I_T: magnified image
I_R: reference image
σ: magnification of motion fields

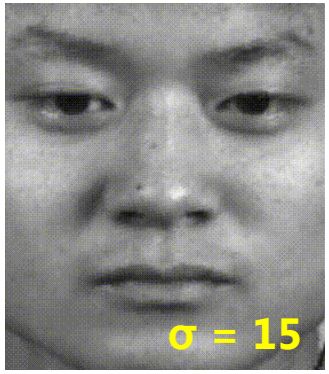


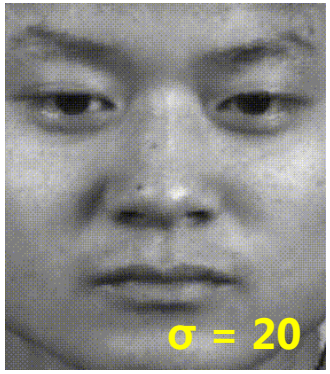
Visual Comparison

AEMM

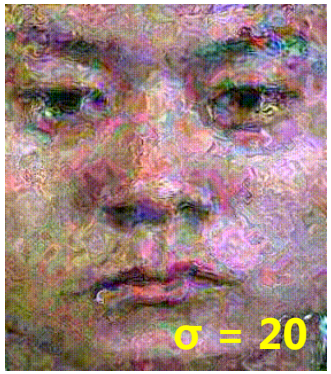




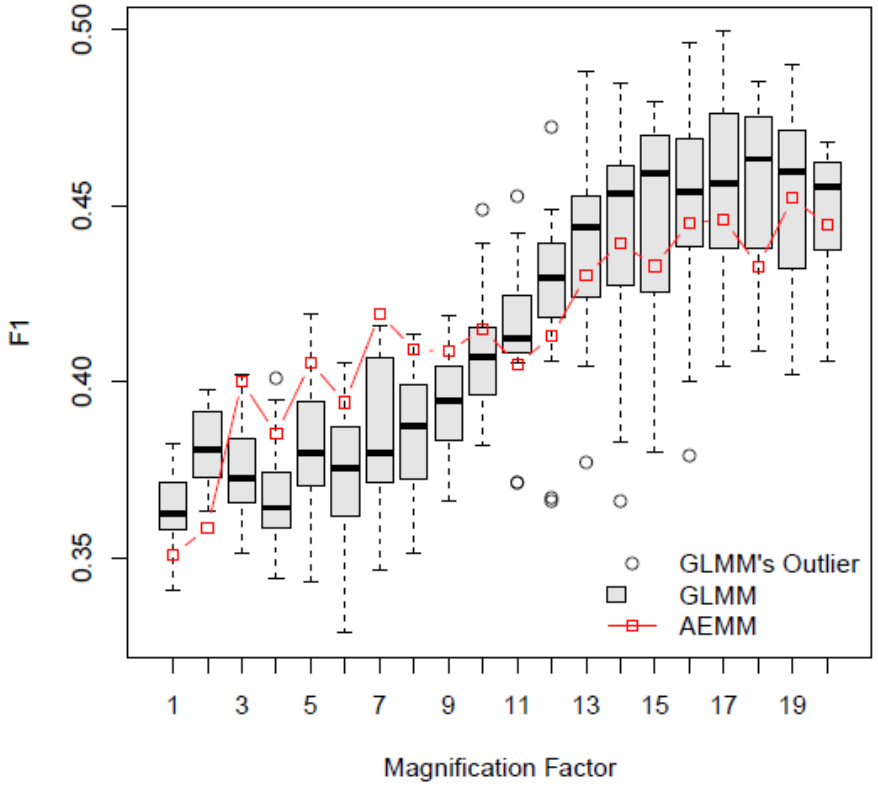




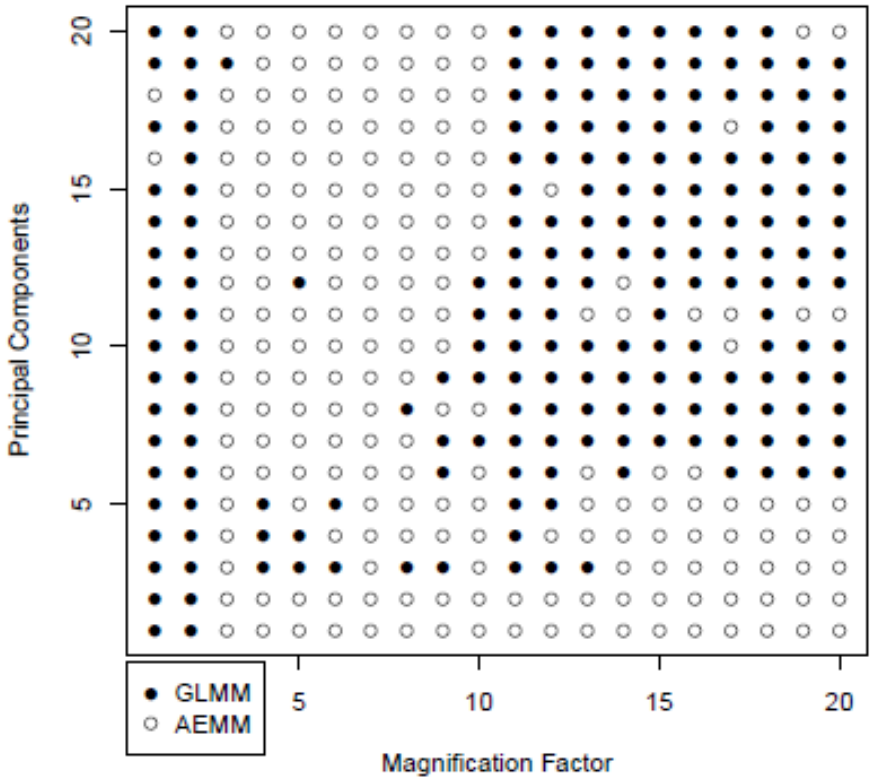
GLMM (k=9)



AEMM vs. GLMM: Results



Box Plot



Go-Chart

Reinforces the benefits of motion magnification towards ME recognition performance

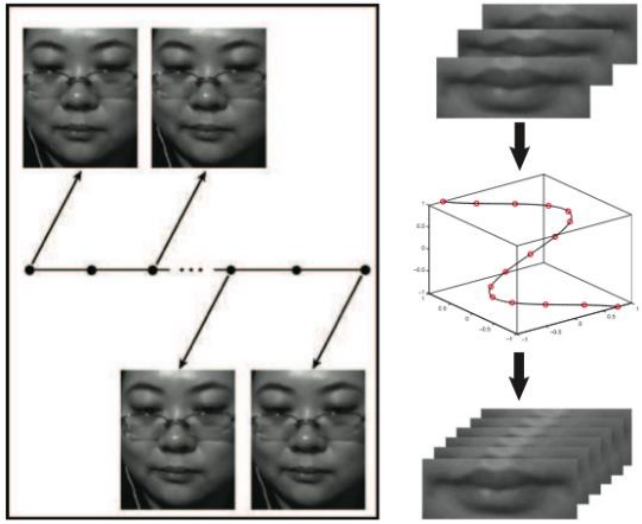
Offers GLMM as an alternative for amplifying subtle changes in MEs

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



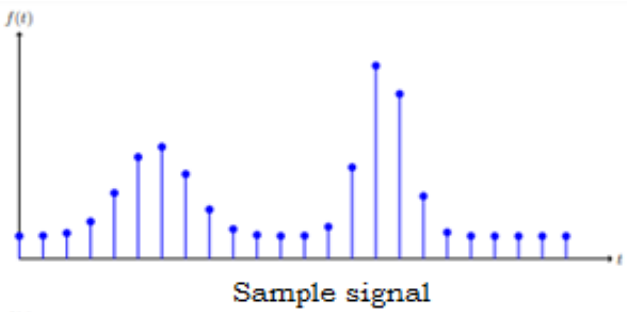
Dynamic selection: Reduce and compress

Interpolation/extrapolation is a "blanket" operation

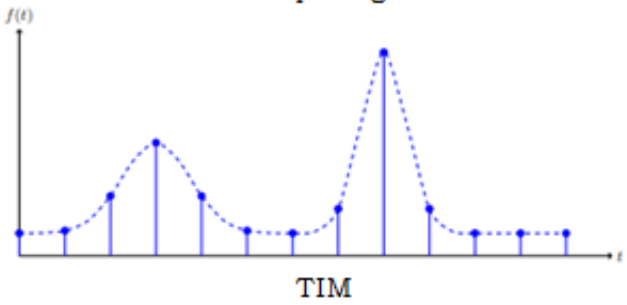
- Does not consider intrinsic dynamics in each video
- Selection based on # frames does not generalize well to MEs exhibited by different people and emotion types

Intuition: Reduce-and-compress

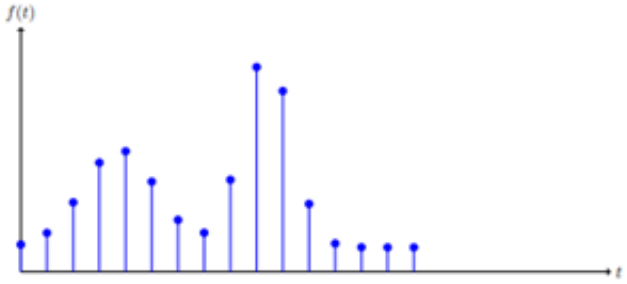
- In speech processing/lip reading, informative samples is more certain after trimming, TIM is acceptable ➔ Interpolation can be done on the originally assumed manifold
- **What we want**: Find informative information based on sparse constraints, and make a reduced size selection (subset selection <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



Sample signal



TIM

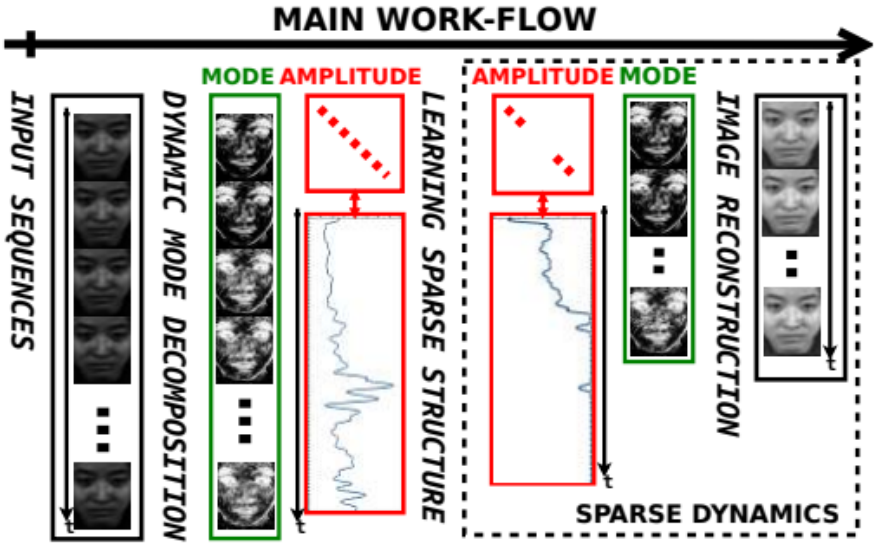


What we want

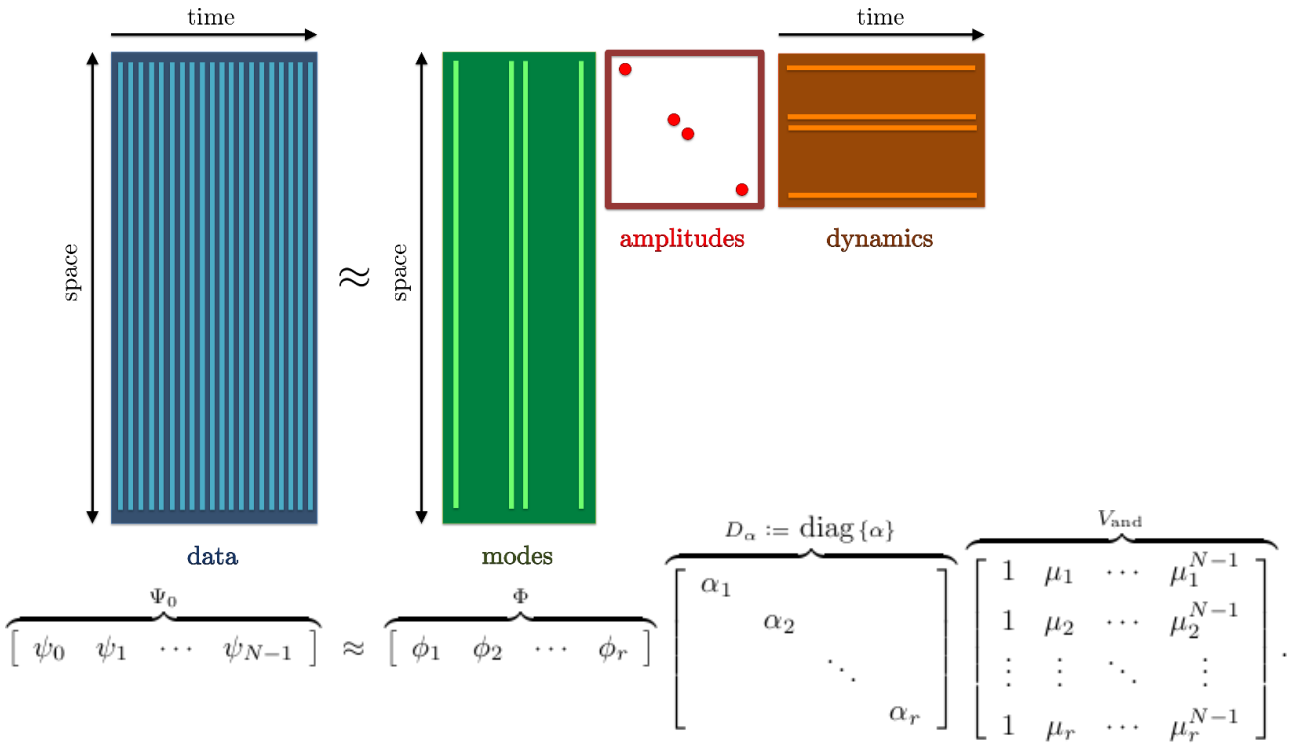
Sparsity-Promoting Dynamic Mode Decomposition (DMDSP)

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

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

$$\boldsymbol{\alpha}_{\boldsymbol{dmdsp}} = \begin{bmatrix} \boldsymbol{I} & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{P} & \boldsymbol{E} \\ \boldsymbol{E}^{\boldsymbol{T}} & \boldsymbol{0} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{q} \\ \boldsymbol{0} \end{bmatrix}$$

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

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 S
RA: Random Sampling w.r.t. % length
BL: Baseline (no changes to original sequence)

	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

	CASME II				SMIC			
	ACC	F1	R	P	ACC	F1	R	P
Sparse Sampling	**.49**	**.51**	**.47**	**.55**	**.58**	**.60**	**.60**	**.60**
Huang et at. [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!

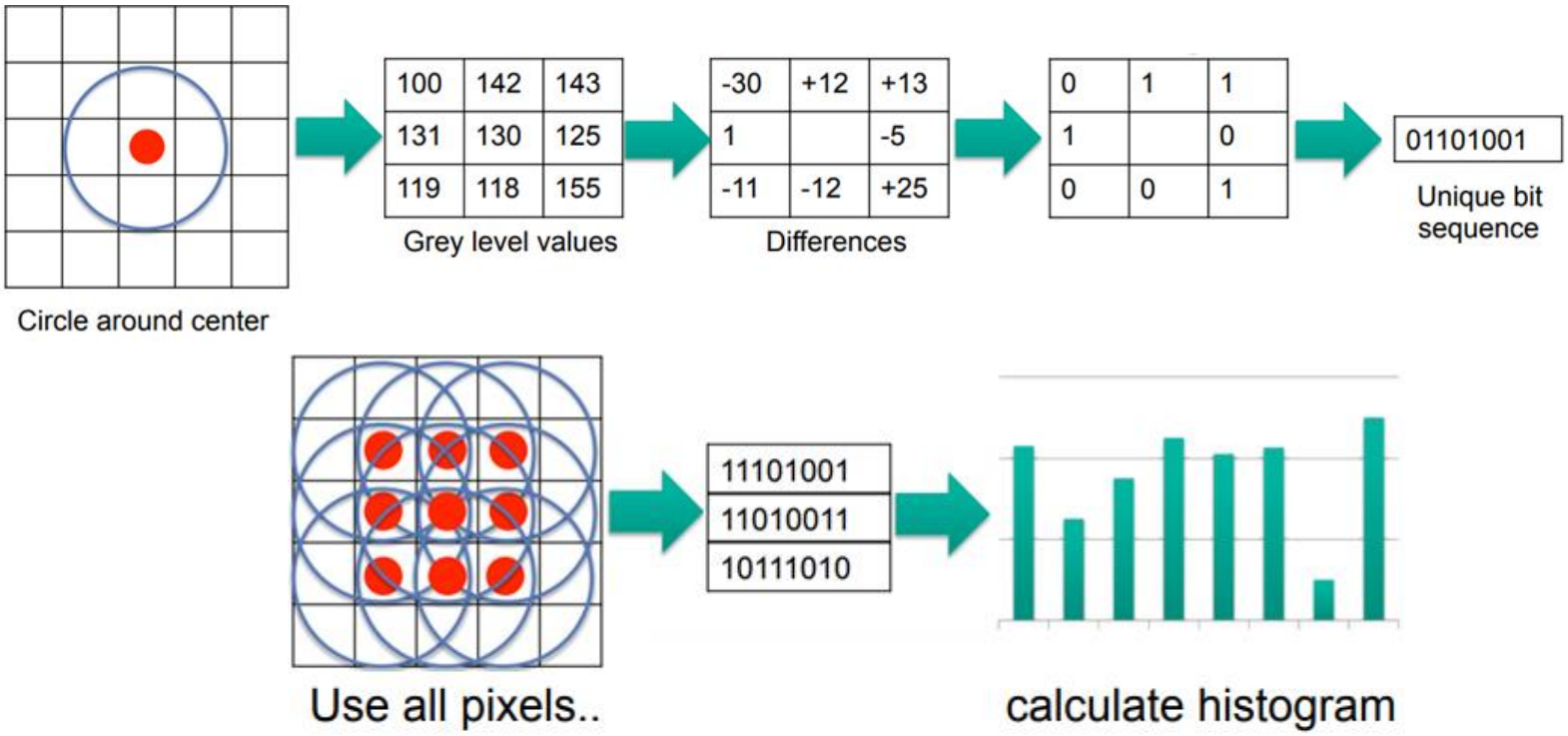
SOTA when published. Now no longer best ☺

Feature Extraction Techniques

I. LBP-TOP, LBP-based methods (texture)
II. Optical Flow-based methods (motion)
III. Other descriptor based methods
IV. Deep learning methods

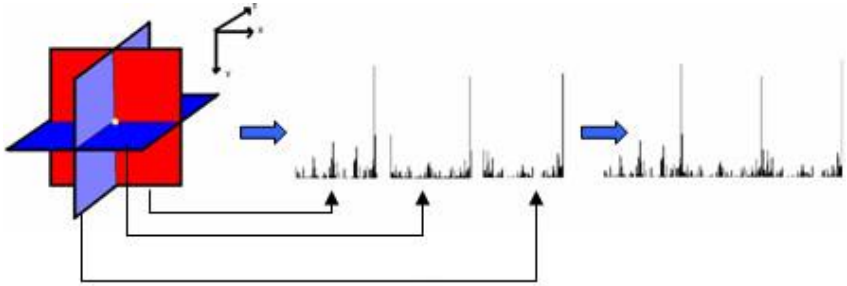
(I) Local Binary Pattern (LBP)

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

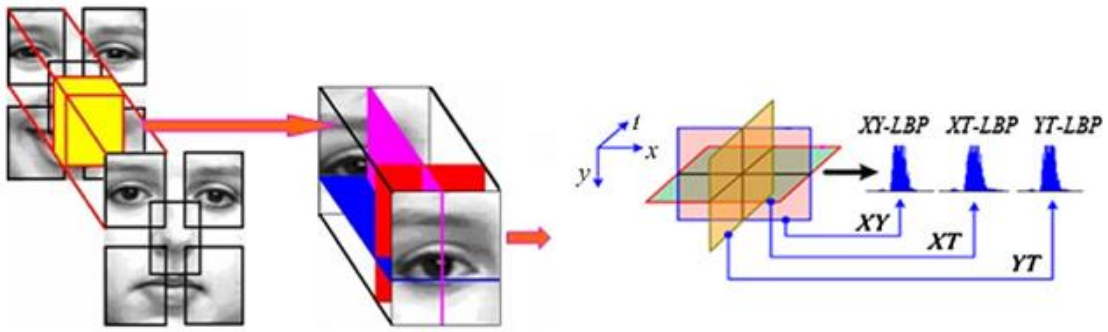


Local Binary Pattern (LBP) on Three Orthogonal Planes (TOP)

- 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

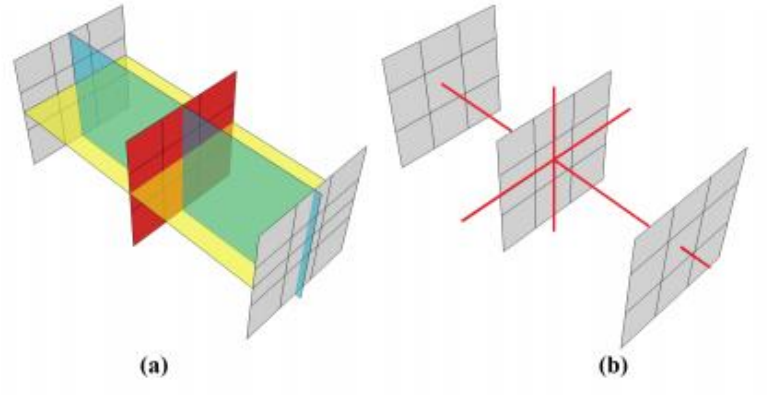


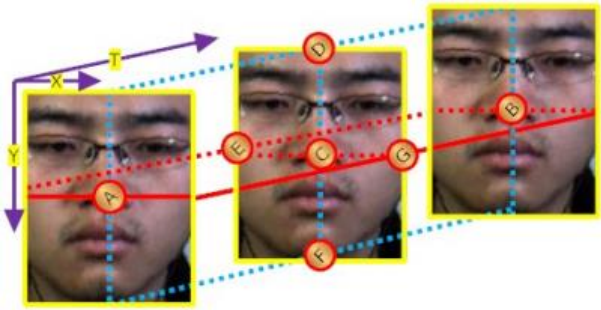
- Block-based LBP-TOP
  - Divide into blocks, each block extracts LBP-TOP histograms, concatenate again



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
$XY \cap XT \cap YT = \{A, B, D, E, F, G\}$

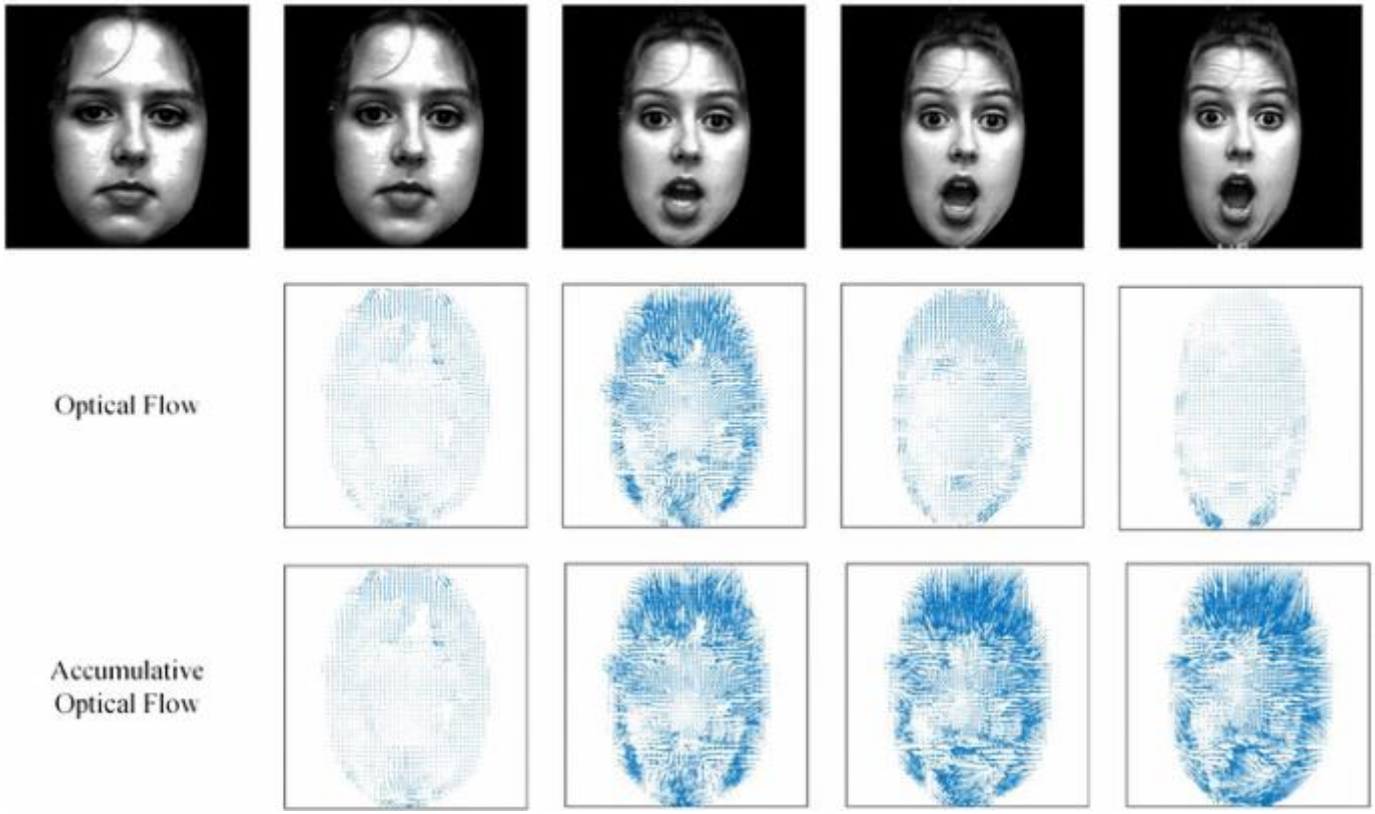
	CASMEII		SMIC	
	LBP-TOP (%)	LBP-SIP (%)	LBP-TOP (%)	LBP-SIP (%)
Linear	62.75	63.56	60.98	64.02
RBF	65.99	66.40	60.98	62.80

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 Randon 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

(II) Optical Flow

- **Optical Flow**: An estimation of the apparent motion of pixel intensities (or edges, surfaces, objects) over time in a video



Zhao, J., Mao, X., & Zhang, J. (2018). Learning deep facial expression features from image and optical flow sequences using 3D CNN. *The Visual Computer*, *34*(10), 1461-1475.

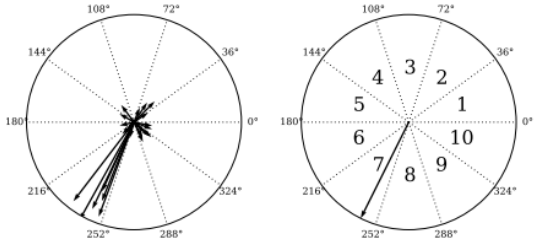
Optical Flow

- Among the Optical Flow **flavours** commonly used in ME motion representation:
  - Lucas-Kanade
  - Horn-Schunck
  - Black-Anandan
  - Dense Optical Flow (Farneback)
  - TV-L1

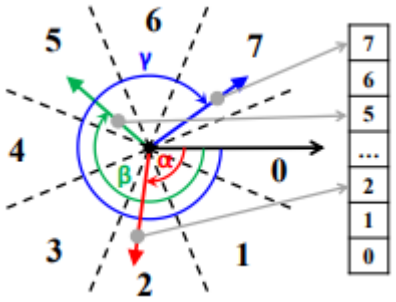
Zhao, Y., & Xu, J. (2019). A Convolutional Neural Network for Compound Micro-Expression Recognition. *Sensors*, *19*(24), 5553.

Selective towards principal directions of flow

- 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

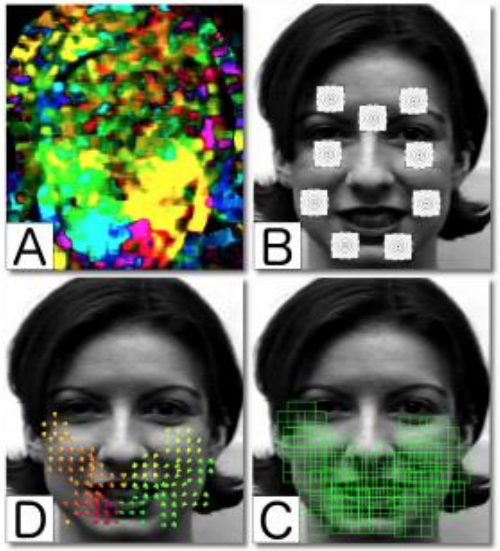


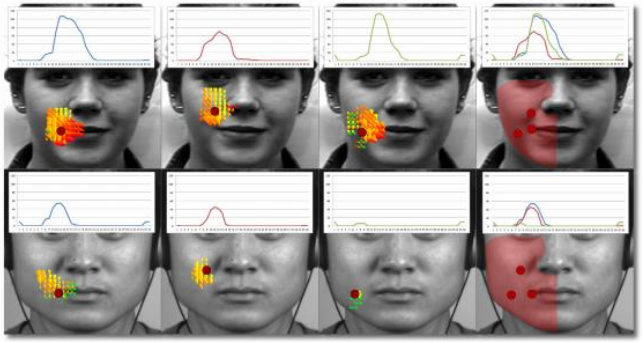
  - **Main Directional Mean Optical-flow (MDMO) (Liu et al, T-AC 2017)**
    - ROI-based normalized statistical feature based on the main direction of the optical flow in polar coordinates
    - 36 ROIs ➔ slim feature dimension of only 36 x 2 = 72

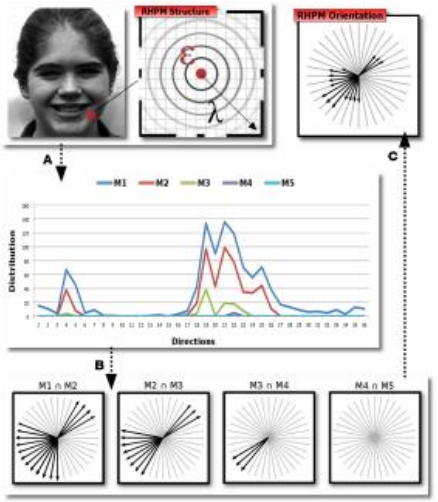


Selective towards regions of consistent flow

- **Allaert et al. (2017)**
  - 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

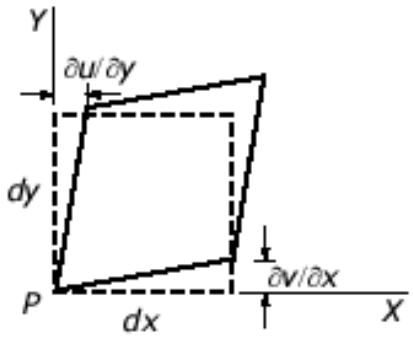






Optical Strain

- Assuming motion is sufficiently small, its corresponding finite strain tensor is defined as



$$\varepsilon = \frac{1}{2}[\nabla \mathbf{u} + (\nabla \mathbf{u})^T]$$

$$\varepsilon = \begin{bmatrix} \varepsilon_{xx} = \frac{\partial u}{\partial x} & \varepsilon_{xy} = \frac{1}{2}(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}) \\ \varepsilon_{yx} = \frac{1}{2}(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}) & \varepsilon_{yy} = \frac{\partial v}{\partial y} \end{bmatrix}$$

Normal strain components

Shear strain components

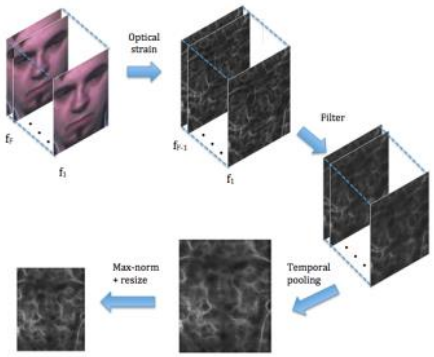
- Optical strain magnitudes for each pixel can be computed by sum of squares of all components:

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

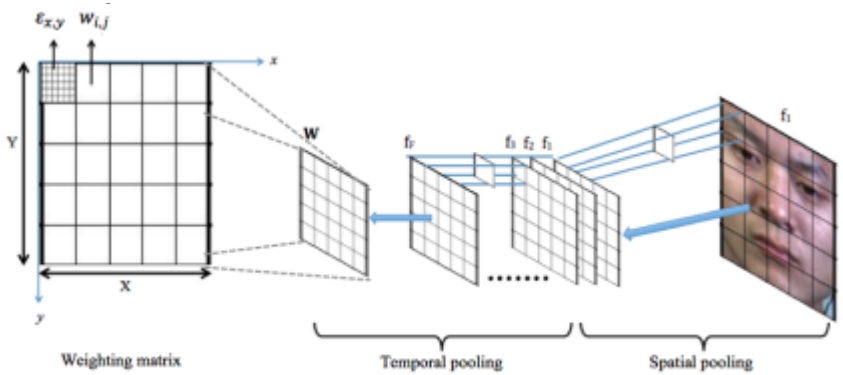
$$= \sqrt{{\frac{\partial u}{\partial x}}^2 + {\frac{\partial v}{\partial y}}^2 + \frac{1}{2}{(\frac{\partial u}{\partial x} + \frac{\partial u}{\partial x})}^2}.$$

Optical Strain

- Original idea by Shreve et al. for identifying macro and micro-expressions
- Optical Strain (OS) was fully modelled by Liong for use in 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



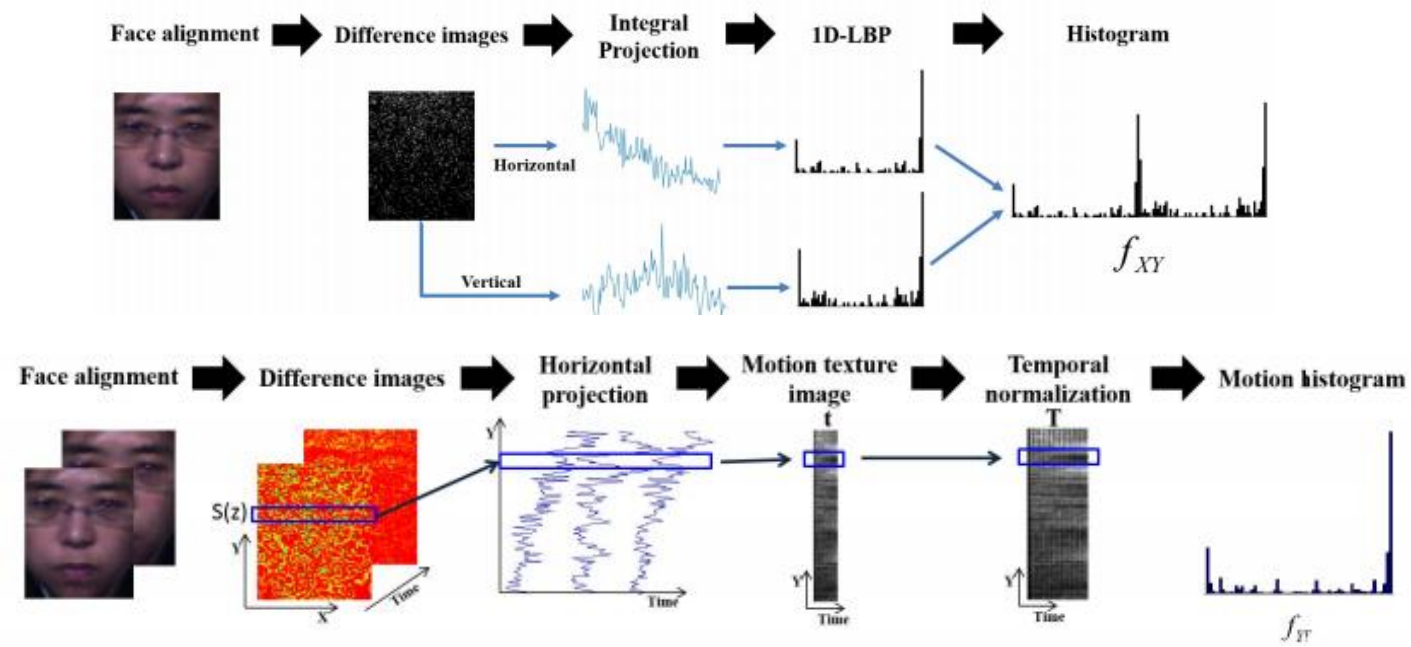
  - OS-weighted LBP-TOP features (Liong et al., ACCV 2014) ➔ allows regions that exhibit active ME motions to be given more significance, increasing discrimin emotion types



Constructing histograms from flow

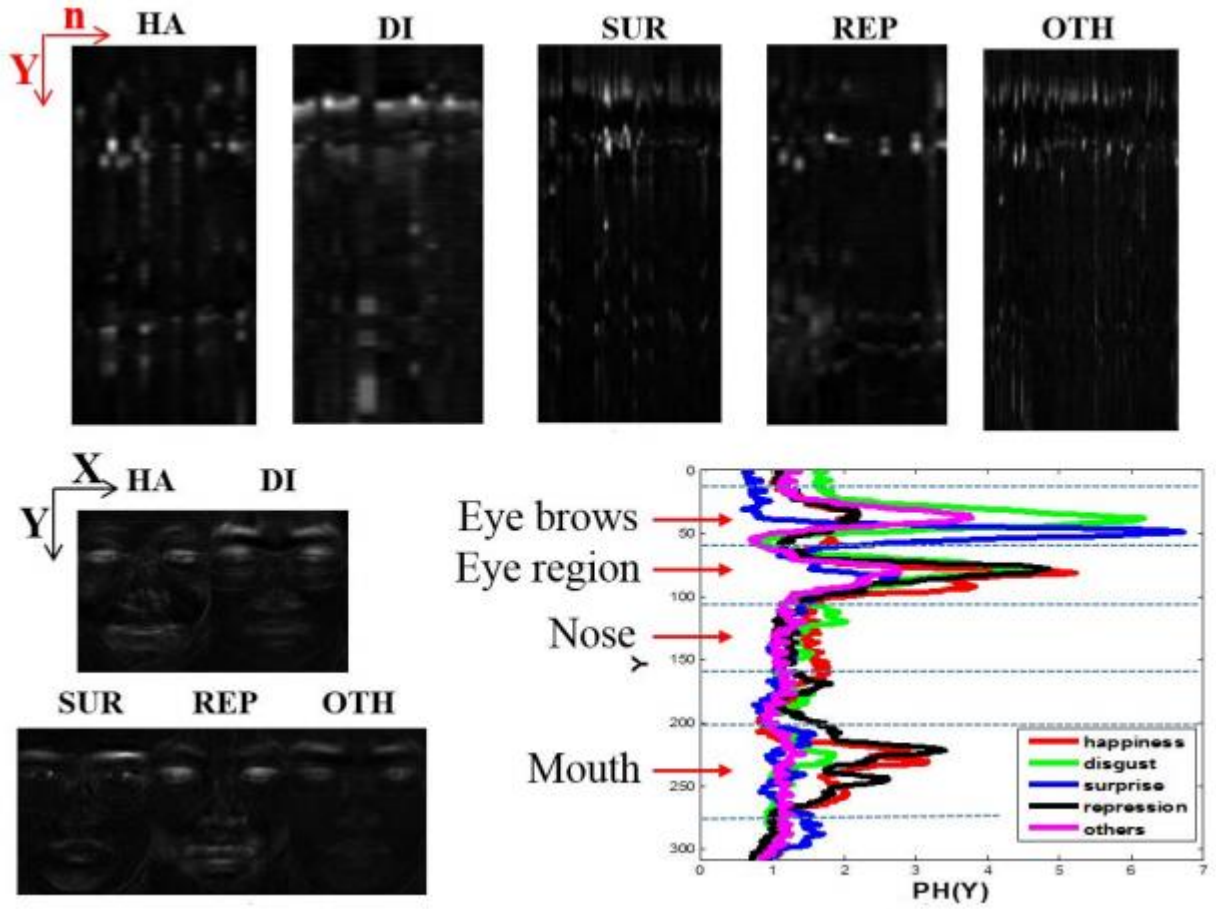
- **Zhang et al., 2017**: Region-by-region Aggregation of Histogram of Oriented Optical Flow (HOOF) and LBP-TOP to construct rich local statistical features
  - Doing it with ROIs yield even better results than globally done
- **Happy & Routray, 2017**: Fuzzy histogram of optical flow orientations (FHOFO)
  - 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



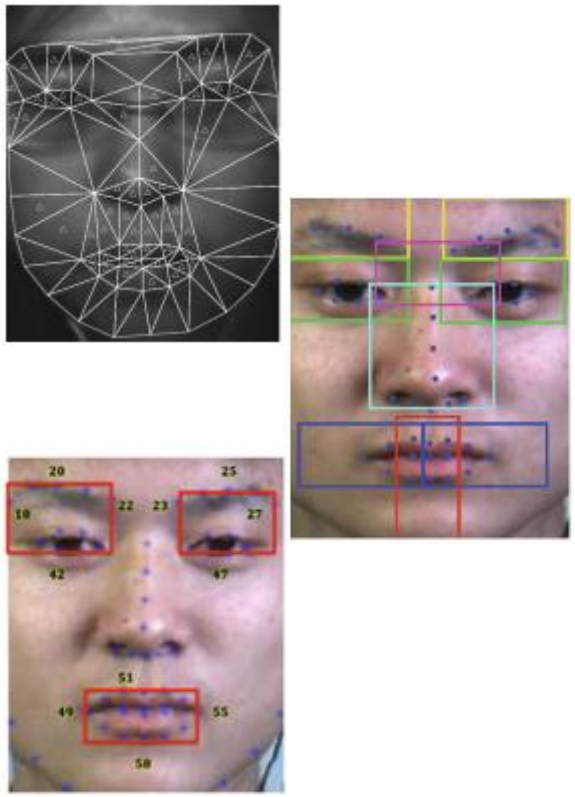
- **Huang et al., ICCV Workshops 2015**
  - Integral projection based on difference images is used to obtain horizontal and vertical projections
  - Apply 1DLBP operators on both projections to obtain features

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



(III) 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

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

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

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

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

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


Do we really need so much information?
???

Prima facie

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

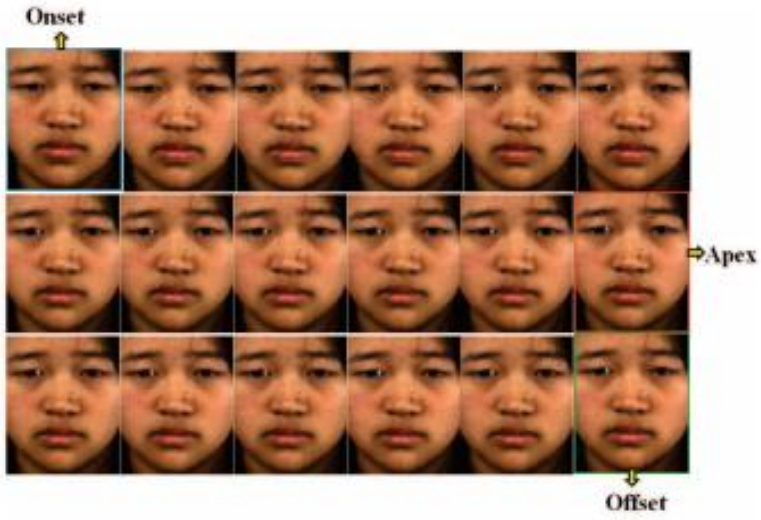
- ✓Ekman: Emotions are characterised by the change in facial contraction.
- ✓Exposito: Visual information (video) conveys poor emotional information, due to cognitive overload.

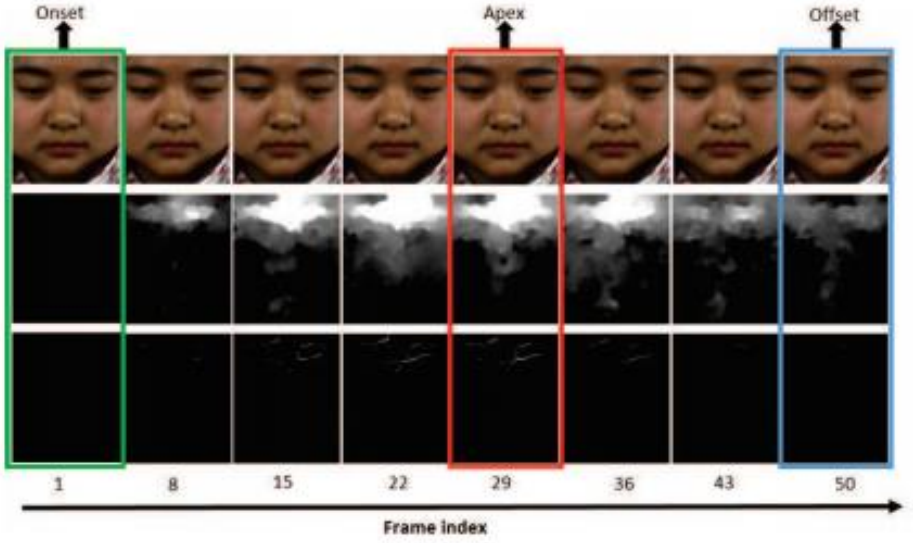
II. The apex frame is sufficient for micro-expression recognition

- ✓"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?

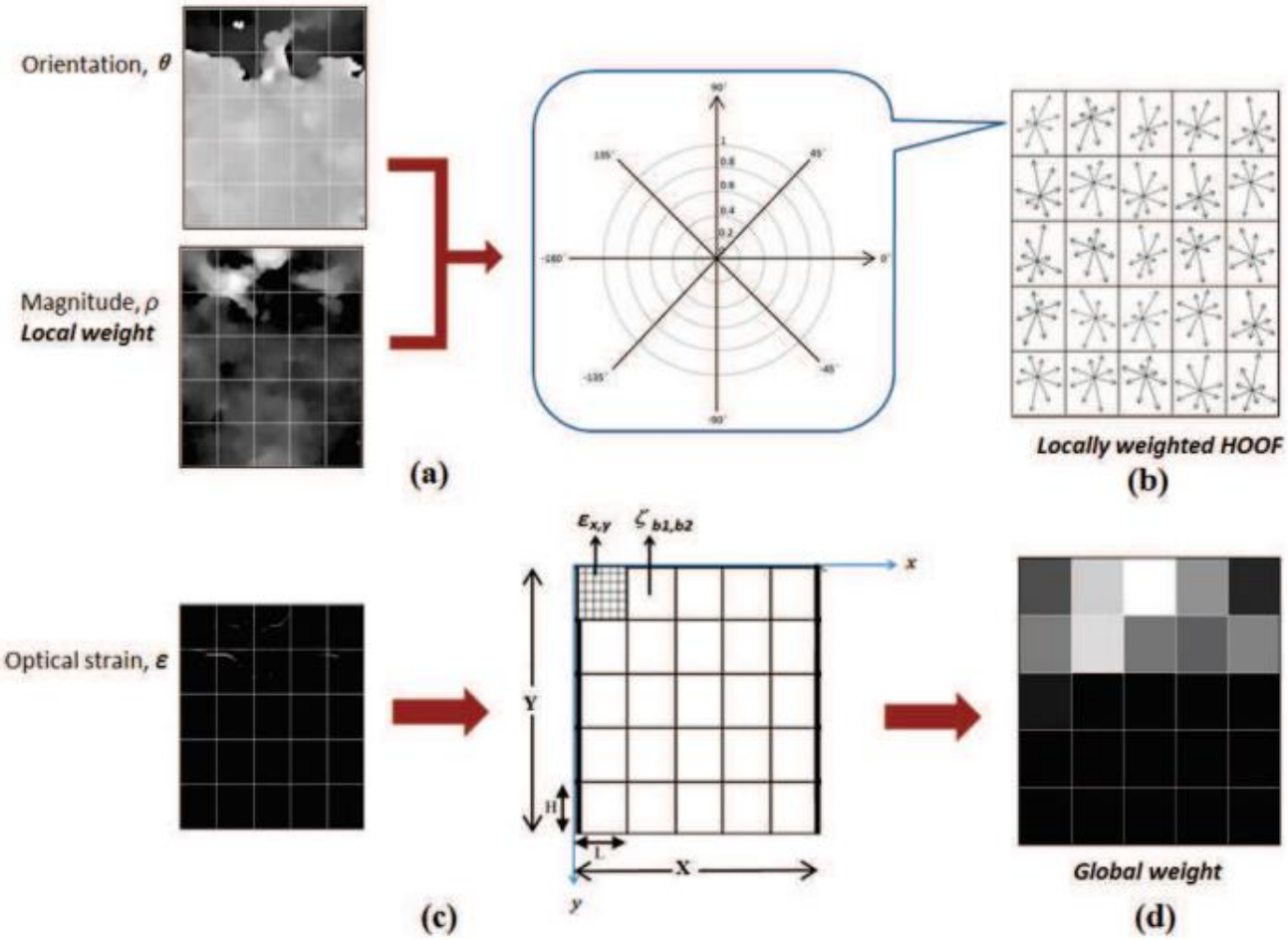




- **Apex:** The frame where the AU reaches the peak or the point of highest intensity of facial motion.
- Optical Flow and Optical Strain shows significant magnitude at the apex.
- Datasets that do not provide the apices (SMICs) required spotting apex[1] in advance. CASME II apex can be directly used.

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

Bi-Weighted Oriented Optical Flow (Bi-WOOF)



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

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 → 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

$$\varepsilon = \frac{1}{2}[\nabla \mathbf{u} + (\nabla \mathbf{u})^T]$$

$$= \begin{bmatrix} \varepsilon_{xx} = \frac{\partial u}{\partial x} & \varepsilon_{xy} = \frac{1}{2}(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}) \\ \varepsilon_{yx} = \frac{1}{2}(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}) & \varepsilon_{yy} = \frac{\partial v}{\partial y} \end{bmatrix}$$

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

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	Accuracy
1	.39	46.09	.46	45.12
2	.61	57.20	.50	50.00
3	.59	55.56	.49	48.78
4	.54	51.03	.58	58.54
5	.60	58.02	.53	54.27
6	.58	54.32	.54	54.27
7	.57	54.32	.50	50.00
8	**.61**	**58.85**	**.62**	**62.20**
9	.59	56.38	.49	49.39
10	.61	59.67	.59	58.54

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

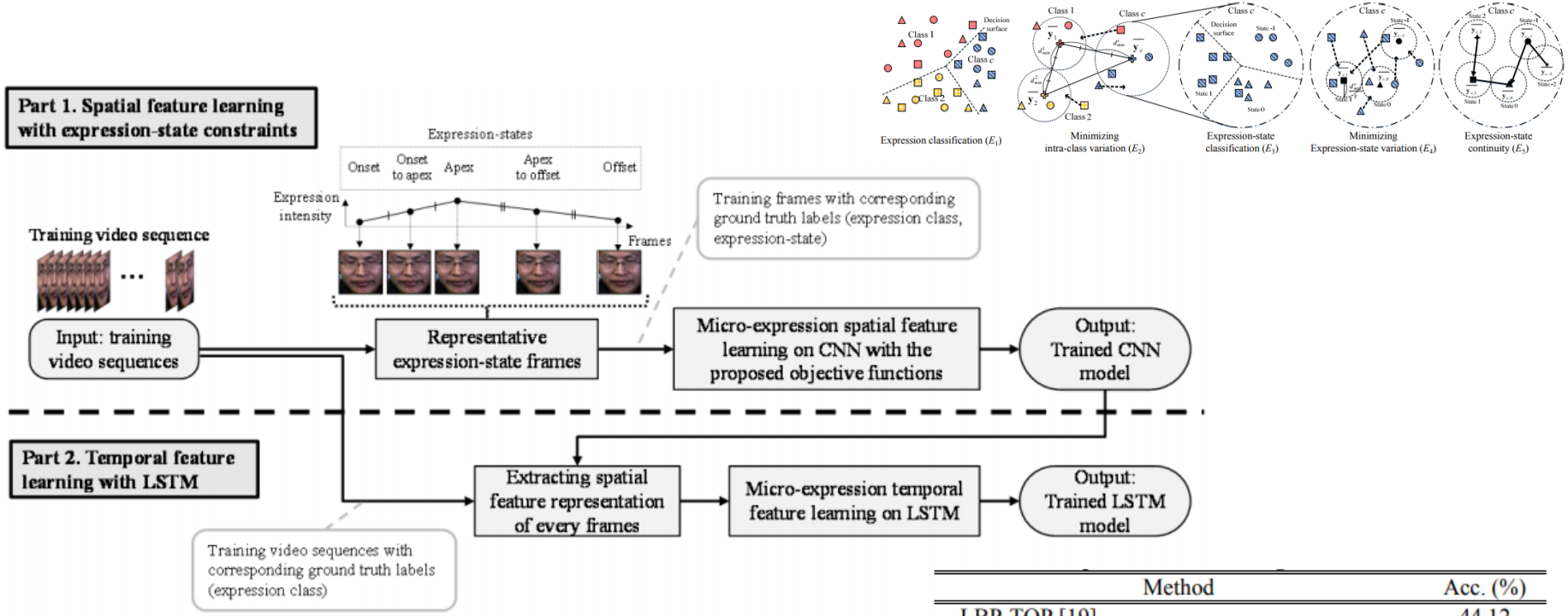
END OF SECTION

(IV) Deep Learning methods

- Deep Learning methods (needs no introduction here!) have been slow in adoption for ME recognition but has gained some momentum in recent years.
- **Key problems**:
  - Low number of samples (CASME II: 247, SAMM: 159) ➔ Very low in DL standards!
  - Databases have different number of classes (CASME II: 5, SMIC: 3, SAMM: 5, 6 or 7) ➔ Inconsistent benchmarking
  - Existing architectures were built with large-scale natural "in-the-wild" images in mind (ImageNet, Places365, LFW) ➔ Limited suitability even for transfer learning
- **Some possible remedies**:
  - The closest models that we could find are those trained for face recognition and facial expression recognition.
  - Merging of datasets
  - Data augmentation

Deep Learning methods: Early attempts

- **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)

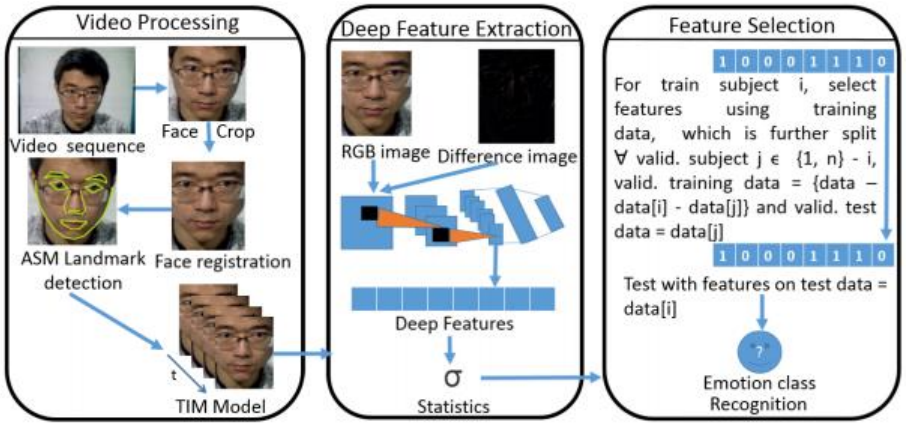


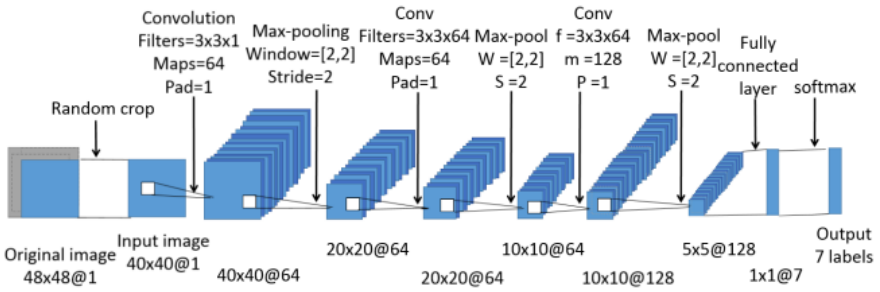
Method	Acc. (%)
LBP-TOP [19]	44.12
LBP-TOP + adaptive motion magnification [13]	51.91
LBP-MOP [16]	45.75
Riesz wavelet [12]	46.15
Proposed method	**60.98**

Kim, D. H., Baddar, W. J., & Ro, Y. M. (2016). Micro-expression recognition with expression-state constrained spatio-temporal feature representations. In Proceedings of the 24th ACM international conference on Multimedia (pp. 382-386).

Deep Learning methods: Early attempts

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





Approach	Database		
	VIS	HS	Casme II
Our approach	56.3	53.6	47.3
LBP-TOP (baseline) [14]	52	48	38
STCLQ [8]	-	64	59

Database	CNN	
	Object	FE
VIS	29.5	56.3
HS	-	53.6

- 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

Patel, D., Hong, X., & Zhao, G. (2016). Selective deep features for micro-expression recognition. In 2016 23rd international conference on pattern recognition (ICPR) (pp. 2258-2263)

(IV) Deep Learning methods

- Multi-Taxonomy of Deep Learning Approaches for ME recognition
- **Input**:
  - Sequence of video frames
  - Single representative frame (e.g. apex frame)
  - + Enriched input

- **Depth**:
  - Deep networks (transfer learning from spatial, sequential)
  - Shallow networks

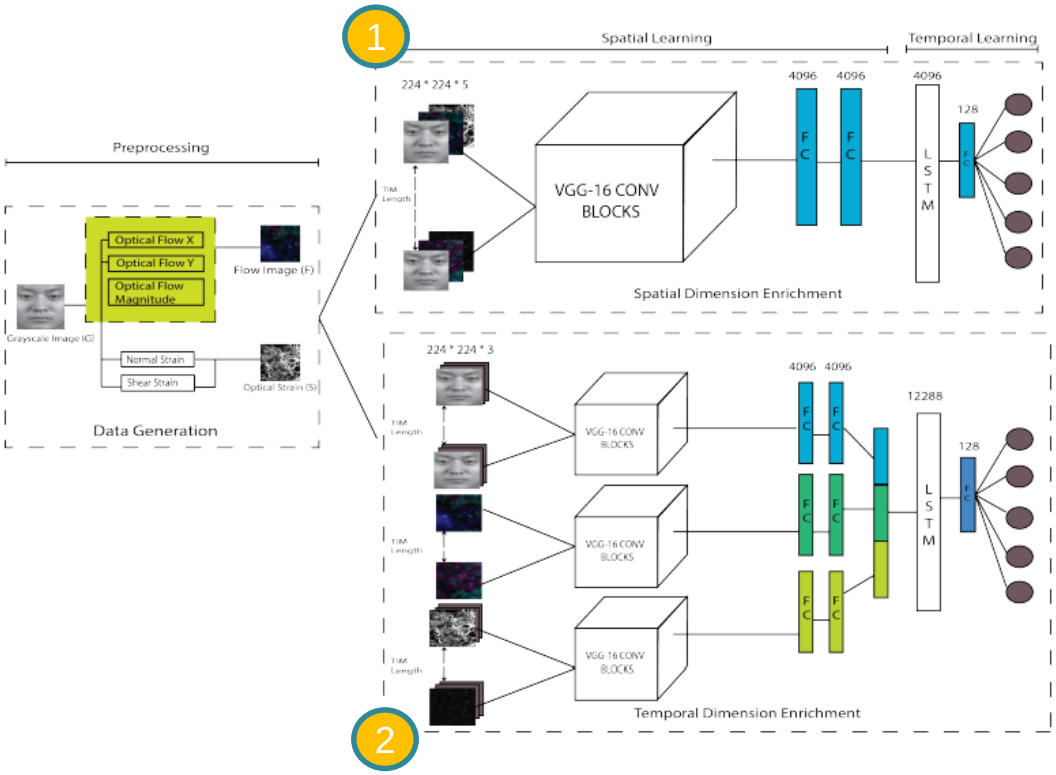
- **Data domain**:
  - Intra-domain learning
  - Inter-domain learning,
    i.e. domain adaptation / regeneration

Factor (I): Input

- **Q**: What kinds of input(s) would be suitable for deep learning in ME?
- Biggest dilemma: ME sequence vs. apex frame
  - ME **Sequence**: Makes sense (!) to capture transitive differences between frames
  - ME **Apex Frame**: *"less is more"* (Liong et al. 2016). Due to the minute differences between frames, using all frames in sequence is redundant and is not necessarily better
- ME images are simply just RGB/gray values, why not **enrich** data by
  - Computing derived signals → optical flow, optical strain, etc.
  - Use different sampling rates
  - Stacking channels, multiple streams/branches

Deep Learning methods: Enriched LRCN

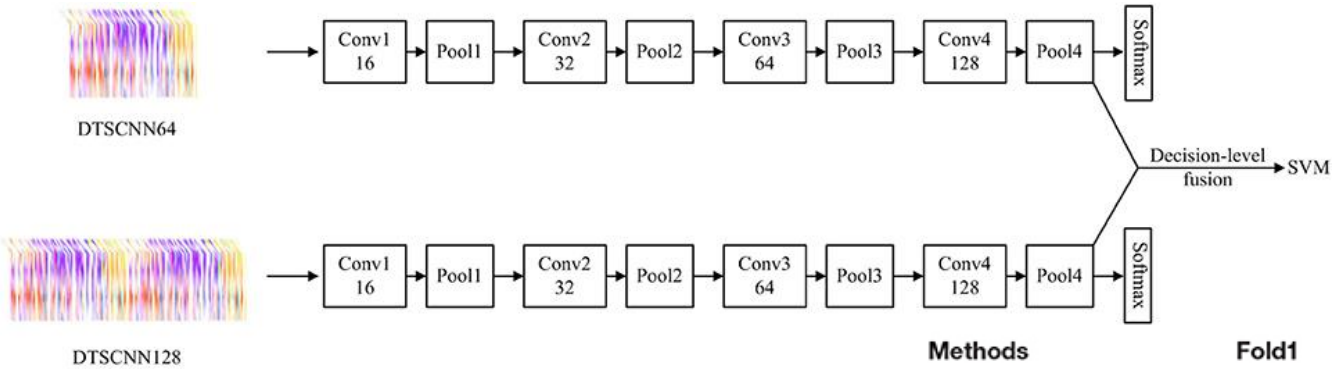
- 2 ways of enriching a CNN-LSTM pairing (also known as LRCN)
  - **Spatially**: Gray, OF, OS images stacked along CNN channel, pass features to LSTM
  - **Temporally**: Separate CNN streams for Gray, OF and OS images , late fusion after FC, pass features to LSTM



Khor, H. Q., See, J., Phan, R. C. W., & Lin, W. (2018). Enriched long-term recurrent convolutional network for facial micro-expression recognition. In 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018) (pp. 667-674)

Deep Learning methods: Dual Temporal Scale CNN

- **Dual Temporal Scale CNN – Peng et al.**
  - 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)

3-fold CV!

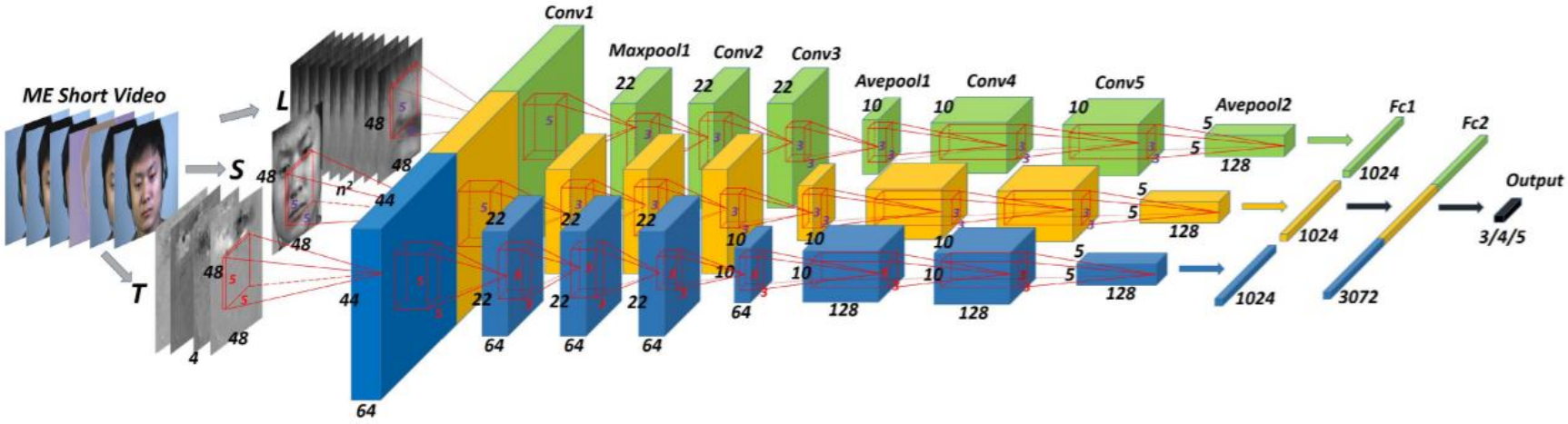
Methods	Fold1	Fold2	Fold3	Average
DTSCNN64 TIM64	65.45	65.45	65.45	65.45
DTSCNN128 TIM128	65.45	66.36	65.45	65.75
DTSCNN (fusion)	67.27	67.27	65.45	**66.67**

  - Why “dual temporal scale”? CASME I is 60fps, 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)

Peng, M., Wang, C., Chen, T., Liu, G., & Fu, X. (2017). Dual temporal scale convolutional neural network for micro-expression recognition. Frontiers in psychology, 8, 1745.

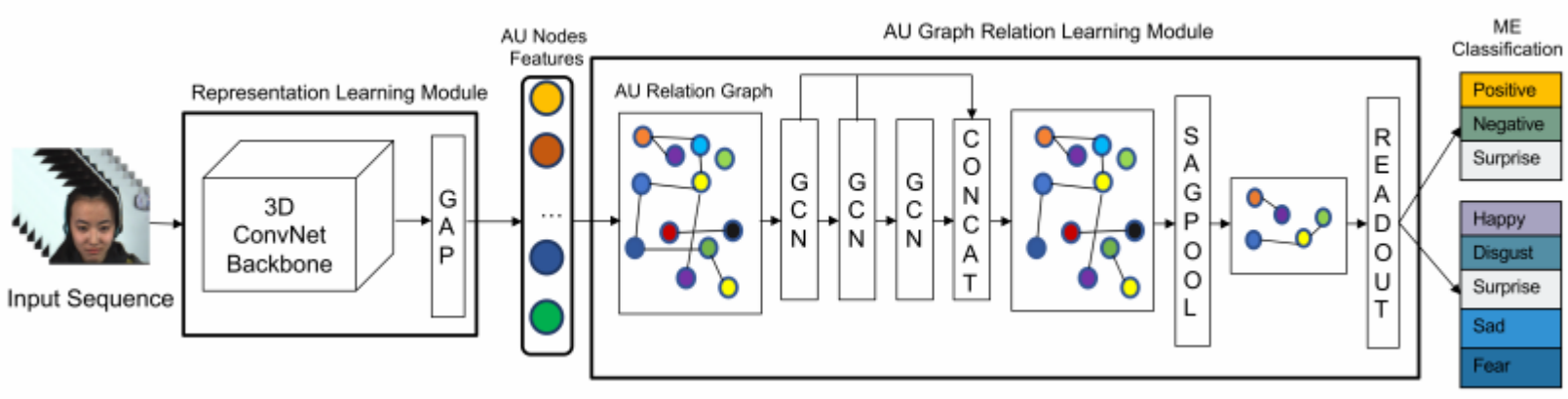
Deep Learning methods: Three Stream CNN

- 3 streams with different information:
  - **Facial Local Feature**, facial region is divided equally into N-blocks.
  - **Static Spatial Feature**, one single face image.
  - **Temporal Feature**, optical flow over T-frames.



Song, B., Li, K., Zong, Y., Zhu, J., Zheng, W., Shi, J., & Zhao, L. (2019). Recognizing Spontaneous Micro-Expression Using a Three-Stream Convolutional Neural Network. IEEE Access, 7, 184537-184551.

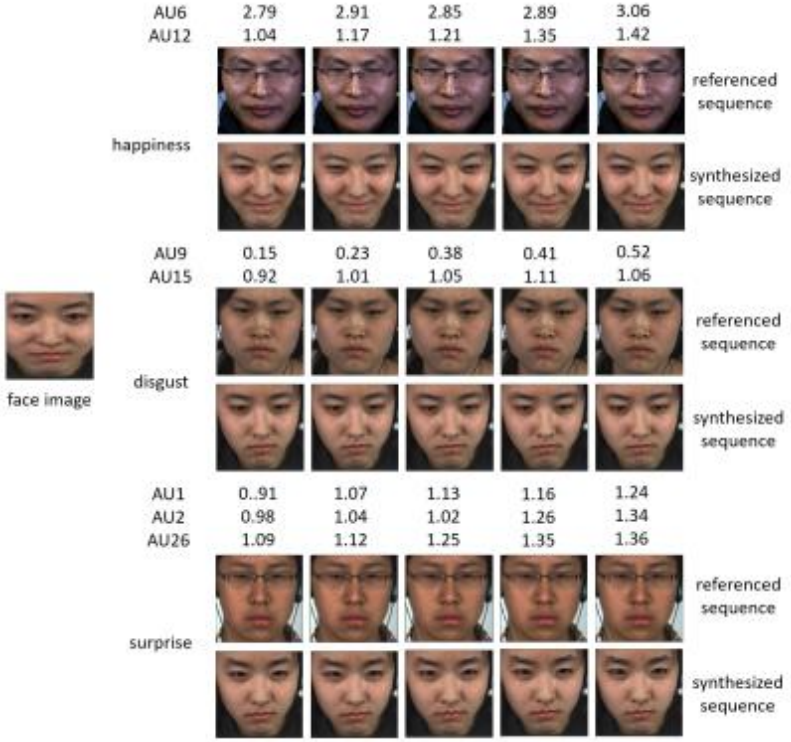
Deep Learning methods: AU assisted Graph CNN



- Two main modules:
  - Representation learning module – typical feature learning + GAP
  - AU graph relation learning module – AU node features passes thru GCNs, essential nodes are filtered through SAGPOOL
- AU-ME supervised loss (AU loss + ME loss)
- Data augmentation: Synthetic data is generated via a GAN using intensity of different AU combinations

Xie, H. X., Lo, L., Shuai, H. H., & Cheng, W. H. (2020). AU-assisted Graph Attention Convolutional Network for Micro-Expression Recognition. In Proceedings of the 28th ACM International Conference on Multimedia (pp. 2871-2880).

Deep Learning methods: AU assisted Graph CNN



Dataset	Happiness	Disgust	Surprise	Fear	Sadness	Anger	repression	Contempt	Others	Total
CASME II	32	63	28	2	4	-	27	-	99	255
synthetic CASME II	384	353	388	294	307	-	389	-	317	2432
SAMM	26	9	15	8	6	57	-	12	26	159
synthetic SAMM	264	281	275	282	284	233	-	278	264	2161

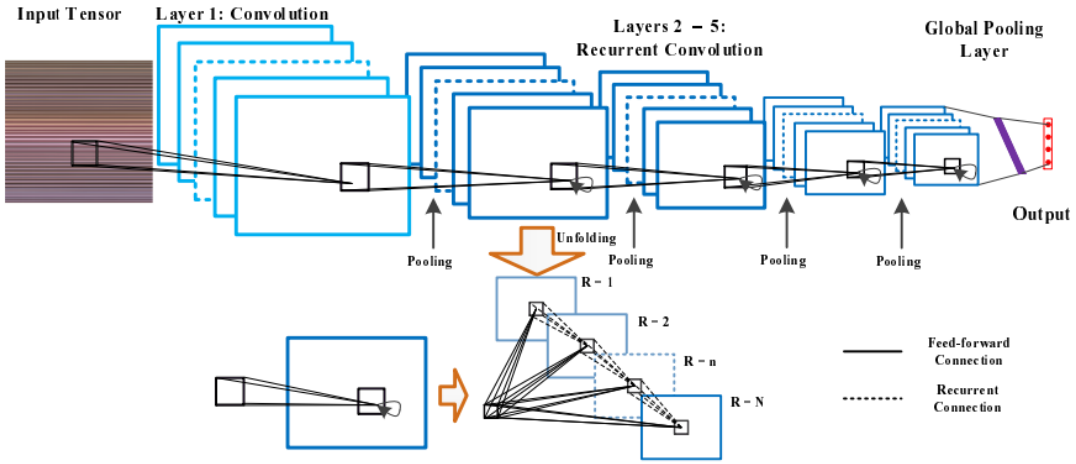
Table 1: A summary of the amount of training samples in real-world and the proposed synthetic dataset.

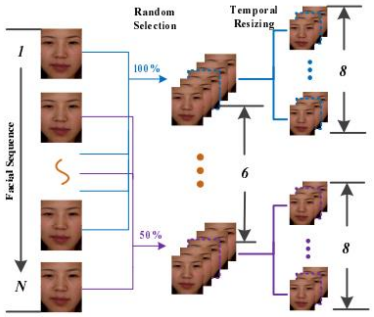
Xie, H. X., Lo, L., Shuai, H. H., & Cheng, W. H. (2020). AU-assisted Graph Attention Convolutional Network for Micro-Expression Recognition. In Proceedings of the 28th ACM International Conference on Multimedia (pp. 2871-2880).

Factor (II): Depth

- **Q**: Are "**deep**" architectures suitable?
  - ME datasets are generally small in size, even with data augmentation, it may still easily over-fit complex models
  - Running LOSO is really taxing. CASME II has 26 subjects ➔ 26-fold, 26 models were trained since 26 experiments required
- Some methods use off-the-shelf architectures (proven great in ImageNet)
- Some methods customize "**shallow**" architectures to cope with the data/experimental scenario, while maintaining convolutional mechanism

Deep Learning methods: CNN with Recurrent Connections



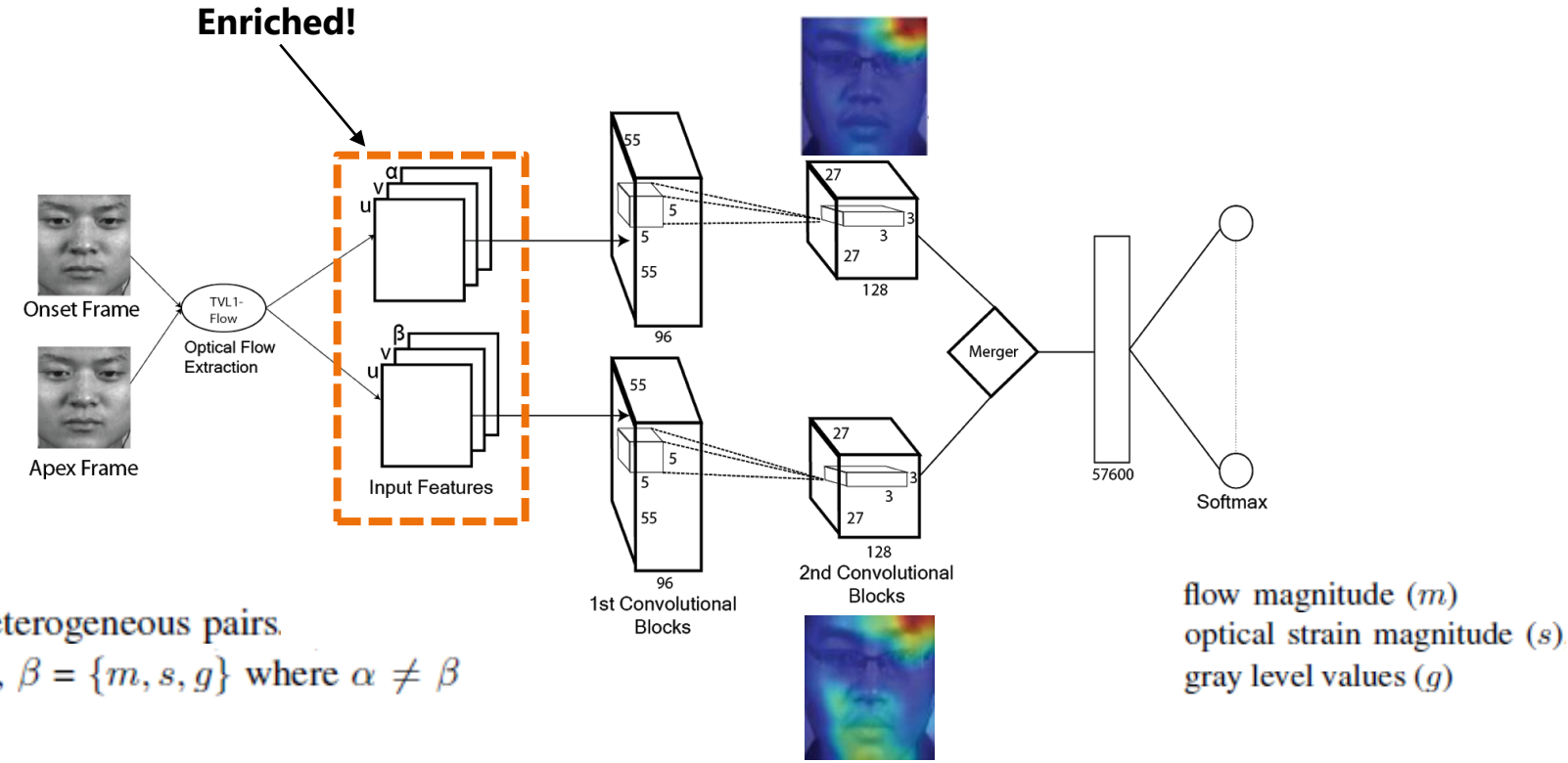


Approaches	Evaluation		
	SMIC	CASME	CASME2
LBP-TOP[11]	0.537	0.577	0.592
LBP-SIP [29]	0.445	0.368	0.466
TICS[28]	0.561	0.618	0.623
MDMO [13]	**0.640**	0.573	0.584
Pre-trained CNNs [30]	0.301	0.376	0.304
CNNs[17]	0.325	0.471	0.491
MER-RCNN (Ours)	0.571	**0.632**	**0.658**

- "Recurrent convolutional layers" (RCL) – Transfer learning from existing object and facial expression based CNN models
- Adding recurrent connections (i.e. RCNNs) within the convolutional layers ➔ capture temporal changes of convolutional features
- Data augmentation strategy ➔ "Temporal jittering" (random selection of % of frames + down/up-sampling)

Xia, Z., Feng, X., Hong, X., & Zhao, G. (2018, November). Spontaneous facial micro-expression recognition via deep convolutional network. In 2018 Eighth International Conference on Image Processing Theory, Tools and Applications (IPTA) (pp. 1-6).

Deep Learning methods: Shallow CNNs



- Motivation: Shallow networks can overcome over-fitting issue in most ME datasets

Khor, H. Q., See, J., Liong, S. T., Phan, R. C., & Lin, W. (2019). Dual-stream shallow networks for facial micro-expression recognition. In 2019 IEEE International Conference on Image Processing (ICIP) (pp. 36-40).

Deep Learning methods: Dual Stream Shallow Networks

	Methods \ Dataset	CASME II		SMIC		SAMM		Parameters (Million)
		Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	
Hand-crafted	LBP-TOP (baseline)	0.3968	0.3589	0.4338	0.3421	0.3968	0.3589	-
	LBP-SIP	0.4656	0.448	0.4451	0.4492	-	-	-
	Bi-WOOF [2]	0.5789	0.6100	0.6220	0.6200	-	-	-
	Hier. STLBP-IP [6]	0.6383	0.6110	0.6010	0.6130	-	-	-
	Bi-WOOF + Phase [7]	0.6255	0.6500	**0.6829**	**0.6730**	-	-	-
	EVM + HIGO [8]	0.6721	-	**0.6829**	-	-	-	-
Deep learning	ELRCN [13]	0.5244	0.5000	-	-	-	-	219
	CNN-LSTM [12]	0.6098	-	-	-	-	-	4.52
	AlexNet (baseline)	0.6296	0.6675	0.5976	0.6013	0.5294	0.4260	62.38
	SSSN	**0.7119**	0.7151	0.6341	0.6329	0.5662	0.4513	**0.63**
	DSSN	0.7078	**0.7297**	0.6341	0.6462	**0.5735**	**0.4644**	**0.97**

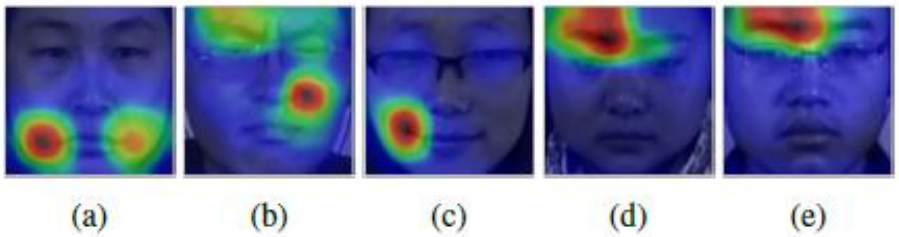
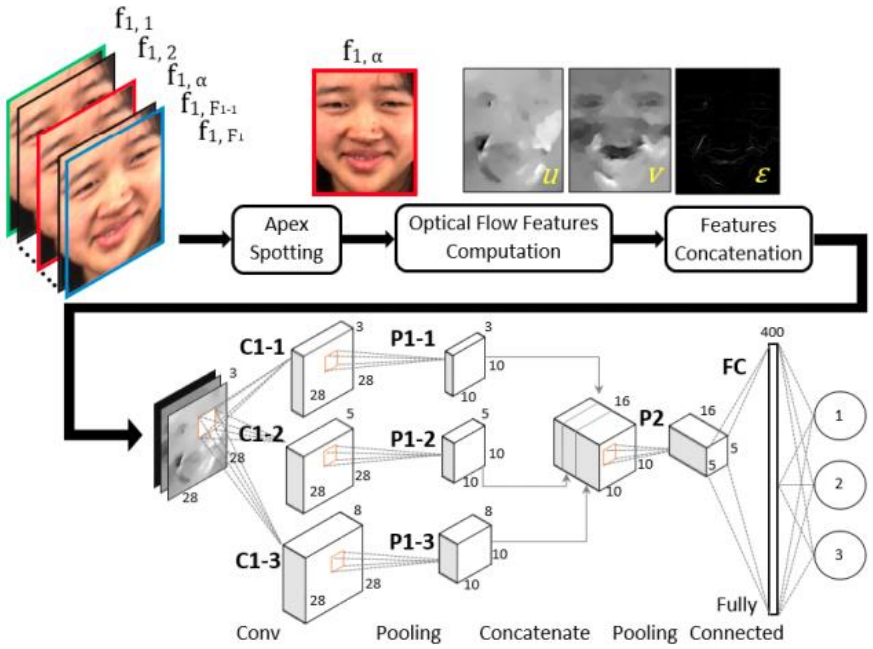


Fig. 5: Visualization of activations for DSSN (on CASME II) after *Multiply* merge on the last conv blocks. (a) Repression, (b) Disgust, (c) Happiness, (d) Others, (e) Surprise. Elaboration on the AUs are provided in the supplementary materials.

Khor, H. Q., See, J., Liong, S. T., Phan, R. C., & Lin, W. (2019). Dual-stream shallow networks for facial micro-expression recognition. In 2019 IEEE International Conference on Image Processing (ICIP) (pp. 36-40).

Deep Learning methods: Shallow Triple Stream 3D CNN



Network	Depth	Parameter (Million)	Image Input Size	Execution Time (*s*)
STSTNet	**2**	**0.00167**	28 × 28 × 3	5.7366
OFF-ApexNet [7]	5	2.77	28 × 28 × 2	5.5632
AlexNet [12]	8	61	227 × 227 × 3	12.9007
SqueezeNet [10]	18	1.24	227 × 227 × 3	14.3704
GoogLeNet [28]	22	7	224 × 224 × 3	29.3022
VGG16 [27]	16	138	224 × 224 × 3	95.4436

No.	Methods	Full			
		Acc	F1-score	UF1	UAR
1	LBP-TOP baseline	-	-	0.5882	0.5785
2	Bi-WOOF [21]	0.6833	0.6304	0.6296	0.6227
3	AlexNet [12]	0.7308	0.6959	0.6933	0.7154
4	SqueezeNet [10]	0.6380	0.5964	0.5930	0.6166
5	GoogLeNet [28]	0.6335	0.5698	0.5573	0.6049
6	VGG16 [27]	0.6833	0.6439	0.6425	0.6516
7	OFF-ApexNet [7]	0.7460	0.7104	0.7196	0.7096
8	**STSTNet**	**0.7692**	**0.7389**	**0.7353**	**0.7605**

Layer	Filter size	# Filters	Stride	Padding	Output size
C1-1	3 × 3 × 3	3	[1,1]	1	28 × 28 × 3
C1-2	3 × 3 × 3	5	[1,1]	1	28 × 28 × 5
C1-3	3 × 3 × 3	8	[1,1]	1	28 × 28 × 8
P1-1	3 × 3	-	[3,3]	1	10 × 10 × 3
P1-2	3 × 3	-	[3,3]	1	10 × 10 × 5
P1-3	3 × 3	-	[3,3]	1	10 × 10 × 8
P2	2 × 2	-	[2,2]	0	5 × 5 × 16
FC	-	-	-	-	400 × 1
Softmax	-	-	-	-	3 × 1

- 3D convolutions on 5-channel input volume, all homogenous 3x3 filters, small FC layer.
- Just over a thousand parameters. Efficient!
- 2nd place in MEGC 2019 Recognition Challenge (composite database)

Liong, S. T., Gan, Y. S., See, J., Khor, H. Q., & Huang, Y. C. (2019). Shallow triple stream three-dimensional cnn (ststnet) for micro-expression recognition. In 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019) (pp. 1-5).

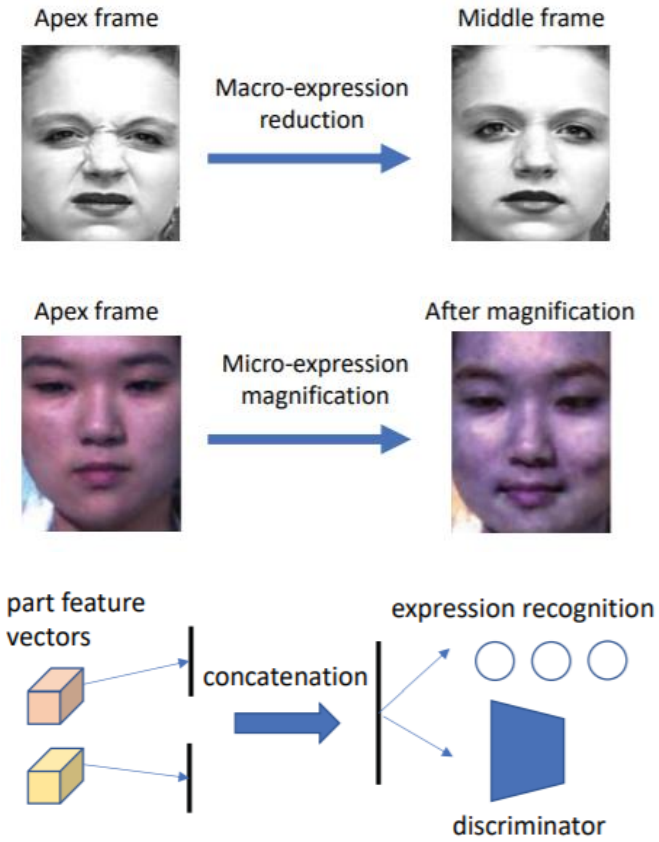
Factor (III): Data Domain

- **Q**: Should models be learned from a similar domain or from multiple (related) domains?
  - **Intra**-domain: Learning from micro-expression domain only
  - **Inter**-domain: Learning from other relevant domains – macro- (or 'normal') expressions, synthetic data/avatars, etc.

- Transferring macro-expression model to micro-expression
  - Training is done by acquiring ~10K images from several macro-expression datasets (**CK+, Oulu-CASIA, Jaffe, MUGFE**), most frames near the apex frame were selected.
  - Then, transfer learning is done on the micro-expression datasets
  - Best method in MEGC 2018 on both HDE and CDE protocols

Peng, M., Wu, Z., Zhang, Z., & Chen, T. (2018). From macro to micro expression recognition: Deep learning on small datasets using transfer learning. In 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018) (pp. 657-661).

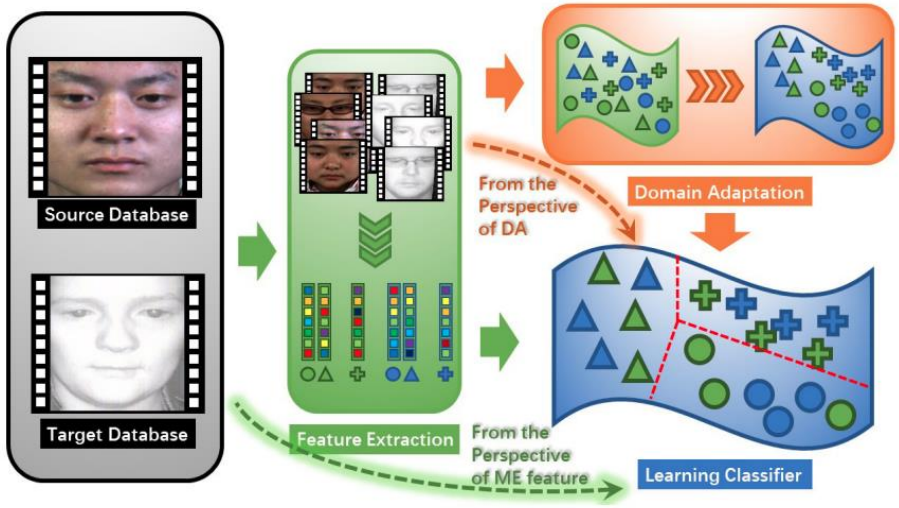
Deep Learning methods: Domain Adversarial Network



- Motivated by Ganin's "domain adversarial networks"
- Bridging macro- and micro-expression domains via **Expression Magnification and Reduction (EMR)**
  - Reduction: Picking middle frame between onset and apex of macro sample (relabeled **CK+** dataset)
  - Magnification: Magnifying micro sample
- Upper and lower part of tensor are extracted separately, merged at FC level
- Adversarial loss is computed between last feature vector of macro- and micro- sample
- 1st place in MEGC 2019 Recognition Challenge (composite database)

Liu, Y., Du, H., Zheng, L., & Gedeon, T. (2019). A neural micro-expression recognizer. In 2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019) (pp. 1-4).

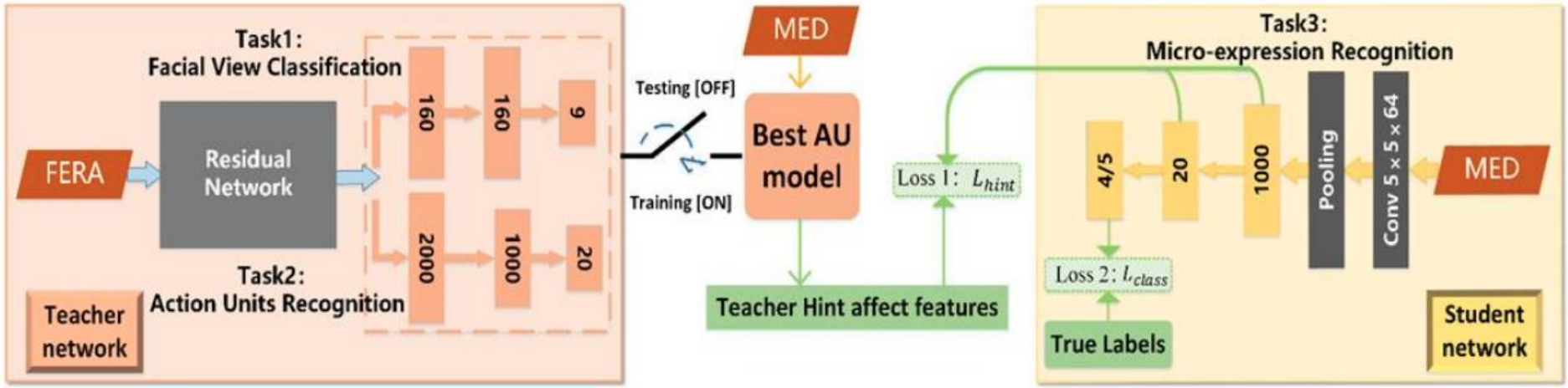
Deep Learning methods: Domain Regeneration for Cross-DB



- Feature distribution from a source database is regenerated closely to the feature distribution of target database via subspace learning.
- Initiation requires both the features to be mapped into a Reproduced Kernel Hilbert Space (RKHS)
- Feature distribution is minimized via optimizing Maximum Mean Discrepancy (MMD) defined in RKHS.

Zong, Y., Zheng, W., Huang, X., Shi, J., Cui, Z., & Zhao, G. (2018). Domain regeneration for cross-database micro-expression recognition. IEEE Transactions on Image Processing, 27(5), 2484-2498.

Deep Learning methods: Knowledge Distillation



(A)

SMIC2	Training from scratch	Fine-tune-AUCNN	TS-AUCNN
Average ACC(%)	58.82	-	**76.06**
F_1-score	0.54	-	**0.71**

(B)

CASME	Training from scratch	Fine-tune-AUCNN	TS-AUCNN
Average ACC(%)	66.67	75.1	**81.8**
F_1-score	0.58	0.72	**0.77**

(C)

CASME II	Training from scratch	Fine-tune-AUCNN	TS-AUCNN
Average ACC(%)	58.38	66.55	**72.61**
F_1-score	0.45	0.60	**0.67**

(D)

SAMM	Training from scratch	Fine-tune-AUCNN	TS-AUCNN
Average ACC(%)	66.67	84.08	**86.74**
F_1-score	0.61	0.79	**0.83**

- Also concluded that "crucial" temporal sequences is better than whole video for ME recognition!

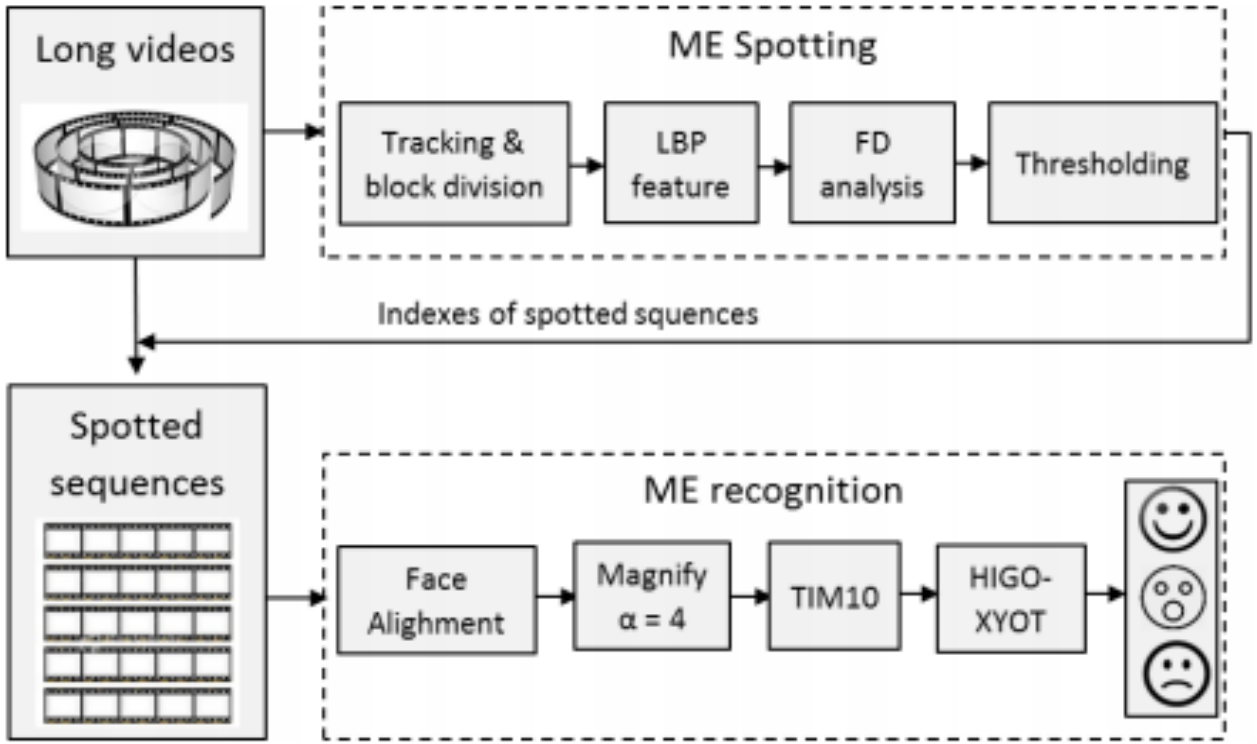
- **(Teacher)** A larger network fine-tuned with larger datasets. (**FERA2017**)
- **(Student)** A smaller network fine-tuned on micro-expression, with *hint information* on the FC-level.

Sun, B., Cao, S., Li, D., He, J., & Yu, L. (2020). Dynamic Micro-Expression Recognition Using Knowledge Distillation. IEEE Transactions on Affective Computing.

“Spot-then-recognize” approaches

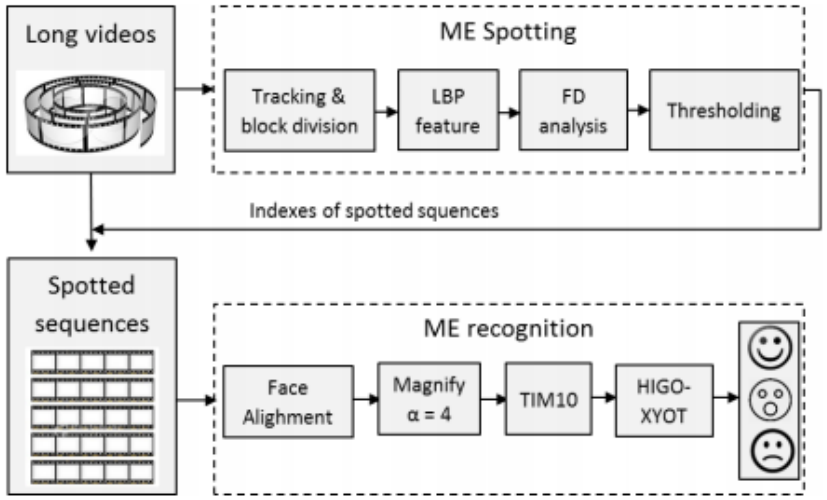
- Both ME analysis tasks (spotting, recognition) often treated individually
- Realistic expectation and application: spot-**THEN**-recognize
  - “Where” and “What” is the emotion
- A seamless framework will allow fully automated ME analysis systems to be built

Spot-then-Recognise Pipeline



Li, X., Hong, X., Moilanen, A., Huang, X., Pfister, T., Zhao, G., & Pietikäinen, M. (2017). Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods. IEEE transactions on affective computing, 9(4), 563-577.

Spot-then-Recognise Approach



- First attempt at spot-then-recognise scheme
- **Spotting TPR = 74.86%**
- **"Spot-then-recognize" accuracy = 56.67%** using correctly spotted ME sequences
- **Overall system performance** = 74.86 x 56.67 = **42.42%**

Li, X., Hong, X., Moilanen, A., Huang, X., Pfister, T., Zhao, G., & Pietikäinen, M. (2017). Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods. IEEE transactions on affective computing, 9(4), 563-577.

Spot-then-Recognise Approach

(Li et al., 2017)

	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, i.e. without spotting)

= **67.21%** (CASMEII)

Li, X., Hong, X., Moilanen, A., Huang, X., Pfister, T., Zhao, G., & Pietikäinen, M. (2017). Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods. IEEE transactions on affective computing, 9(4), 563-577.

Spot-then-Recognise Approach

(Li et al., 2017)

Benchmarking via Human Test

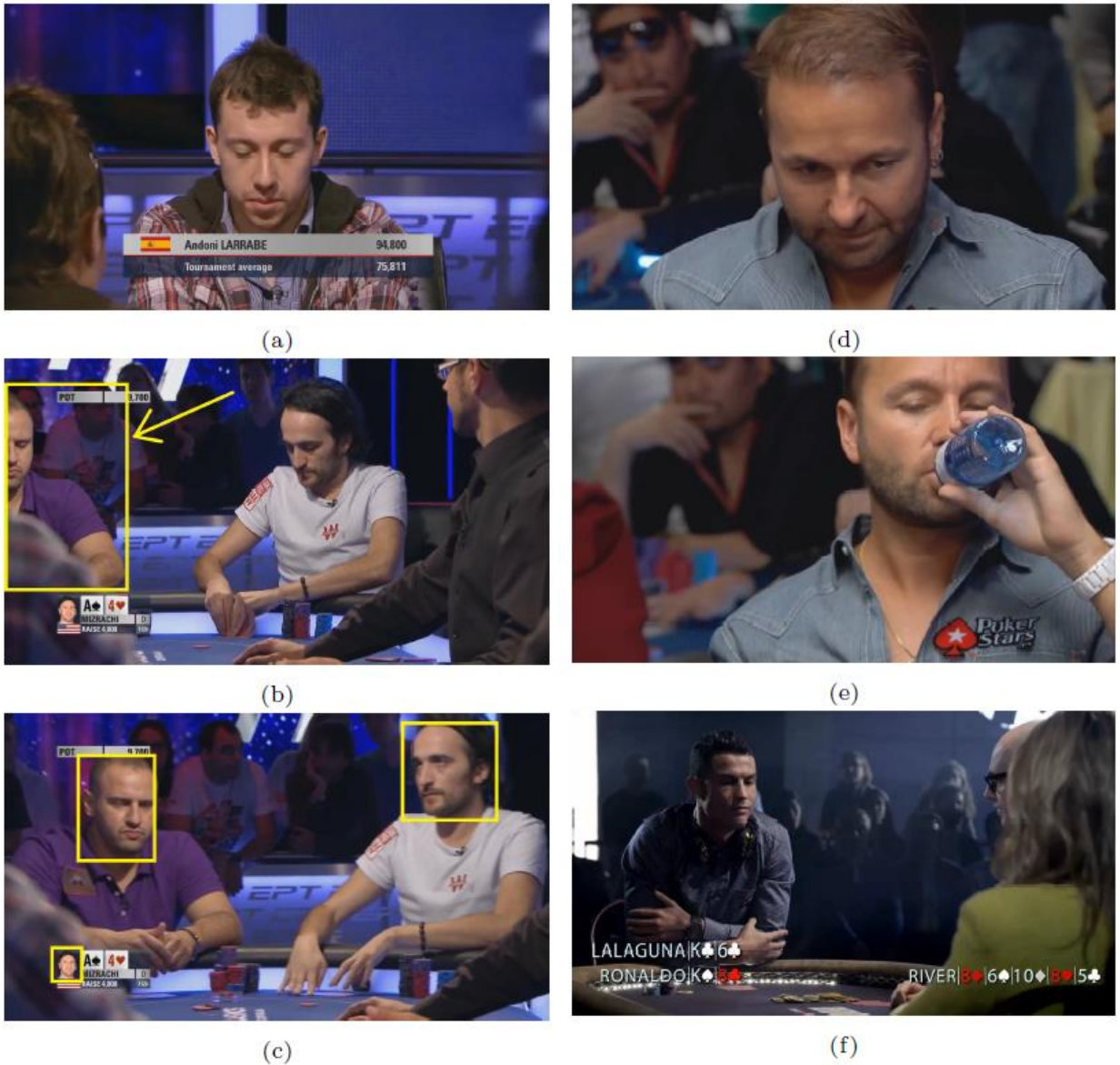
- 15 subjects (avg. age 28.5 years, 10 male, 5 females)
- Definition of emotions explained, ME clips from SMIC-VIS were shown, subjects asked to select their answers after watching them
- Mean accuracy = **72.11%** (SMIC-VIS accuracy using proposed method = **81.69%)**

Insights:

- A very first attempt at a combined spotting and recognition pipeline
- Pros: Gives a glimpse of possibility of real-world practical applications
- Limitations: Problems in spotting (fixed spotting intervals, non-ME movements) hamper recognition capability

Li, X., Hong, X., Moilanen, A., Huang, X., Pfister, T., Zhao, G., & Pietikäinen, M. (2017). Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods. IEEE transactions on affective computing, 9(4), 563-577.

Spotting “in-the-wild” on MEVIEW database



a) Eye blinking
b) Partially occluded faces;
c) Multiple faces;
d) Head movement;
e) Drinking water
f) Inconsistency lighting

P. Husak, J. Cech, J. Matas, Spotting facial micro-expressions \in the wild", in: 22nd Computer Vision Winter Workshop (Retz), 2017.

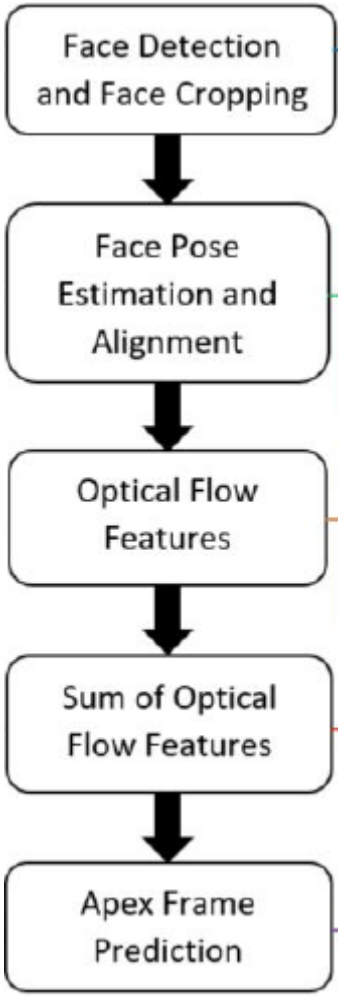
Spotting “in-the-wild” on MEVIEW database

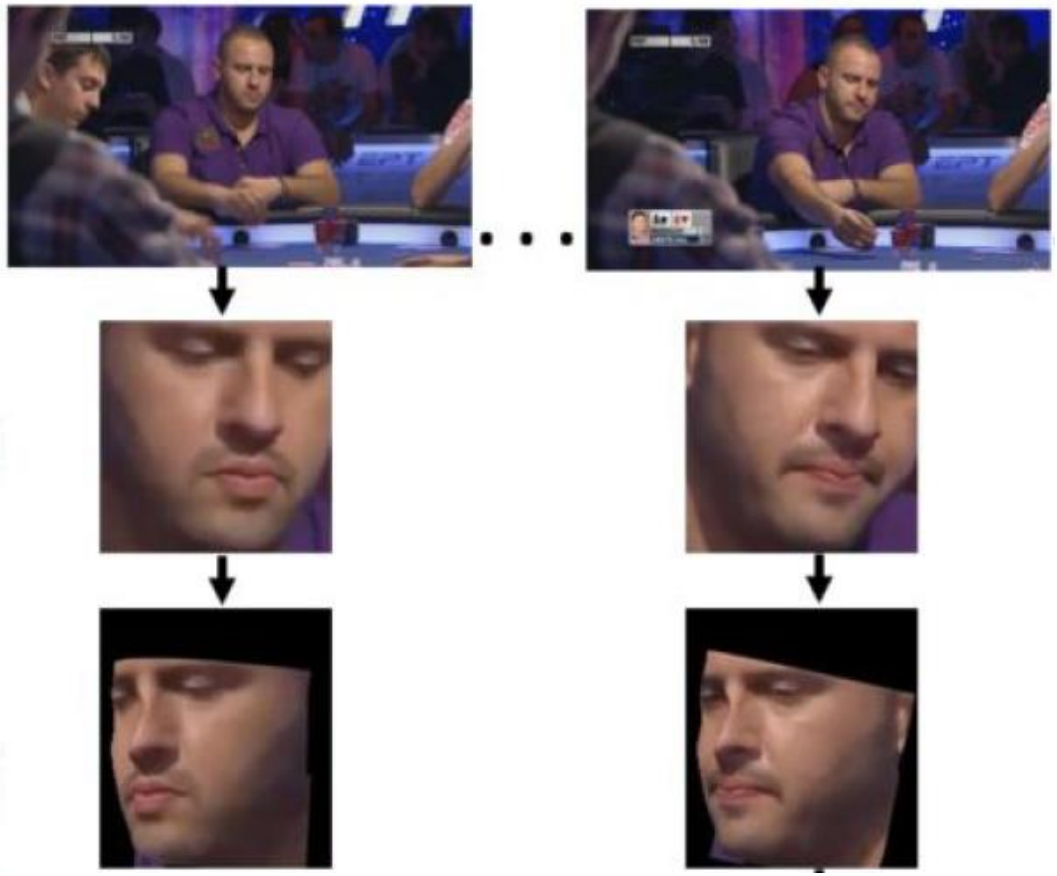
Database	SMIC-E-NIR [10]	SMIC-E-VIS [10]	SMIC-E-HS [10]	CASME II-RAW [9]	MEVIEW [13]
Year	2013	2013	2013	2014	2017
Subject	8	8	16	26	14
Sample	71	71	157	246	21
Frame rate (*fps*)	25	25	100	200	30
Elicitation environment	Constrained lab condition	Constrained lab condition	Constrained lab condition	Constrained lab condition	In-the-wild

- Political interview
- Poker venues in Barcelona, Dublin, Prague, Austria, and Las Vegas.

P. Husak, J. Cech, J. Matas, Spotting facial micro-expressions \in the wild", in: 22nd Computer Vision Winter Workshop (Retz), 2017.

Spotting “in-the-wild” on MEVIEW database





P. Husak, J. Cech, J. Matas, Spotting facial micro-expressions \in the wild", in: 22nd Computer Vision Winter Workshop (Retz), 2017.

Spotting “in-the-wild” on MEVIEW database

	Database	Apex Spotting	Recognition
		ASR	F1-score
Constrained Lab Condition	SMIC-E-VIS [10]	0.28	0.53
	SMIC-E-NIR [10]	0.27	0.43
	SMIC-E-HS [10]	0.38	0.47
	CASME II-RAW [9]	0.82	0.59
In-the-wild	**MEVIEW** [13]	**0.33**	**0.67**

End of Part 4

Questions?