Building the Facial Expressions Recognition System Based on RGB-D Images in High Performance

Trung Truong and Ngoc Quoc Ly
Faculty of Information Technology
University of Science, VNUHCM-HCMUS

Outline

- Motivation
- Problem Statement
- Method
- Experiments
- Conclusions
- Future work

- Develop security cameras & surveillance system
- Expression is a part of behavior. It contributes to enhance the understanding of image and video in intelligent system.
- Behavioral science research
  - Automation of objective measurement of facial activity

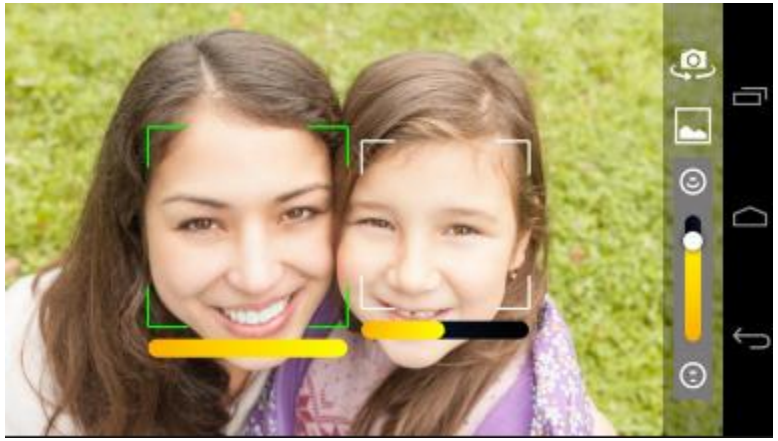
Application:

- More human-like human-computer, and human-robot interaction (e.g.: elder care robot).
- Detect terrorist attacks (e.g.: security at airport)
- Face Retrieval based on attributes system

Application:

Robot teaches autistic kids interact

human-computer-interaction



SmileCam - smile detector

- Input: the facial expression one in face images
  - Grayscale: $I = \begin{bmatrix} a_{00} & \cdots & a_{0m} \\ \vdots & \ddots & \vdots \\ a_{n0} & \cdots & a_{nm} \end{bmatrix}$
  - Depth image: $D = \begin{bmatrix} z_{00} & \cdots & z_{0m} \\ \vdots & \ddots & \vdots \\ z_{n0} & \cdots & z_{nm} \end{bmatrix}$
- Ouput: Define a classification function f so that:
  - $f(I, D) \in$ {anger, fear, surprise, sadness, joy and disgust}

Let: a_{ij}, z_{ij} are a point in grayscale and depth image respectively.

Traditional method based on 2D image

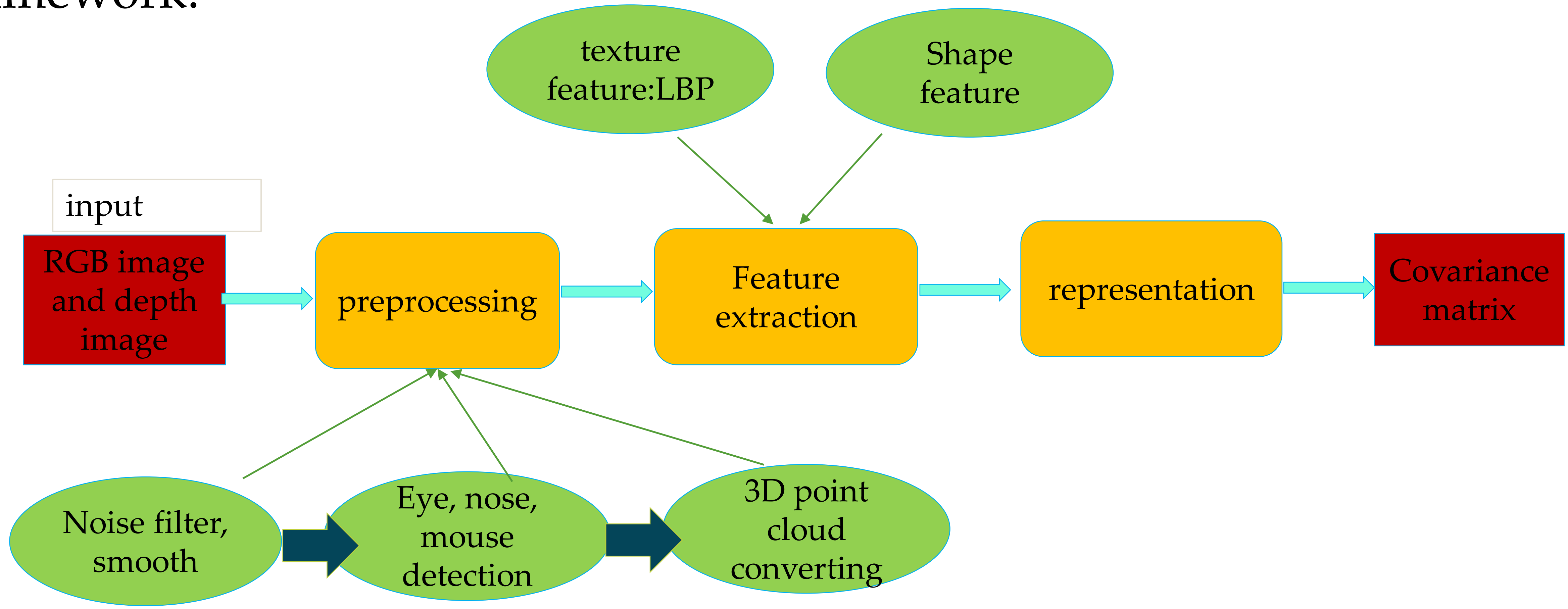
Recent method based on 3D image

- Challenges in 2D image:
  - Pose
  - Illumination
  - Occlusion

- Challenges in 3D image:
  - ~~Pose~~
  - ~~Illumination~~
  - Occlusion

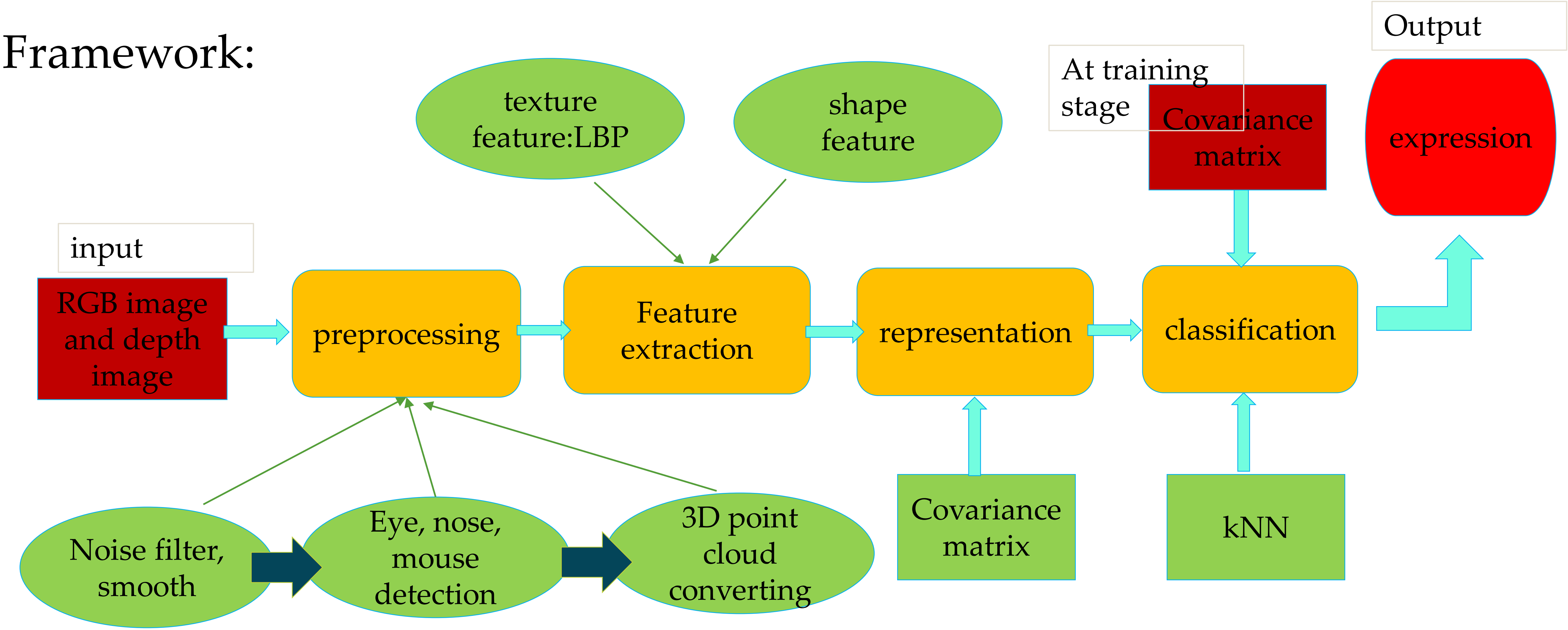
Kinect device

Framework:



Training stage

Framework:



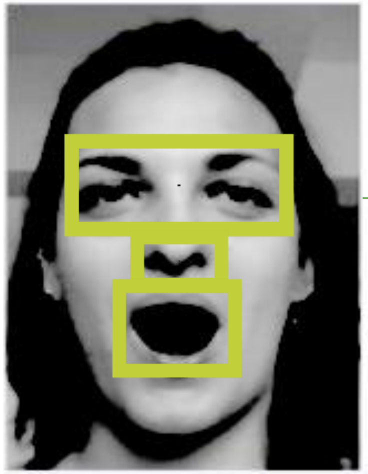
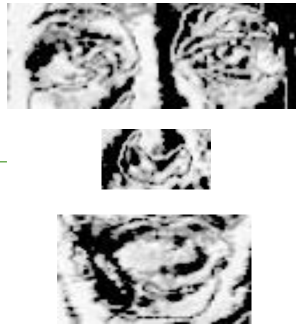
Testing stage

- **Eye, nose, mouth are detected by utilizing Active Shape Models with Stasm lib ver 4.1.0. [6]**
- **Feature extraction in RGB image**
  - Local binary pattern
  - Introduced by Ojala et al. [2]
- $LBP_{8,1}(x_c, y_c) = \sum_{i=0}^{7} s(I_i - I_c) * 2^i$

Where

- I_i, I_c are the grey level values at c, i.
- i is the label of parts around the center pixel location (xc, yc) and $s(x) = \begin{cases} 1 \; if \; x \geq 0 \\ 0 \; if \; x < 0 \end{cases}$

- **Pros:** computational simplicity, extracting texture feature (i.e.: furrow..)
- **Cons:** illumination, local feature

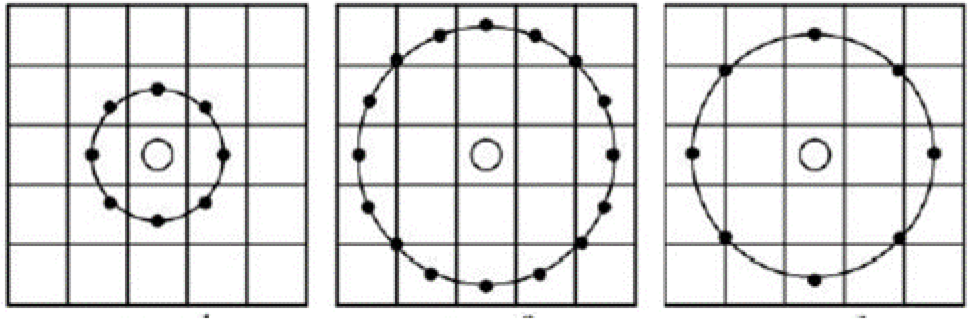
Original image and $LBP_{8,1}$ image

Feature extraction in RGB image

- *Using Neighborhoods of Different Sizes Extended LBP*
- *Multi-scale LBPs*

We propose:

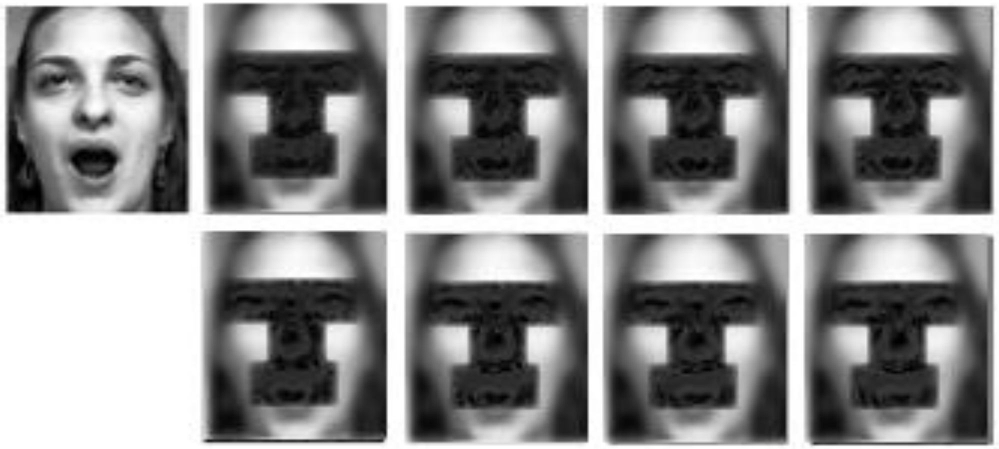
- $LBP_1^{ms}(8, R)$ with R = 1..8.
- $LBP_2^{ms}(8, R)$, LBP (8, R)
- with R=1, 2 and LBP(16,R) with R=2, 3, 4.



LBP (8, 1) LBP (16, 2) LBP (8, 2)

Feature extraction in RGB image

- *Using Neighborhoods of Different Sizes Extended LBP*
- *Multi-scale LBPs*



$LBP_1^{ms}(8, R)$ with R = 1..8.

Left to right, original image and $LBP_2^{ms}(8, R)$

Pros:

- increase discriminative power
- handle variations in illumination
- real-time processing
- computational simplicity

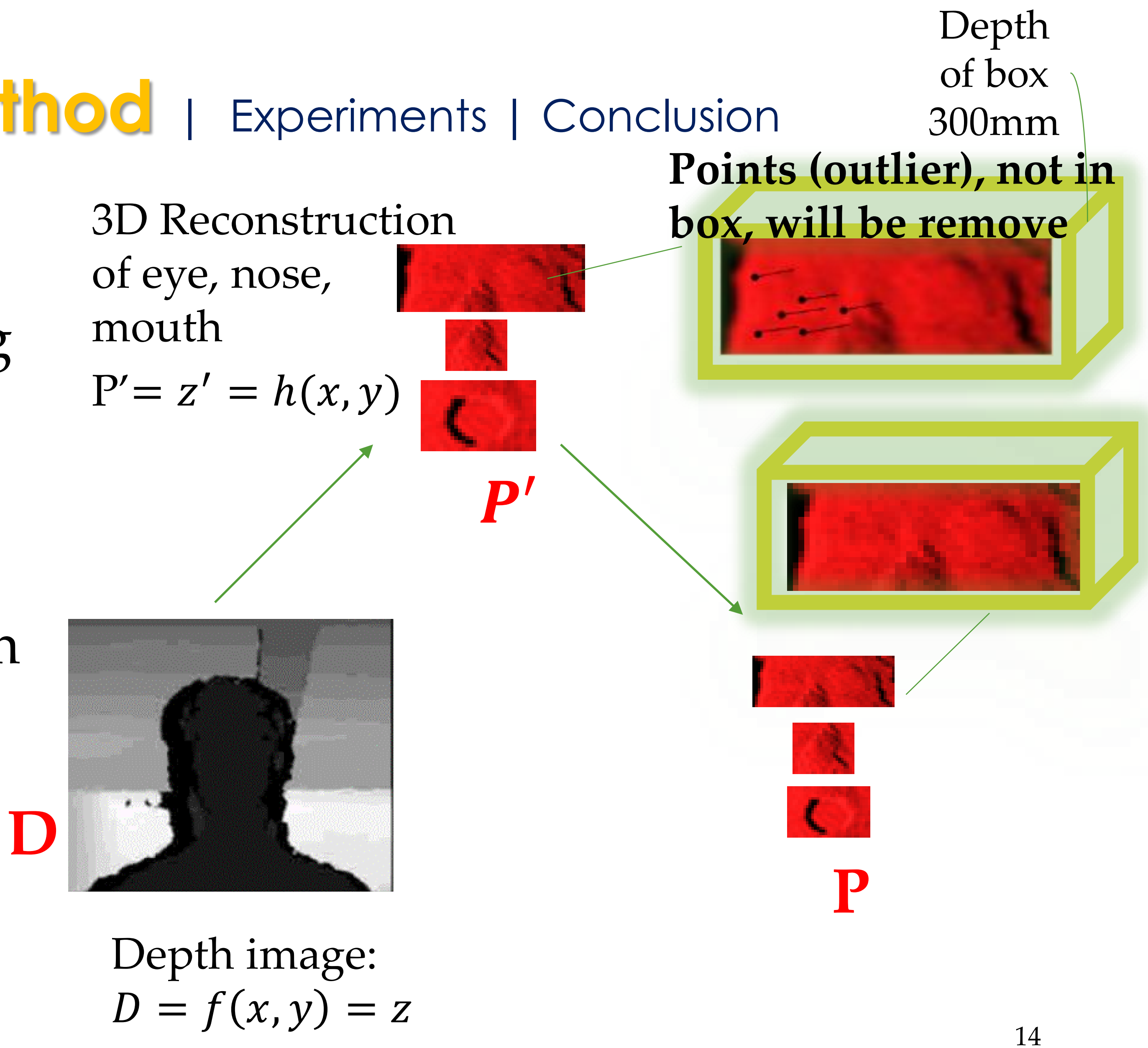
Normal Estimation on Point Clouds

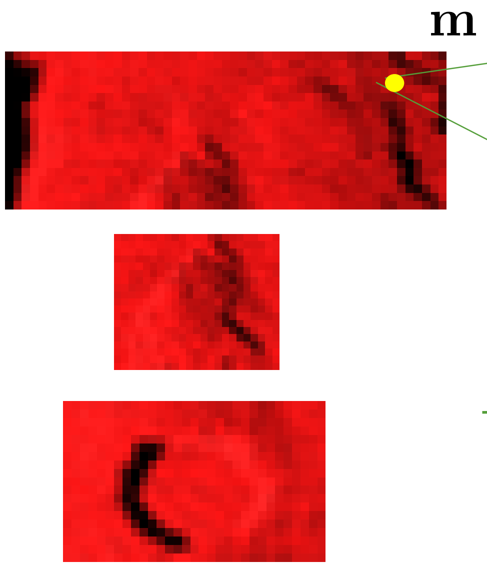
Depth image

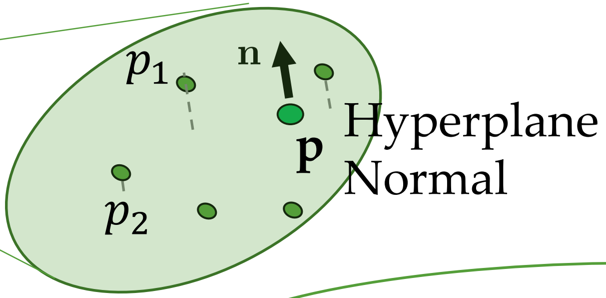
3D point cloud

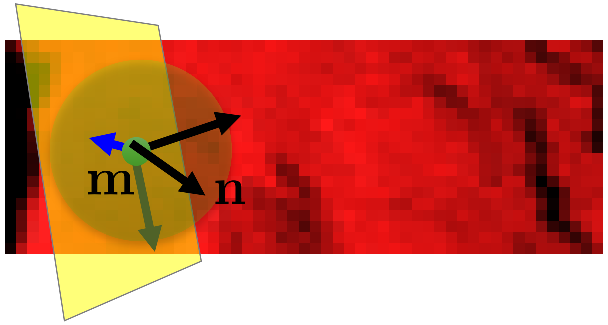
Normal descriptor

- Input: depth image 3D
- Applying morphological closing fill holes.
- Gaussian filter is applied for smoothing depth image
- Reconstruct 3D point cloud from Camera Calibration and depth image P'
- Remove outlier: P





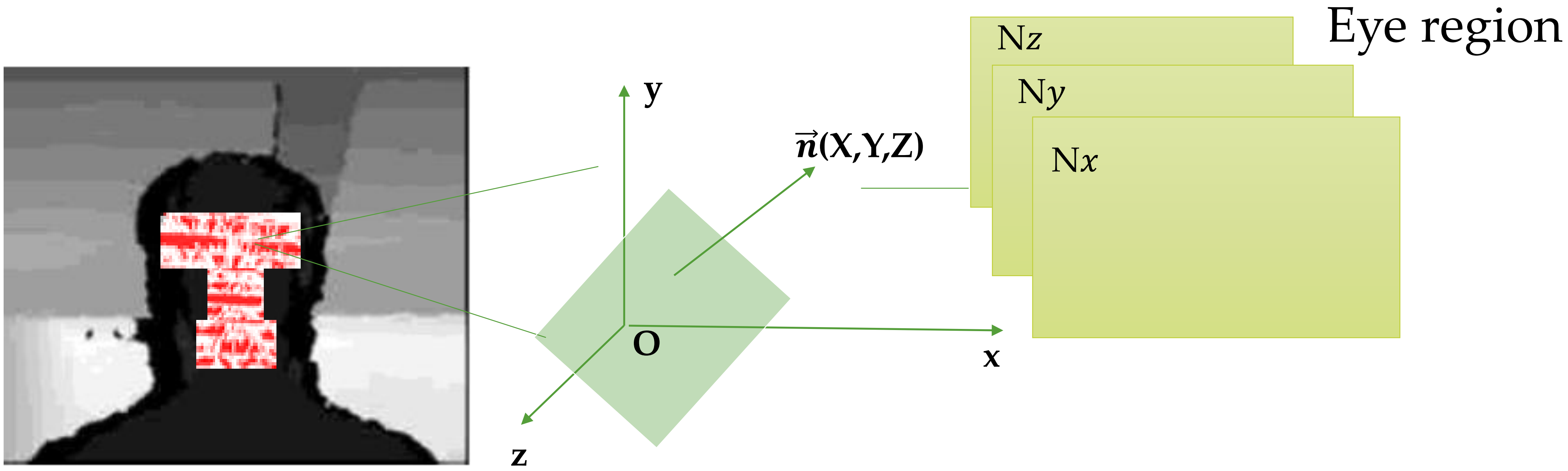




- Point cloud P
- Normal vector at point p.
- Its exists: $p_1, p_2.., p_k \in P^k$ are number of nearest neighbors of point p

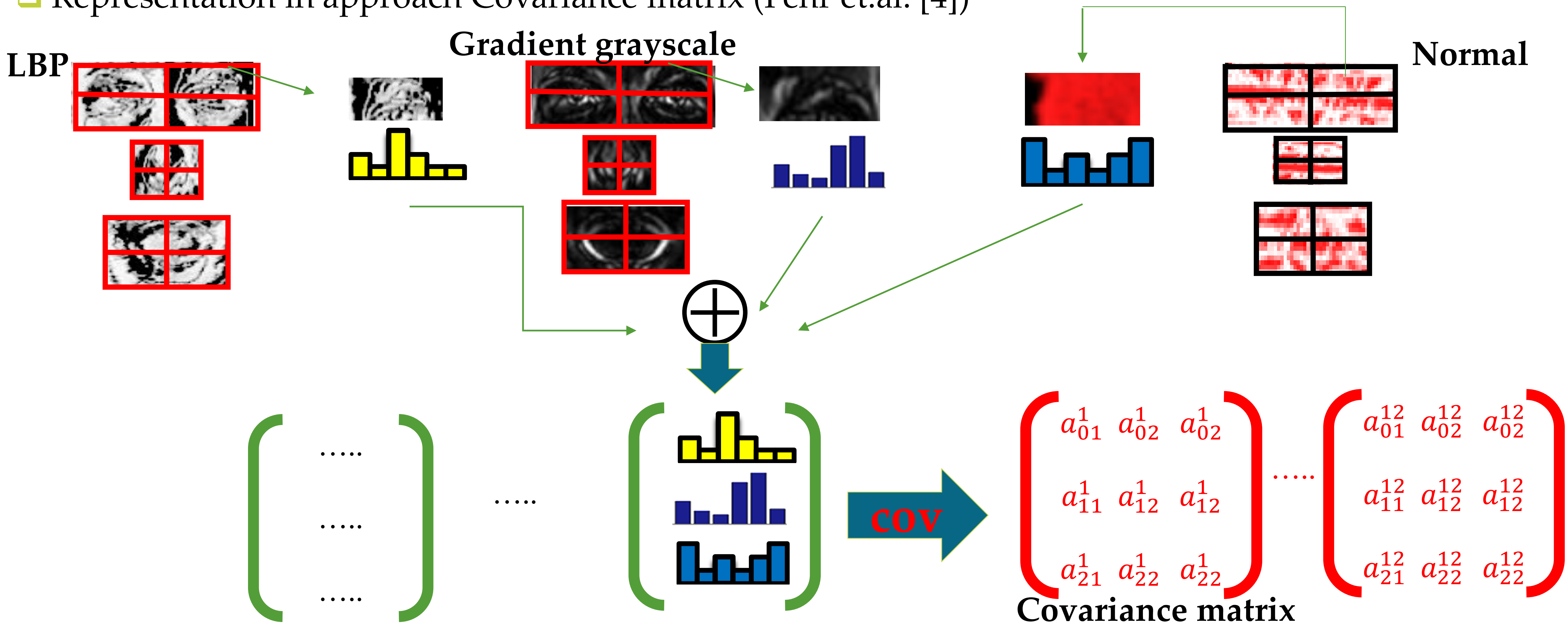
- ❑ Jusu [7] propose theoretical primer for estimating surface normals in a point cloud.
- ❑ Compute centroid = $\bar{p} = \frac{1}{k}\sum_{i=1}^{k} p_i$
- ❑ Compute scatter matrix
  - $C = \frac{1}{k}\sum_{i=1}^{k}(p_i - \bar{p})(p_i - \bar{p})^T$
  - $C * \vec{v}_j = \boldsymbol{\lambda}_j * \vec{v}_j, \mathrm{j} \in \{0,1,2\}$

  Where, $\boldsymbol{\lambda}_j$ is the j-th eigenvalue of the covariance matrix, and $\vec{v}_j$ the j-th eigenvector
  - The vector corresponding to the smallest eigenvalue is **the normal n**



Let (Nx, Ny, Nz) are the first, second and third coordinate of the normal vector.

- Representation in approach Covariance matrix (Fehr et.al. [4])



Discussion

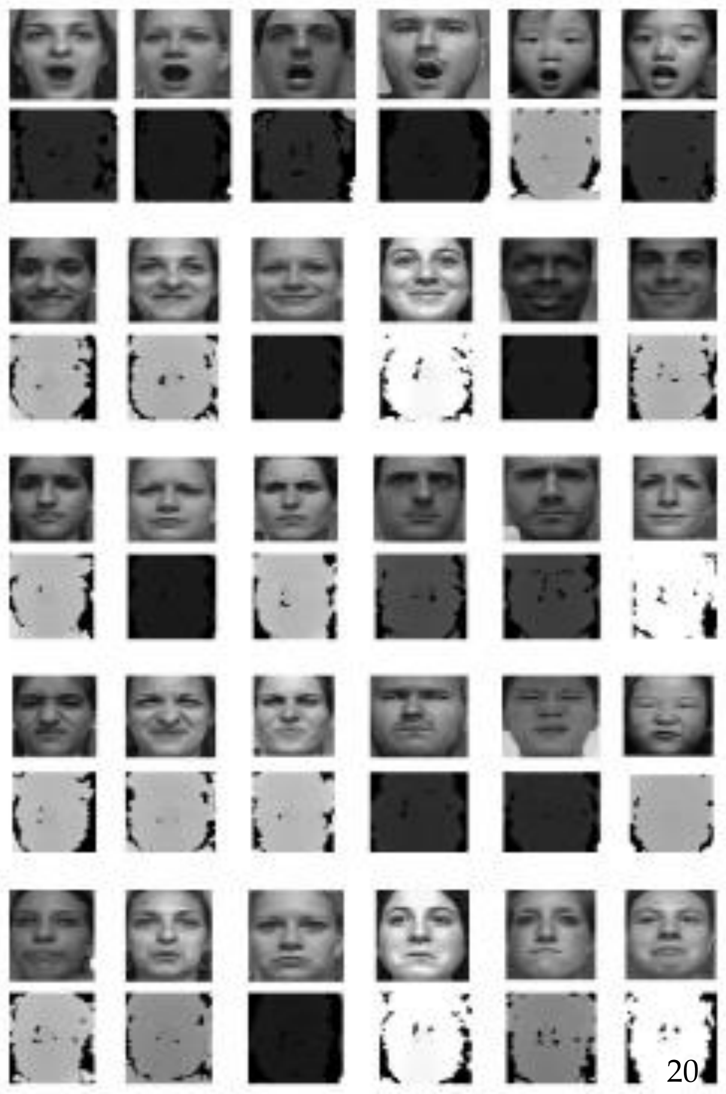
- Why the earlier combination of LBP and normal descriptor by represented by covariance descriptor attain high performance than others?
- LBP is a kind of texture feature
- Normal descriptor is a kind of shape feature

- Utilities kNN algorithm for facial expression recognition (Fehr et.al. [4])
- In range of k = 9..14 have highest accuracy
- Formula: $\rho(C^G, C^P) = \sum_{i=1}^{12} \rho\left(C_i^G, C_i^P\right) - max_i[\rho\left(C_i^G, C_i^P\right)]$

where

- $\rho(C_1, C_2) = \sqrt{\sum_1^d ln^2 \boldsymbol{\lambda}_i (C_1, C_2),}$
- C_i^G, C_i^P are sub–regions, located in the same regions such as eyes, nose and mouth regions.

- Dataset: FaceWarehouse database [2]
- Using a Kinect RGBD camera
- Participant: 150 individuals aged 7-80 from various ethnic backgrounds
- In this paper:
- Recognize 6 facial expression: anger, fear, surprise, sadness, joy and disgust.
- Total: 42 color images and depth images per facial expression



- Using FaceWarehouse database for our approach
- Conduct 3 experiments under various illumination
  - Normalized illumination
  - Low, high illumination

Images are captured under low illuminant.

Images are captured under highly illuminant

Approach in grayscale and point cloud only under good illumination condition

Feature	Accuracy %
$F1 = [x\ y\ I\ I_x I_y\ \frac{I_x}{I_y}\ LBP(8, 2)]$ + COV + kNN	84.13
$F2 = [x\ y\ n_x n_y n_z]$ + COV + kNN	72.22

Recognition rate (%) for recognition facial expression in 3-Fold and leaveone- out cross-validation

Datasets	Accuracy	
	Low illumination	High illumination
RGB dataset	53.97	74.98
Depth dataset	72.22	72.22
RGBD dataset	**73.81**	**76.57**

2 comparative methods based on FaceWarehouse database

- Mao et al [3] ‘s Approach
  - Method: Aus, FPPs, FPs
  - Accuracy: <72% (4 expresions)

- Our approach
  - Method: Normal descriptor, COV, kNN
  - Accuracy: 72.22% (6 expressions)

❑ Our contribution:

- Propose the novel feature based on the covariance representation by fusing all of LBP, gradient and normal descriptor.
- Prove that the features are extracted in combining depth and texture images to improves the performance more considerably than both depth image-only method and color image-only method under high and poor lighting condition.

Future work

- We leave as future research the study of experiments which are based on an extension of our proposed framework in the challenge of head pose and occlusion.
- Furthermore, the proposed framework used in this paper will be investigated for a video.

References

1. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. J. Pattern Anal. Mach. Intell. **24**(7), 971–987 (2002)
2. Cao, C., Weng, Y., Zhou, S., Tong, Y., Zhou, K.: Facewarehouse: a 3D facial expression database for visual computing. IEEE Trans. J. Vis. Comput. Graph. **20**(3), 413–425 (2014)
3. Mao, Q., Pan, X., Zhan, Y., Shen, X.: Using kinect for real-time emotion recognition via facial expressions. J. Front. Inf. Tech. Electr. Eng. **16**, 272–282 (2015)
4. Fehr, D.A.: Covariance based point cloud descriptors for object detection and classification, University of Minnosita, August (2013)
5. Forstner, W., Moonen, B.: A metric for covariance matrices. In: Grafarend, E.W., Krumm, F.W., Schwarze, V.S. (eds.) Journal of Geodesy-The Challenge of the 3rd Millennium, pp. 299–309. Springer, Heidelberg (2003)
6. Milborrow, S., and F. Nicolls. "Active shape models with sift descriptors and mars. VISAPP, 2014."
7. D Rusu, R.B., Cousins, S.: 3D is here: point cloud library (PCL). In: IEEE International Conference on Proceedings of Robotics and Automation (ICRA), Shanghai, China, 9–13 May 2011, pp. 1–4. IEEE (2011)