FACIAL EXPRESSIONS IN THE WILD

Databases and Experiment Protocols

Emotion Recognition In The Wild Challenge and Workshop (EmotiW 2013)

Partial Least Squares Regression on Grassmannian Manifold for Emotion Recognition

Mengyi Liu, Ruiping Wang, Zhiwu Huang, Shiguang Shan, Xilin Chen

Institute of Computing Technology, Chinese Academy of Sciences

Outline

- Problem
- Related work
- Our Method
- Experiments
- Conclusion

Outline

- Problem
- Related work
- Our Method
- Experiments
- Conclusion

Emotion recognition in the wild

- Challenges
  - Large data variations
    - head pose, illumination, partial occlusion, etc.
  - Lack of labeled data
    - Manual annotation is hard as spontaneous expression is ambiguous in the real world.

Outline

- Problem
- Related work
- Our Method
- Experiments
- Conclusion

Video-based emotion recognition

- Acoustic information based
  - Time domain and frequency domain
    - e.g. pitch, intensity, pitch contour, Low Short-time Energy Ratio (LSTER), maximum bandwidth, …
- Vision information based
  - Spatial space and temporal space
    - e.g. Optical flow, 3D descriptor (LBP-TOP, HOG 3D), tracking based (AAM, CLM), probabilistic graph model (HMM, CRF), …

Outline

- Problem
- Related work
- Our Method
- Experiments
- Conclusion

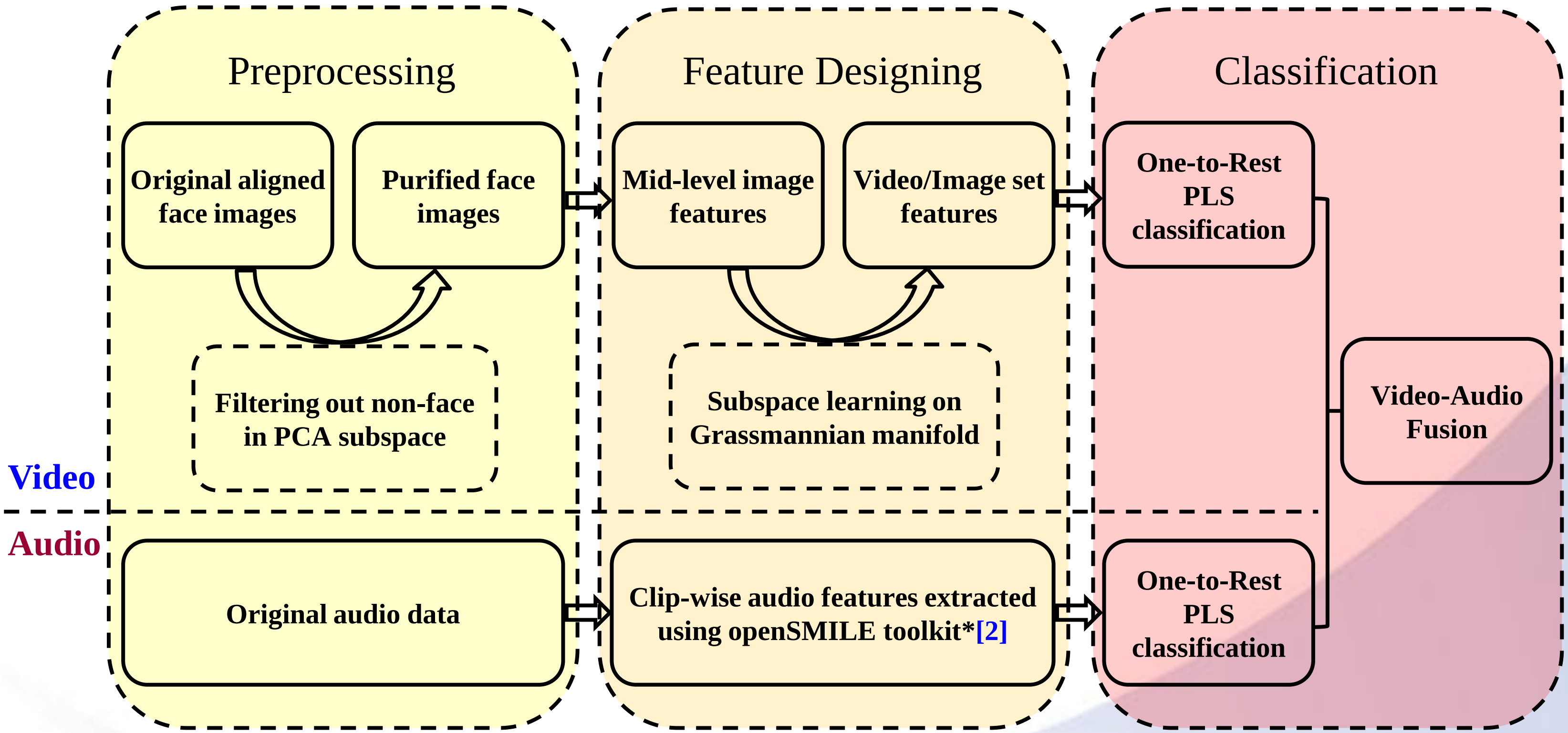
Our method

- Key issue
  - How to model the emotion video clip?
- Motivation
  - Alleviate the effect of mis-alignment of facial images
  - Encode the data variations among video frames
- Basic idea
  - Inspired by recent progress of image set-based face recognition [1]
  - Treat the video clip as an image set, i.e., a collection of frames
  - Linear subspace for video (image set) modeling

[1] R. Wang, H. Guo, L. S. Davis, and Q. Dai. Covariance discriminative learning: A natural and efficient approach to image set classification. CVPR, 2012.

Our method

- An overview



[2] F. Eyben, M. Wollmer, and B. Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. ACM MM, 2010.

Our method

- Preprocessing
  - Original face alignment using MoPS [3] (*provided by organizer*)
  - Purification of face images
    - Original aligned face images set: $X = \{x_1, x_2, \dots, x_n\}, x_i \in R^D$.
    - PCA projection learned on X by preserving low energy: W.
    - Mean reconstruction error of each image:

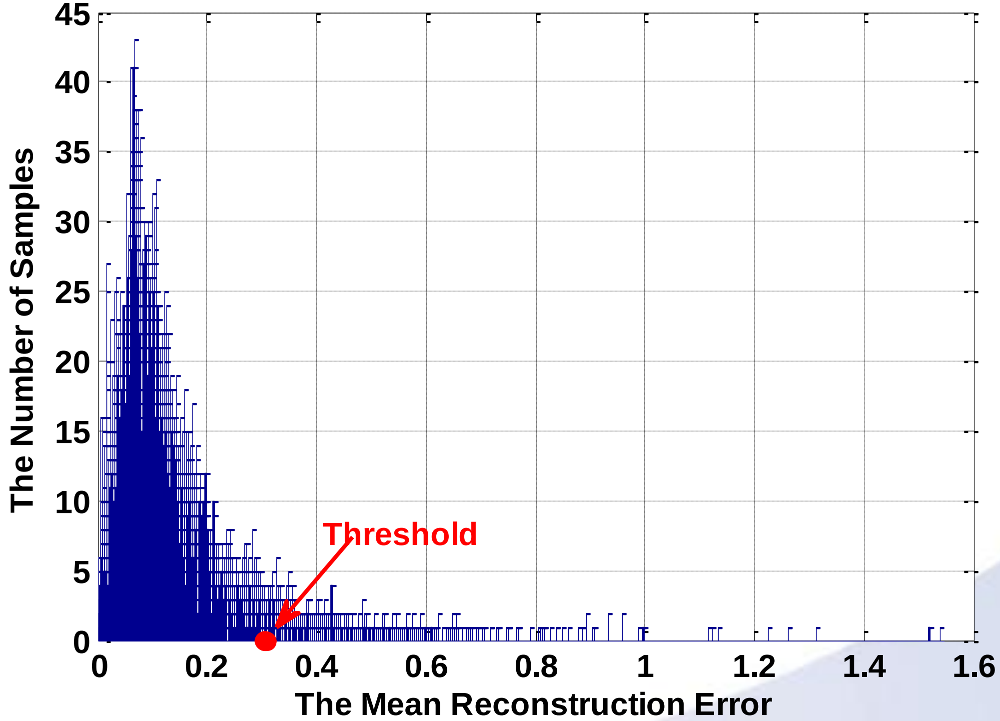
$$MeanErr_t = \frac{1}{D} \left|\left| x_t - W^T W x_t \right|\right|^2$$

    - Non-face/Badly-aligned face images tend to have large $MeanErr_t$.

[3] X. Zhu, and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. CVPR, 2012.

Our method

- Preprocessing
  - The distribution of $MeanErr_t$ on training set in EmotiW2013.



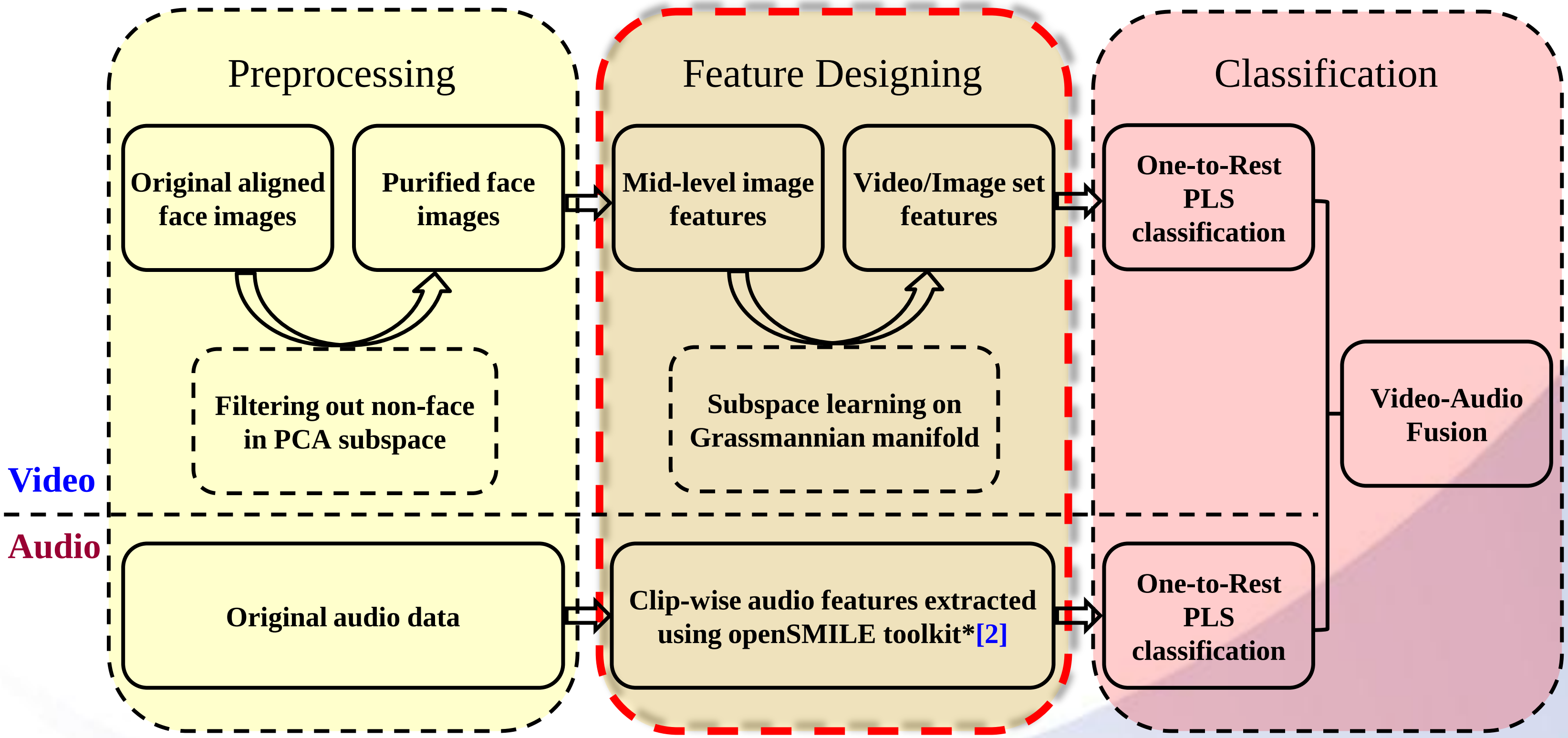
* Threshold is for filtering out non-face in PCA space.

Our method

- Preprocessing
  - An example of 100 samples with largest mean reconstruction error. Most are non-face images or mis-alignment results.

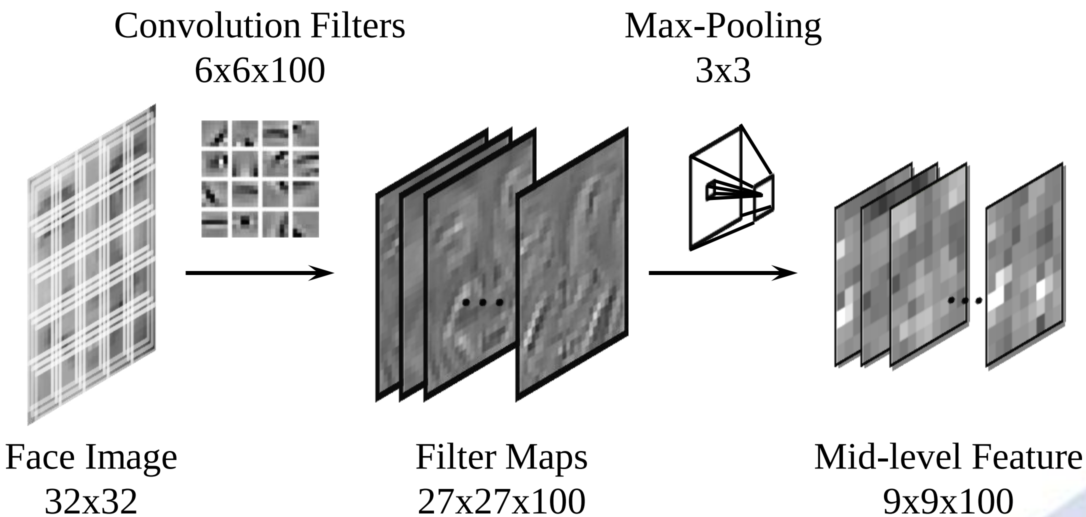
Our method

- An overview



Our method

- Feature designing
  - Image feature [4]



[4] M. Liu, S. Li, S. Shan, X. Chen. AU-aware Deep Networks for Facial Expression Recognition. FG, 2013.

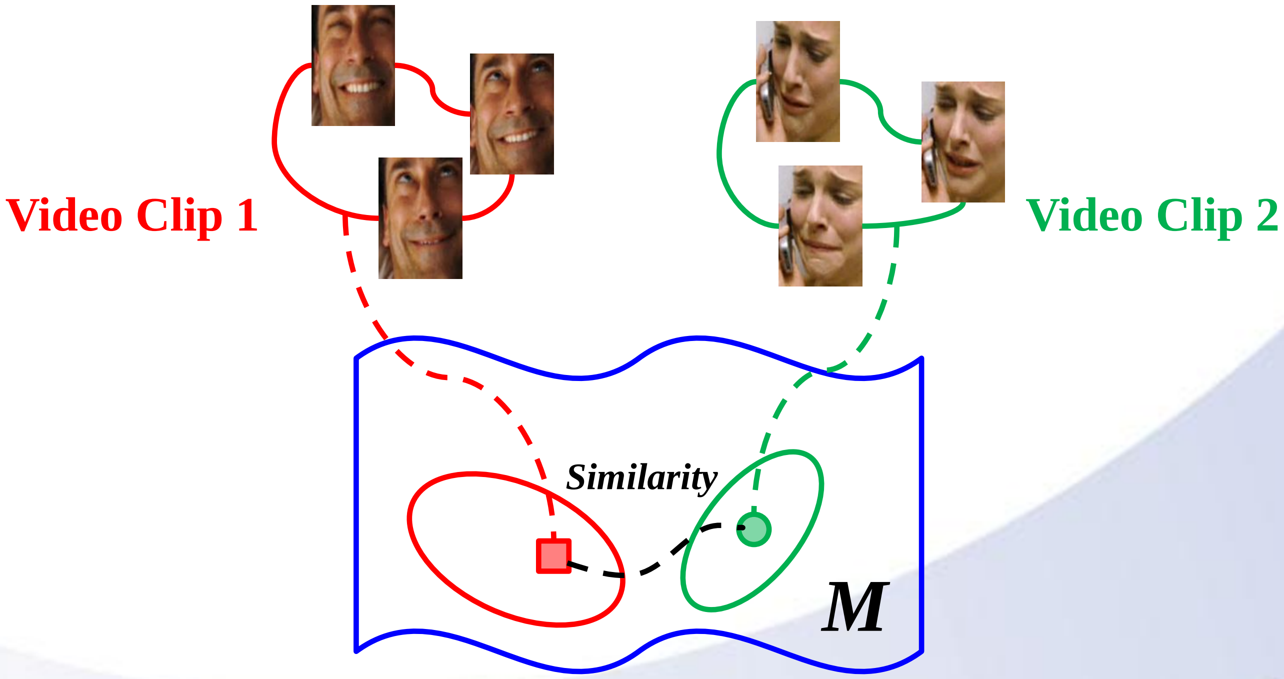
Our method

- Feature designing
  - Video feature
    - Each video clip is a set of images, denoted as $S_i \in R^{f \times n_i}$, where f is the dimension of image feature, and n_i is the number of frames.
    - The video S_i can be represented as a linear subspace P_i, s.t.
      $$S_i S_i^T = P_i \Lambda_i P_i^T$$
    - Thus all the video clips can be modeled as a collection of subspaces, which are also the points on Grassmannian manifold.

中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

Our method

- Feature designing
  - Video feature
    - An illustration of subspaces on Grassmannian manifold



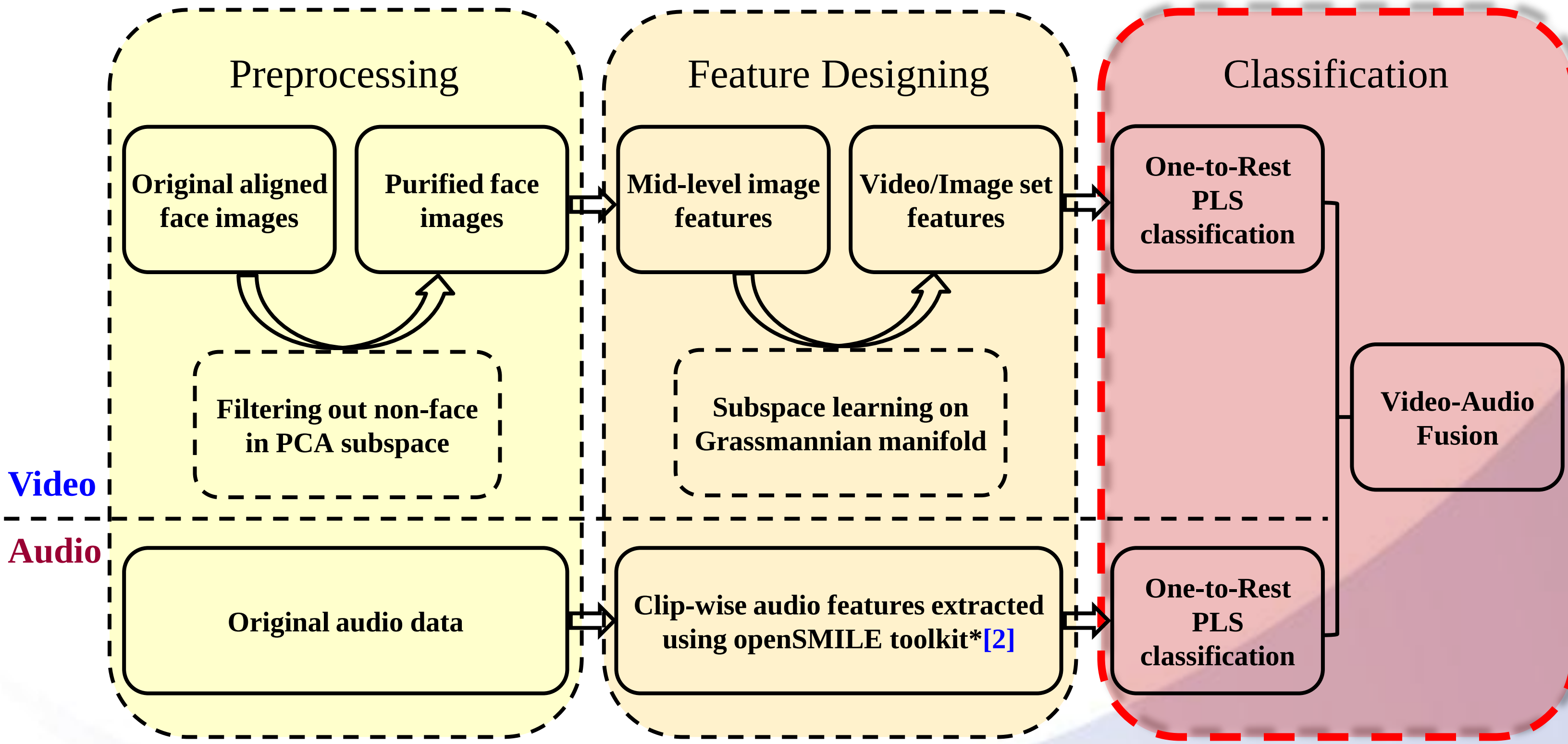
Our method

- Feature designing
  - Video feature
    - The similarity between two points $\boldsymbol{P_i}$ and $\boldsymbol{P_j}$ on manifold M can be measured by a linear combination of Grassmannian kernels.
      - Projection kernel[5]: $k_{ij}^{[proj]} = ||P_i^T P_j||_F^2$.
      - Canonical correlation kernel[6]: $k_{ij}^{[CC]} = max_{a_p \in span(P_i)} max_{b_q \in span(P_j)} a_p^T b_q$.
      - Linear combination: $k_{ij}^{[com]} = k_{ij}^{[proj]} + \alpha k_{ij}^{[CC]}$.
    - The kernels of each point (i.e., each video) to all training points serve as its final feature representation for classification.

[5] J. Hamm, D. Lee. Grassmann discriminant analysis: a unifying view on subspace-based learning. ICML, 2008.
[6] M. Harandi, C. Sanderson, S. Shirazi, B.C. Lovell. Graph embedding discriminant analysis on Grassmannian manifolds for improved image set matching. CVPR, 2011.

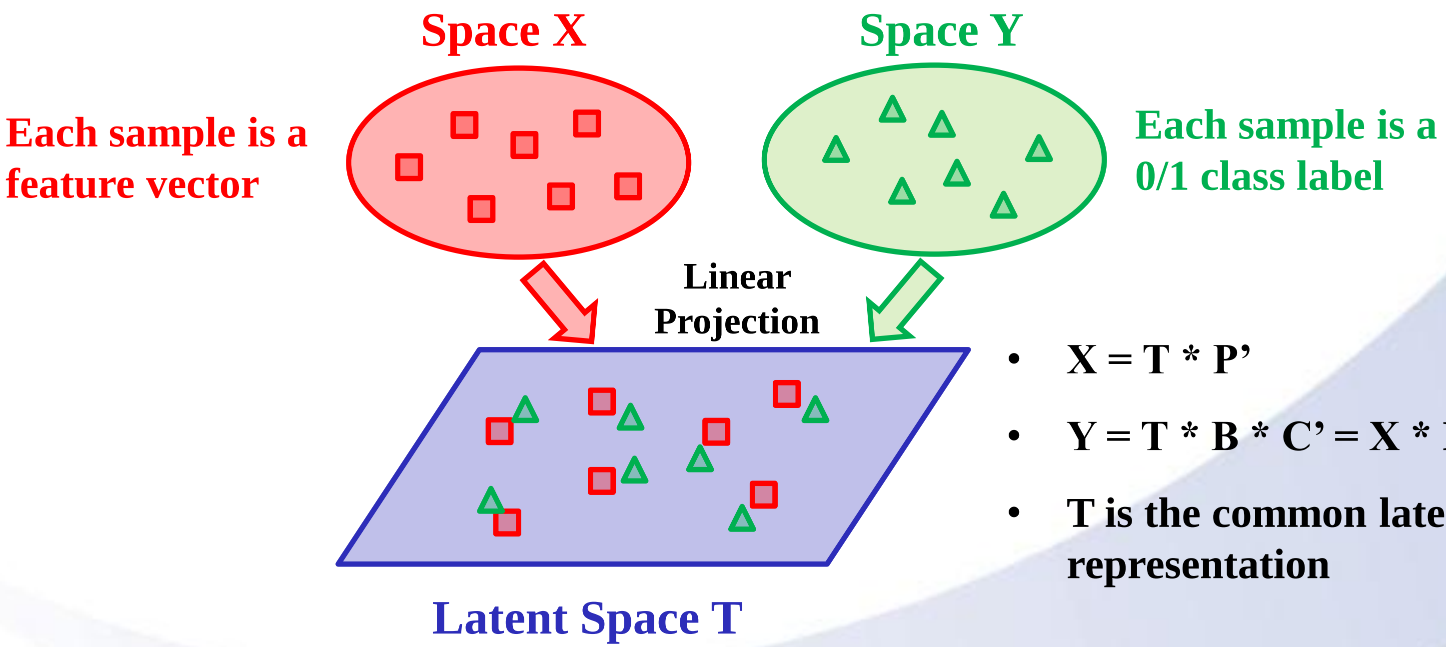
Our method

- An overview



Our method

- Classification
  - Partial Least Squares (PLS) for classification [1]
    - Maximize the covariance between observations and class labels

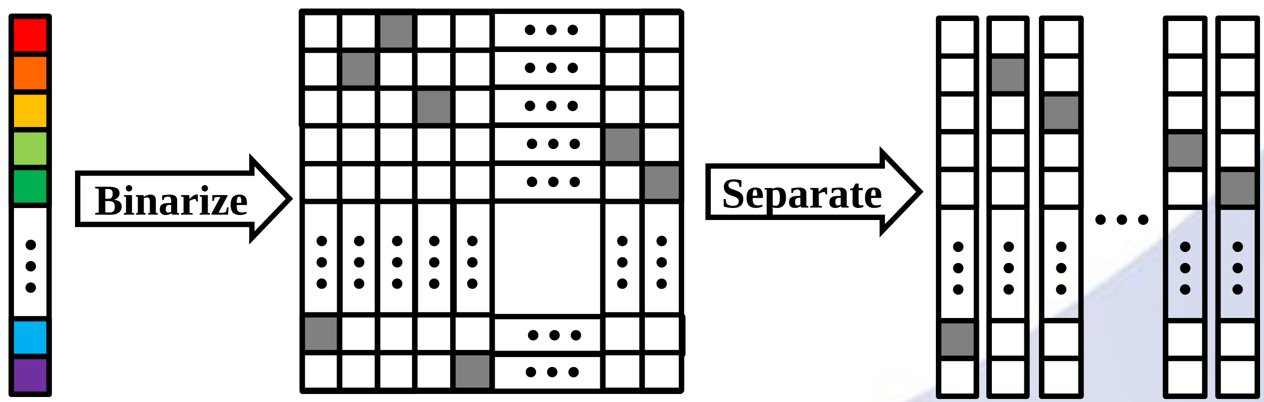


- $X = T * P'$
- $Y = T * B * C' = X * B\_pls$
- T is the common latent representation

[1] R. Wang, H. Guo, L. S. Davis, and Q. Dai. Covariance discriminative learning: A natural and efficient approach to image set classification. CVPR, 2012.

Our method

- Classification
  - One-to-Rest PLS
    - Suppose there are ***c*** categories and ***N*** training samples, we train ***c*** One-to-Rest PLS classifiers to predict each class simultaneously.
    - Effectively to handle the hard classes, e.g. “*Sad*” vs. “*Disgust*”



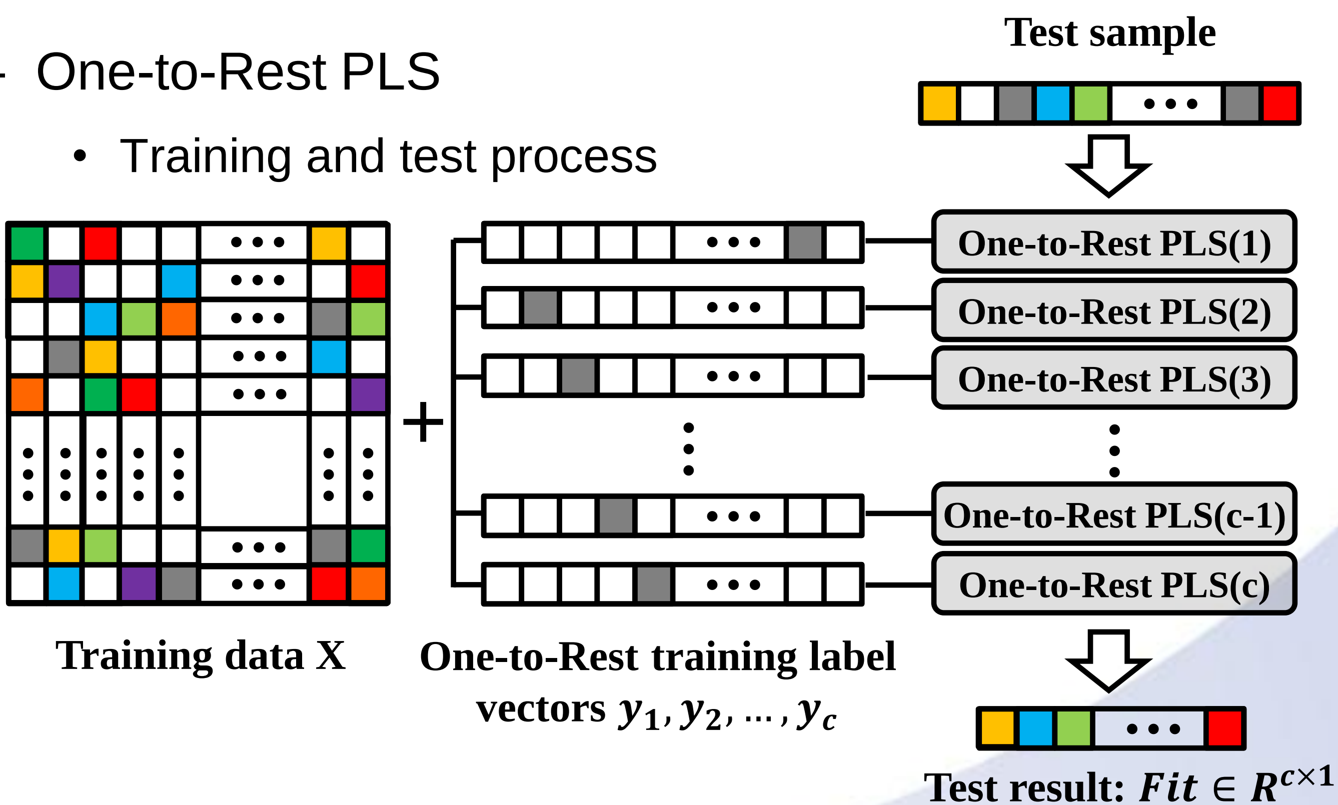
Original training label vector $\boldsymbol{Y} \in \boldsymbol{R}^{N\times 1}$

Binary training label matrix $\boldsymbol{Y} \in \boldsymbol{R}^{N\times c}$

One-to-Rest training label vectors, $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_c \in \boldsymbol{R}^{N\times 1}$

Our method

- Classification
  - One-to-Rest PLS
    - Training and test process



Our method

- Classification
  - Video-Audio fusion for final test output
    - For a given test video, using the *c* PLS classifiers for video and audio respectively, we obtain two prediction vectors $Fit^{video}, Fit^{audio} \in R^{c\times 1}$.
    - We conduct a linear fusion at decision level using weighted parameter λ

$$Fit^{fusion} = (1-\lambda)\, Fit^{video} + \lambda Fit^{audio}.$$

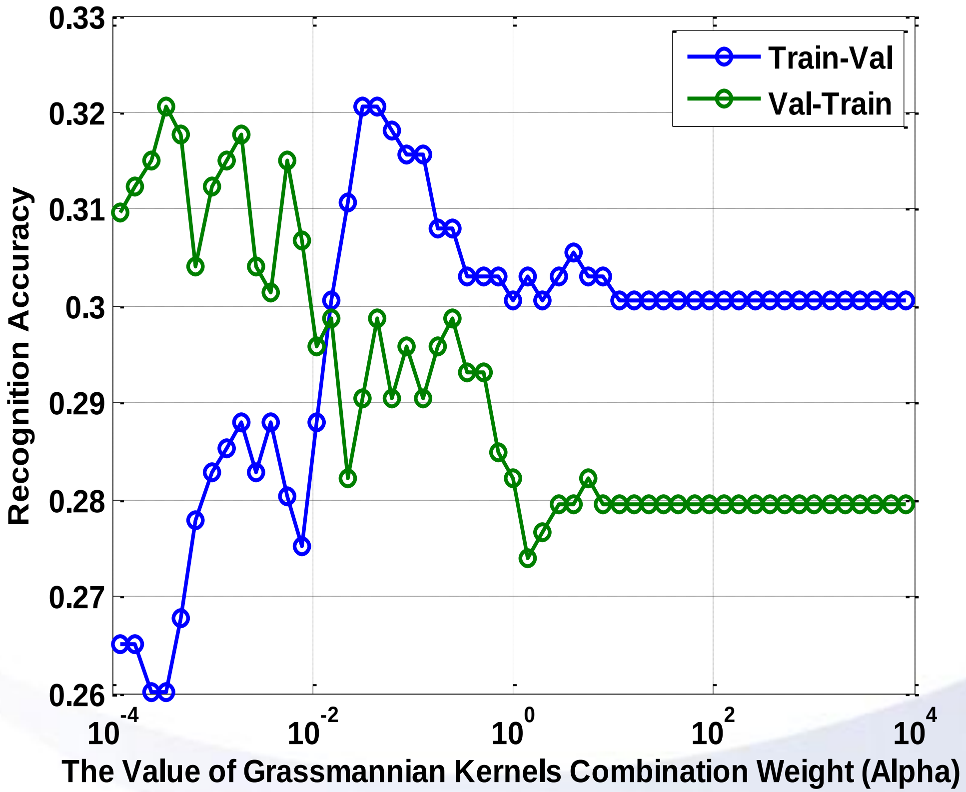
    - The category corresponding to the maximum value in $\boldsymbol{Fit^{fusion}}$ is determined to be the recognition result.

Outline

- Problem
- Related work
- Our Method
- Experiments
- Conclusion

Experiments

- Discussion of Parameters
  - The fusion weights of Grassmannian kernels



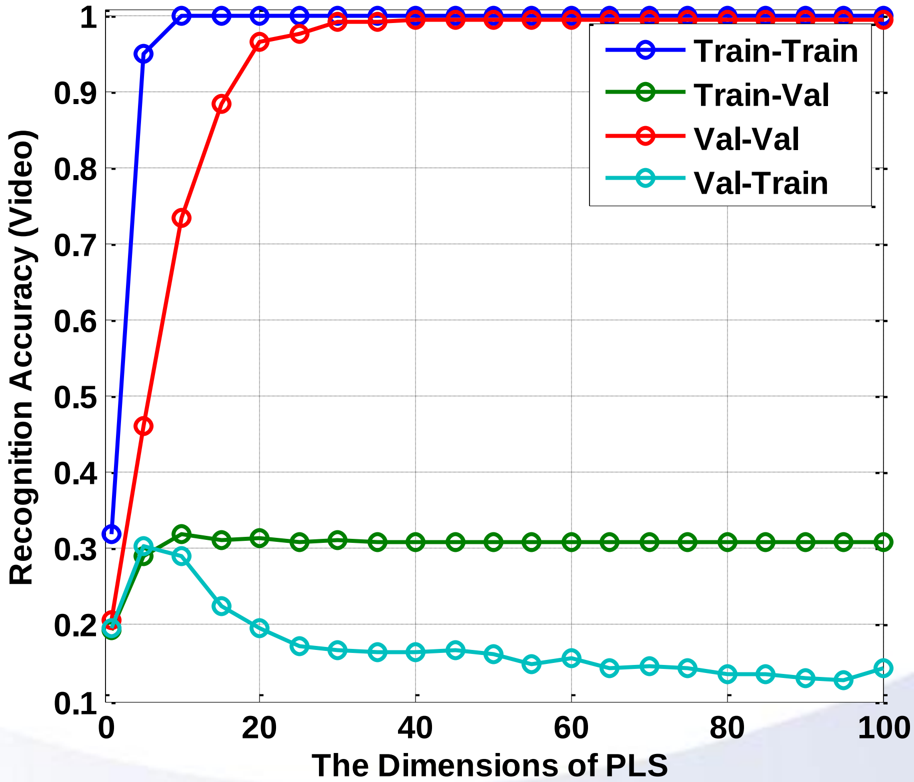
$$k_{ij}^{[com]} = k_{ij}^{[proj]} + \alpha k_{ij}^{[CC]}$$

Train-Val: $@\alpha = 2^{-6}, 2^{-5}$

Val-Train: $@\alpha = 2^{-10}$

Experiments

- Discussion of Parameters
  - The dimension of One-to-Rest PLS (video)

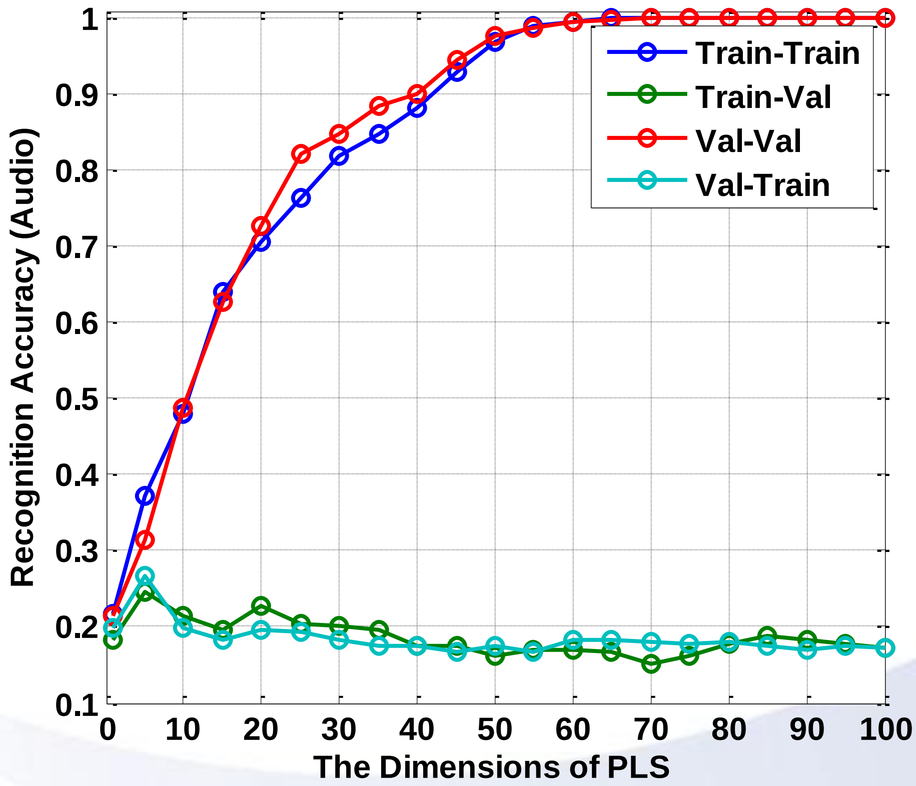


Train-Val: $@dim = 10$

Val-Train: $@dim = 5$

Experiments

- Discussion of Parameters
  - The dimension of One-to-Rest PLS (audio)

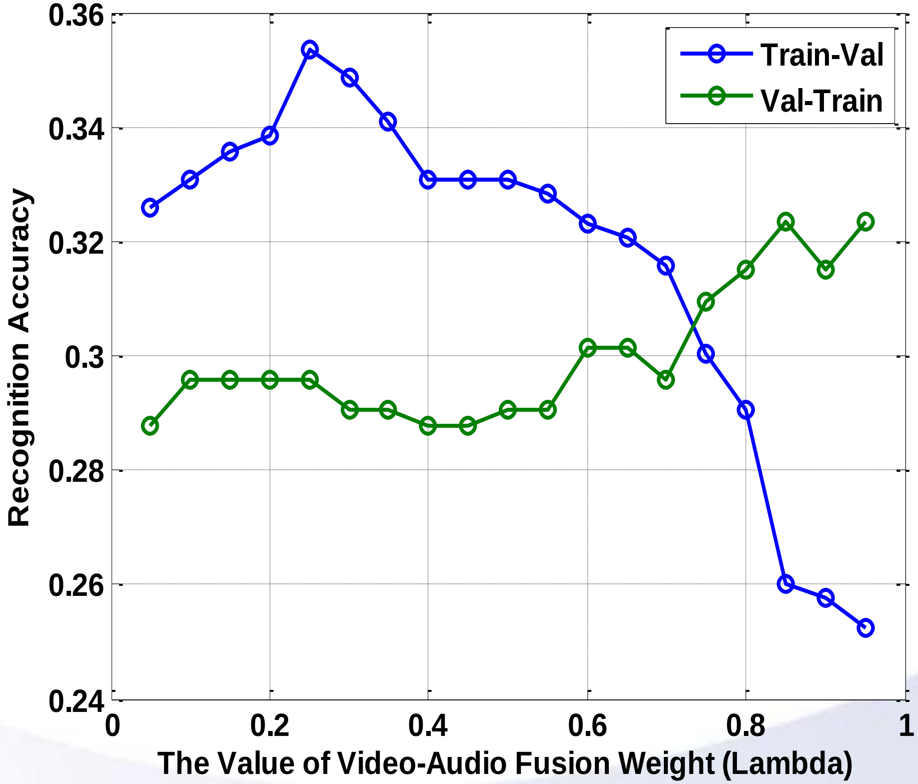


Train-Val: @$dim = 5$

Val-Train: @$dim = 5$

Experiments

- Discussion of Parameters
  - The fusion weights of video and audio modalities



$$Fit^{fusion} = (1 - \lambda) Fit^{video} + \lambda Fit^{audio}$$

Train-Val: @$\lambda = 0.25$

Val-Train: @$\lambda = 0.85$

Experiments

- Results comparison

Performance Comparison		Audio only	Video only		Audio + Video			
					Original data			Purified data
					Feature-level fusion		Decision-level fusion	Decision-level fusion
		One-to-Rest PLS	Grassmannian Discriminant Analysis [6]	Grassmannian Kernels + One-to-Rest PLS	Multi-class LR	One-to-Rest PLS	One-to-Rest PLS	One-to-Rest PLS
Ours	***Val***	24.49 %	30.81%	32.07%	22.48%	24.24%	34.34%	**35.86%**
	Test*	--	**24.04%**	--	--	**26.28%**	**33.01%**	**34.61%**
Baseline	***Val***	19.95%	27.27%		22.22%			
	Test	22.44%	22.75%		27.56%			

[6] M. Harandi, C. Sanderson, S. Shirazi, B.C. Lovell. Graph embedding discriminant analysis on Grassmannian manifolds for improved image set matching. CVPR, 2011.

Outline

- Problem
- Related work
- Our Method
- Experiments
- Conclusion

Conclusion

- Key points of the current method
  - PCA-based data purifying to filter out mis-alignment faces
  - Linear subspace modeling of video data variations
  - Multiple video features fusion by Grassmannian kernels combination
  - Multi-modality fusion at decision level of video and audio
- Issues to further address
  - Exploration of video temporal dynamics information
  - More sophisticated video modeling
  - More effective fusion at feature level
  - …

Thank you.

Question?