中国科学院
CHINESE ACADEMY OF SCIENCES

# Combining Multiple Kernel Methods on Riemannian Manifold for Emotion Recognition in the Wild

**Mengyi Liu, Ruiping Wang, Shaoxin Li, Shiguang Shan, Zhiwu Huang, and Xilin Chen**

Visual Information Processing and Learning (VIPL) Group
@ Institute of Computing Technology, Chinese Academy of Sciences



中国科学院计算技术研究所
Institute of Computing Technology,Chinese Academy of Sciences

# Outline

- Introduction
- Data and protocols
- Proposed method
- Experiments
- Conclusion

中国科学院 Chinese Academy of Sciences

# Introduction (1/1)

- Task
  - Classify audio-video clips into seven categories
    - Neutral, anger, disgust, fear, happy, sad, surprise
- Challenge
  - Close-to-real-world conditions
    - Large variations: head pose, illumination, occlusion, etc.

# Outline

- Introduction
- Data and protocols
- Proposed method
- Experiments
- Conclusion

中国科学院 Chinese Academy of Sciences

# Data and protocols (1/2)

- Challenge data
  - Acted Facial Expression in Wild (AFEW 4.0)
    - Over 1,000 audio-video clips collected from movies showing close-to-real-world conditions

**Tab.1 Attributes of EmotiW 2014 challenge data [1]**

| Attribute | Description |
|---|---|
| Length of sequences | 300-5400ms |
| Number of annotators | 3 |
| Emotion categories | Anger, disgust, fear, happiness, neutral, sadness, and surprise |
| Audio/Video format | Audio: WAV; Video: AVI |
| # of samples | 1368 |
| # of subjects | 428 |
| # of movies | 111 |

[1] Abhinav Dhall, Roland Goecke, Jyoti Joshi, Karan Sikka and Tom Gedeon, Emotion Recognition In The Wild Challenge 2014: Baseline, Data and Protocol, ACM ICMI 2014.

# Data and protocols (2/2)

- Evaluation protocols
  - Dataset division: training, validation, and testing
  - The test labels were unknown.
  - Either audio/video modality or both can be used.

**Tab.2 Attributes of subjects of training, validation, and testing sets.**

| Set | # of subjects | Min. Age | Max. Age | Avg. Age | # of Males | # of Females |
|---|---|---|---|---|---|---|
| **Train** | 177 | 5 | 76 | 34 | 102 | 75 |
| **Val** | 136 | 10 | 70 | 35 | 78 | 58 |
| **Test** | 115 | 5 | 88 | 34 | 64 | 51 |

**Tab.3 The numbers of samples for each emotion category in the training, validation and testing sets.**

| | Anger | Digust | Fear | Happiness | Neutral | Sadness | Surprise |
|---|---|---|---|---|---|---|---|
| **Train** | 92 | 66 | 66 | 105 | 102 | 82 | 54 |
| **Val** | 59 | 39 | 44 | 63 | 61 | 59 | 46 |
| **Test** | 58 | 26 | 46 | 81 | 117 | 53 | 26 |

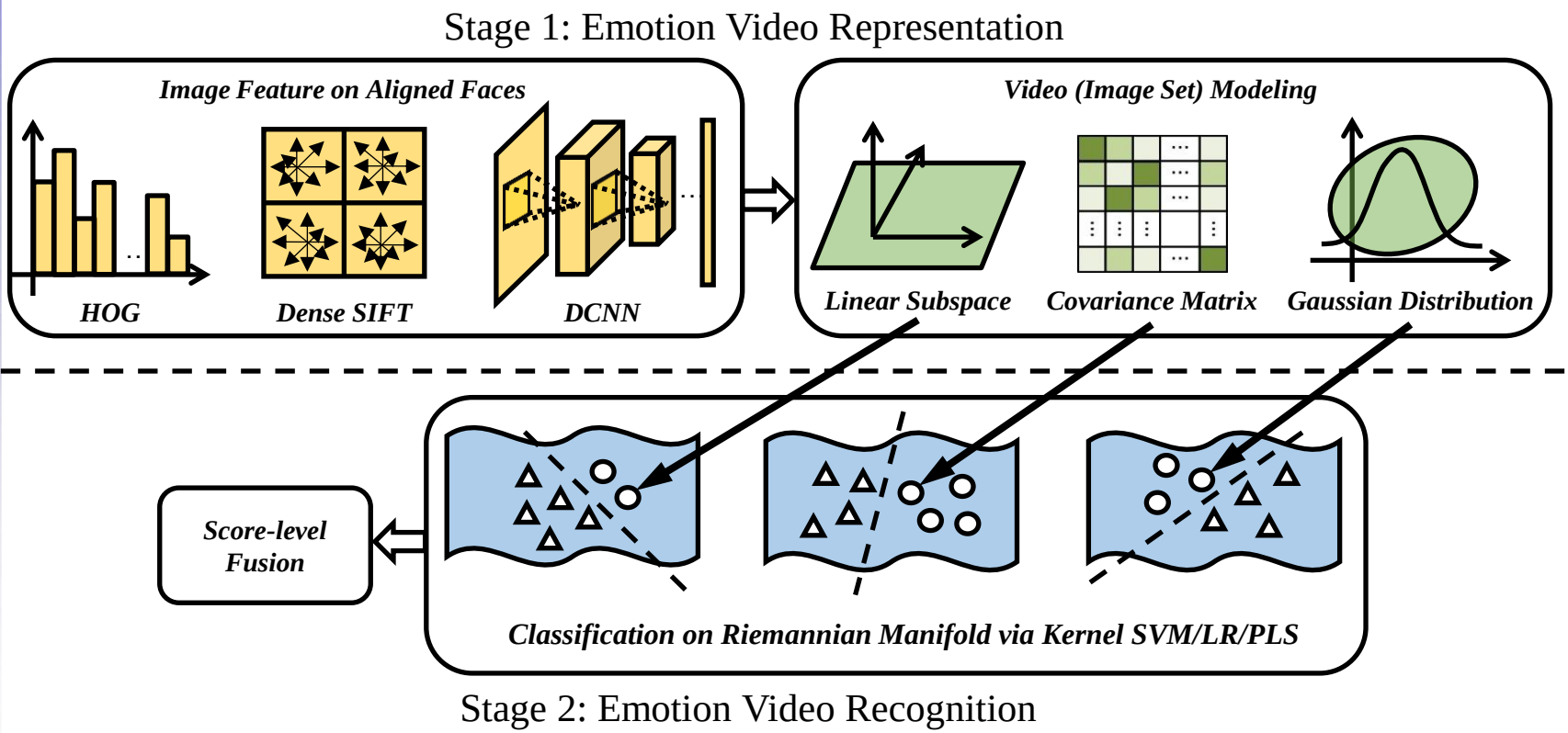
# Outline

- Introduction
- Data and protocols
- Proposed method
- Experiments
- Conclusion

中国科学院 Chinese Academy of Sciences

# Proposed method (1/9)

- Framework



# Proposed method (2/9)

- Image features
  - Aligned face images: 64x64;  Features: HOG, dense SIFT, DCNN.
  - HOG
    - Block size: 16x16; stride: 8; # of blocks: 7x7=49
    - # of cells per block: 2x2; # of bins: 9; # of total dims: **2x2x9x49=1764**
  - Dense SIFT
    - Block size: 16x16; stride: 8; # of points: 7x7=49
    - # of dims per point: 4x4x8=128; # of total dims: **128x49=6272**
  - DCNN
    - CaffeNet trained on CFW database (over 150,000 face images from 1520 subjects, identities are served as supervised label in the deep networks)
    - Deep Architecture 3@237x237 > 96@57x57 > 96@28x28 > 256@28x28 > 384@14x14 > 256@14x14 > **256@7x7** > 4096 > 1520
    - The **256x7x7=12544** nodes values output from the last convolutional layer are used for final image features

# Proposed method (3/9)

- Video (image set) modeling
  - Given frame feature set

$$F = \{f_1, f_2, \dots, f_n\} \ (f_i \in R^d)$$

    - Linear subspace

$$P \in R^{d \times r} \ \text{s.t.} \sum_{i=1}^{n} f_i f_i^T = P \Lambda P^T \ , P = [p_1, p_2, \dots, p_r]$$

    - Covariance matrix

$$C = \frac{1}{n-1} \sum_{i=1}^{n} (f_i - \bar{f})(f_i - \bar{f})^T$$

    - Gaussian distribution
      - Suppose $f_1, f_2, \dots, f_n$ follow a $k$-dimensional Gaussian $\mathcal{N}(\mu, \Sigma)$

$$\mu = E(f_i) = \frac{1}{n} \sum_{i=1}^{n} f_i$$

$$\Sigma = E[(f_i - \mu)(f_i - \mu)^T] = \frac{1}{n-1} \sum_{i=1}^{n} (f_i - \bar{f})(f_i - \bar{f})^T$$

# Proposed method (4/9)

- Riemannian kernels
  - Kernels for linear subspace
    - The similarity of two subspaces can be measured via mapping Grassmann manifold to Euclidean space by **Projection kernel** originated from **principle angles**
    - Then the form of **polynomial kernel** can be calculated as

$$\mathcal{K}_{i,j}^{Proj.-Poly.} = \left(\gamma \cdot \left|\left|P_i^T P_j\right|\right|_F^2\right)^{\alpha}$$

    - Define $\Phi_{Proj.} = P_i P_j^T$. Then the form of **RBF kernel** is

$$\mathcal{K}_{i,j}^{proj.-RBF} = \exp(-\gamma \left|\left|\Phi_{Proj.}(P_i) - \Phi_{Proj.}(P_j)\right|\right|_F^2)$$

# Proposed method (5/9)

- Riemannian kernels
  - Kernels for linear subspace

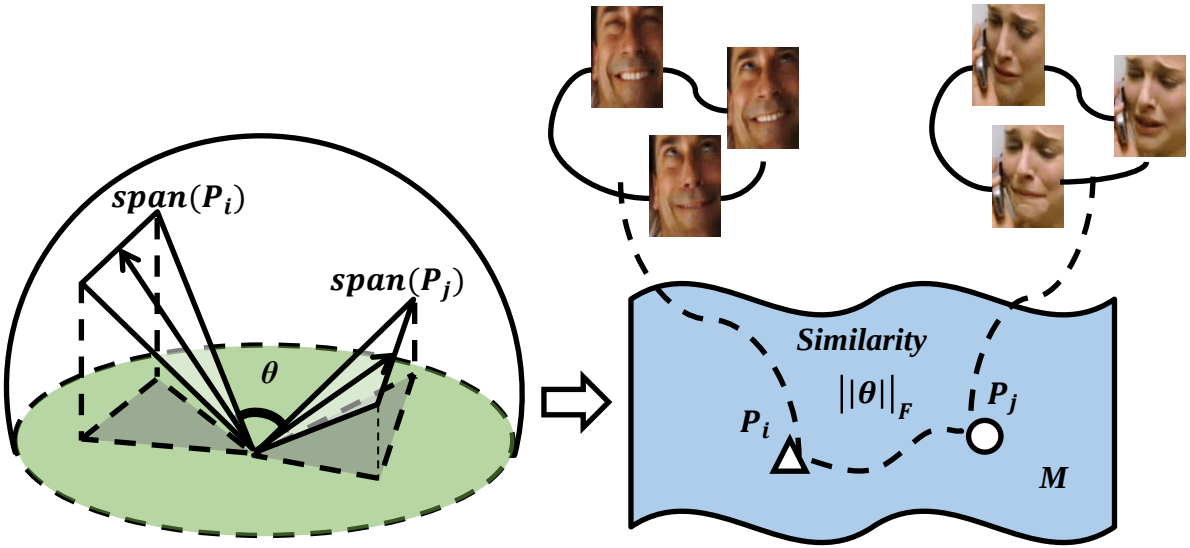



**Fig.2 An illustration of principal angles of linear subspaces and their projection metric distances on Grassmann manifold *M*.**

# Proposed method (6/9)

- Riemannian kernels
  - Kernels for covariance matrix
    - The SPD matrix on Riemannian manifold can be mapped to vector space via ordinary matrix logarithm operator originated from the **Log-Euclidean Distance (LED)**
    - Then the **polynomial kernel** can be calculated as

$$\mathcal{K}_{i,j}^{LED-Poly.} = \left(\gamma \cdot trace\left[\log(C_i) \cdot \log(C_j)\right]\right)^{\alpha}$$

    - The corresponding mapping is $\Phi_{LED} = \log(C_i)$, then the form of **RBF kernel** can be generated using $\Phi_{LED}$ by

$$\mathcal{K}_{i,j}^{LED-RBF} = \exp\left(-\gamma \left|\left|\Phi_{LED}(C_i) - \Phi_{LED}(C_j)\right|\right|_F^2\right)$$

# Proposed method (7/9)

- Riemannian kernels
  - Kernels for covariance matrix

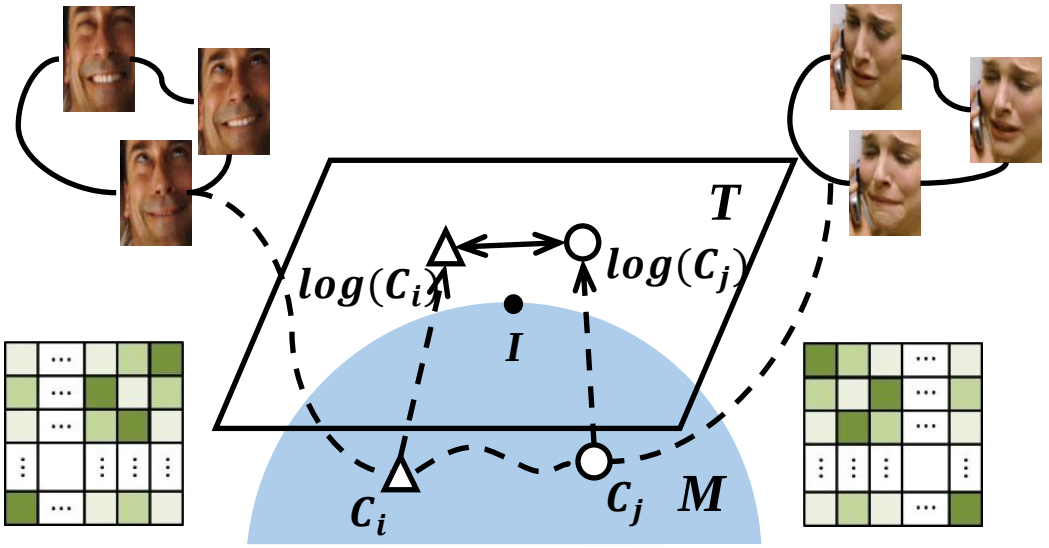



**Fig.3 An illustration of mapping covariance matrices from the SPD Riemannian manifold *M* to the tangent space *T* (which is a vector space) at the point of identity matrix *I* on *M*.**

# Proposed method (8/9)

- Riemannian kernels
  - Kernels for Gaussian distribution
    - The space of $d$-dimensional multivariate Gaussians $\mathcal{N}(\mu, \Sigma)$ lie on Riemannian manifold and can be embedded into the space of **$(d+1)$-dimensional SPD** matrices (denoted as $Sym^{+}_{d+1}$) as follows

      $$\mathcal{N}(\mu, \Sigma) \sim G = |\Sigma|^{-\frac{1}{d+1}} \begin{bmatrix} \Sigma + \mu\mu^T & \mu \\ \mu & 1 \end{bmatrix}$$

    - Similar to covariance matrix, the **polynomial kernel** can be calculated

      $$\mathcal{K}_{i,j}^{LED-Poly.} = \left(\gamma \cdot trace\left[\log(G_i) \cdot \log(G_j)\right]\right)^{\alpha}$$

    - And the form of **RBF kernel** can be generated using $\Phi_{LED}$ by

      $$\mathcal{K}_{i,j}^{LED-RBF} = \exp\left(-\gamma \left|\left|\Phi_{LED}(G_i) - \Phi_{LED}(G_j)\right|\right|_F^2\right)$$

# Proposed method (9/9)

- Different classifiers
  - Kernel SVM
  - Logistic Regression
  - One-vs-rest Partial Least Squares
- Fusion scheme
  - Score-level fusion of different kernel methods
  - Linear weighted score fusion of audio and video modality

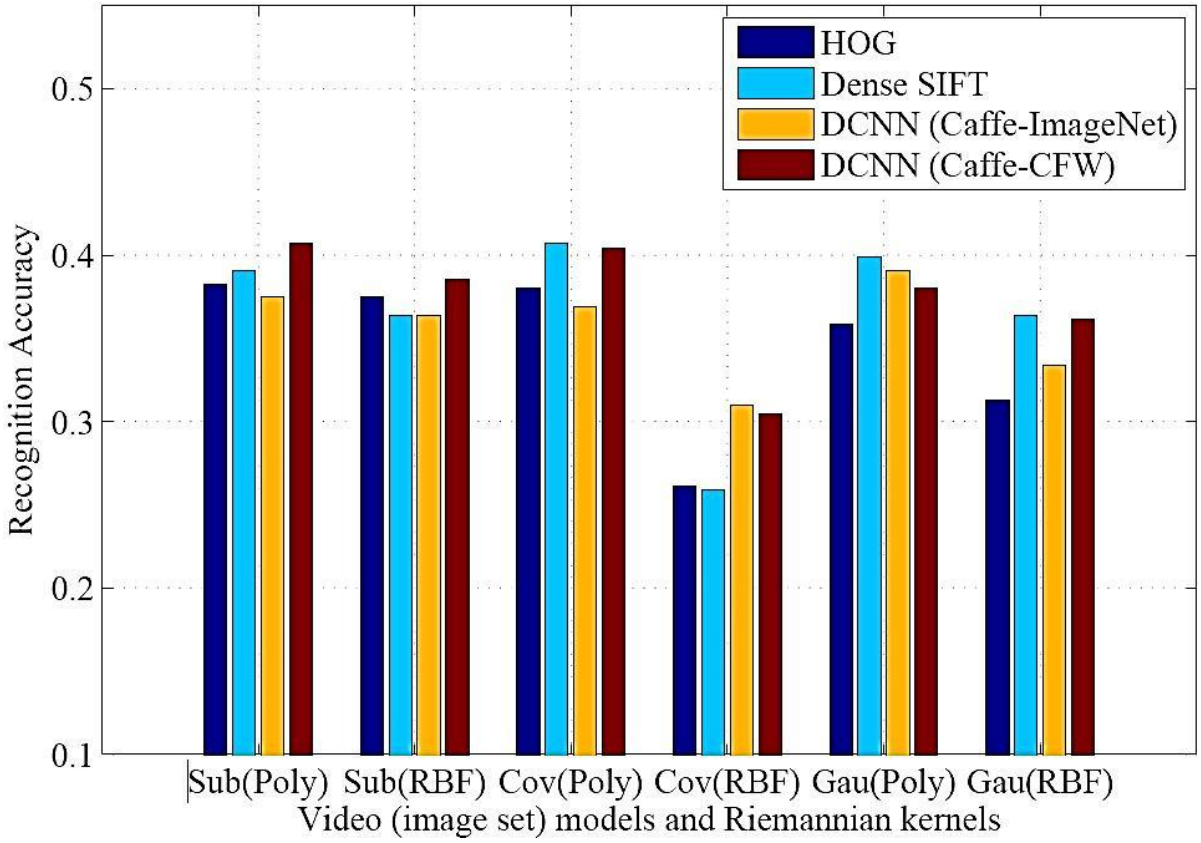
# Outline

- Introduction
- Data and protocols
- Proposed method
- Experiments
- Conclusion

中国科学院 Chinese Academy of Sciences

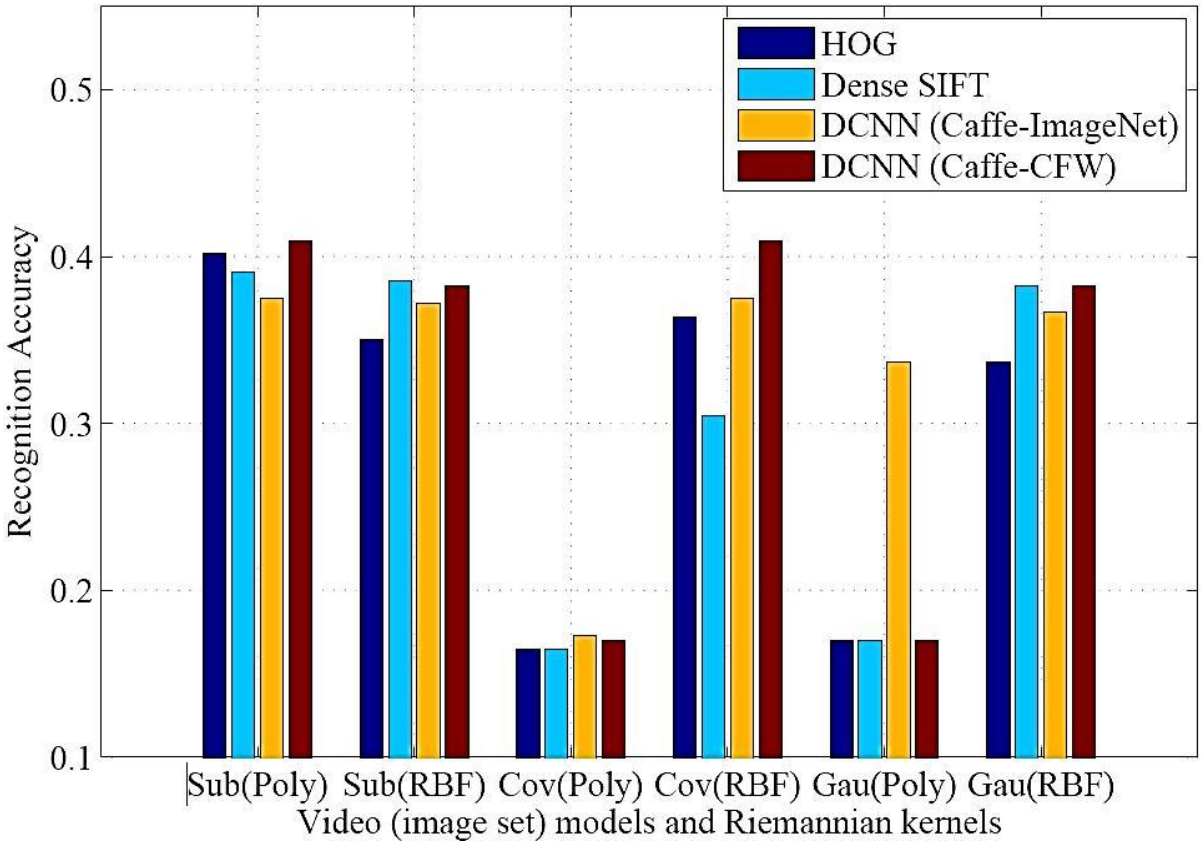
# Experiments (1/7)

- Results comparison – Kernel SVM



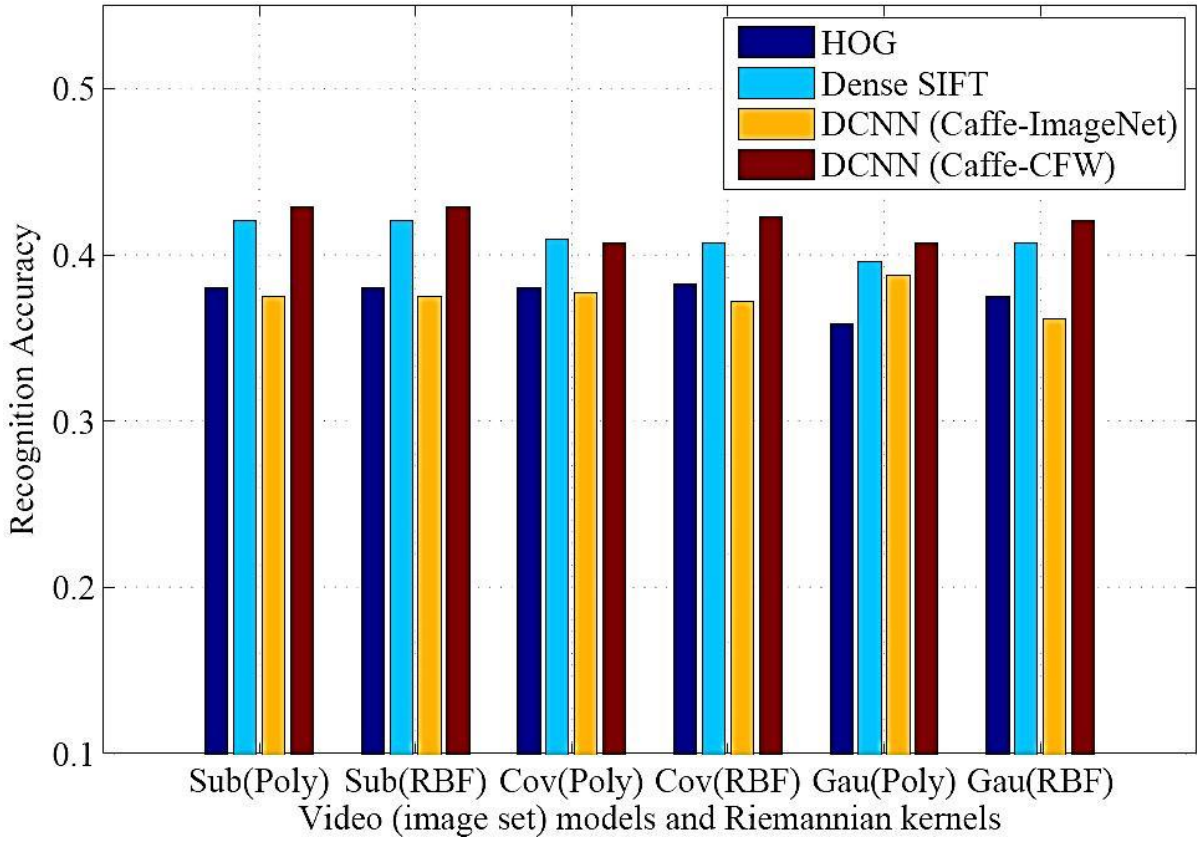
# Experiments (2/7)

- Results comparison – Logistic Regression



# Experiments (3/7)

- Results comparison – One-vs-rest Partial Least Squares



# Experiments (4/7)

- Results comparison – **HOG**

| | Linear Subspace | | Covariance Matrix | | Gaussian Distribution | |
|---|---|---|---|---|---|---|
| | Proj.-Poly. Kernel | Proj.-RBF Kernel | LED-Poly. Kenrel | LED-RBF Kernel | LED-Poly. Kenrel | LED-RBF Kernel |
| Kernel SVM | 38.27 | 37.47 | 38.01 | 26.15 | 35.85 | 31.27 |
| Logistic Regression | 40.16 | 35.04 | 16.44 | 36.39 | 16.98 | 33.69 |
| Partial Least Squares | 38.01 | 38.01 | 38.01 | 38.27 | 35.85 | 37.47 |

- Results comparison – **Dense SIFT**

| | Linear Subspace | | Covariance Matrix | | Gaussian Distribution | |
|---|---|---|---|---|---|---|
| | Proj.-Poly. Kernel | Proj.-RBF Kernel | LED-Poly. Kenrel | LED-RBF Kernel | LED-Poly. Kenrel | LED-RBF Kernel |
| Kernel SVM | 39.08 | 36.39 | 40.70 | 25.88 | 39.89 | 36.39 |
| Logistic Regression | 39.08 | 38.54 | 16.44 | 30.46 | 16.98 | 38.27 |
| Partial Least Squares | 42.05 | 42.05 | 40.97 | 40.70 | 39.62 | 40.70 |

# Experiments (5/7)

- Results comparison – **Caffe trained on ImageNet**

| | Linear Subspace | | Covariance Matrix | | Gaussian Distribution | |
|---|---|---|---|---|---|---|
| | Proj.-Poly. Kernel | Proj.-RBF Kernel | LED-Poly. Kenrel | LED-RBF Kernel | LED-Poly. Kenrel | LED-RBF Kernel |
| Kernel SVM | 37.74 | 36.39 | 36.93 | 31.00 | 39.08 | 33.42 |
| Logistic Regression | 37.47 | 37.20 | 17.25 | 37.47 | 33.69 | 36.66 |
| Partial Least Squares | 37.47 | 37.47 | 37.74 | 37.20 | 38.81 | 36.12 |

- Results comparison – **Caffe trained on CFW**

| | Linear Subspace | | Covariance Matrix | | Gaussian Distribution | |
|---|---|---|---|---|---|---|
| | Proj.-Poly. Kernel | Proj.-RBF Kernel | LED-Poly. Kenrel | LED-RBF Kernel | LED-Poly. Kenrel | LED-RBF Kernel |
| Kernel SVM | 40.70 | 38.54 | 40.43 | 30.46 | 38.01 | 36.12 |
| Logistic Regression | 40.97 | 38.27 | 16.98 | 40.97 | 16.98 | 38.27 |
| Partial Least Squares | 42.86 | 42.86 | 40.70 | 42.32 | 40.70 | 42.05 |

# Experiments (6/7)

| Methods | | Accuracy | |
|---|---|---|---|
| | | Val | Test |
| | Baseline (provided by EmotiW organizers) | 34.4% | 33.7% |
| | Audio (OpenSMILE Toolkit) | 30.73% | -- |
| Video | HOG | 38.01% | -- |
| | Dense SIFT | **43.94%** | -- |
| | DCNN (Caffe-ImageNet) | 39.35% | -- |
| | DCNN (Caffe-CFW) | **43.40%** | -- |
| | HOG + Dense SIFT | 44.47% | -- |
| | HOG + Dense SIFT + DCNN (Caffe-ImageNet) | 44.74% | -- |
| | HOG + Dense SIFT + DCNN (Caffe-CFW) | **45.28%** | -- |
| | Audio + Video ( HOG+Dense SIFT ) | 46.36% | 46.68% |
| | Audio+Video ( HOG + Dense SIFT + DCNN (Caffe-ImageNet) ) | 46.90% | 47.91% |
| | **Audio+Video ( HOG + Dense SIFT + DCNN (Caffe-CFW) )** | **48.52%** | **50.37%** |

# Experiments (6/7)

**Validation Set**

| | Angry | Disgust | Fear | Happy | Neutral | Sad | Surprise |
|---|---|---|---|---|---|---|---|
| Angry | 84.75% | 3.39% | 0.00% | 1.69% | 5.08% | 5.08% | 0.00% |
| Disgust | 10.26% | 17.95% | 2.56% | 28.21% | 33.33% | 5.13% | 2.56% |
| Fear | 27.27% | 6.82% | 27.27% | 13.64% | 11.36% | 9.09% | 4.55% |
| Happy | 4.76% | 0.00% | 0.00% | 82.54% | 9.52% | 3.17% | 0.00% |
| Neutral | 13.11% | 0.00% | 1.64% | 8.20% | 70.49% | 6.56% | 0.00% |
| Sad | 13.56% | 3.39% | 6.78% | 23.73% | 28.81% | 22.03% | 1.69% |
| Surprise | 17.39% | 4.35% | 28.26% | 8.70% | 32.61% | 2.17% | 6.52% |

**Test Set**

| | Angry | Disgust | Fear | Happy | Neutral | Sad | Surprise |
|---|---|---|---|---|---|---|---|
| Angry | 81.03% | 3.45% | 0.00% | 5.17% | 10.34% | 0.00% | 0.00% |
| Disgust | 11.54% | 3.85% | 3.85% | 34.62% | 23.08% | 15.38% | 7.69% |
| Fear | 26.09% | 0.00% | 23.91% | 10.87% | 19.57% | 15.22% | 4.35% |
| Happy | 8.64% | 0.00% | 1.23% | 64.20% | 11.11% | 14.81% | 0.00% |
| Neutral | 7.69% | 1.71% | 5.13% | 9.40% | 63.25% | 11.97% | 0.85% |
| Sad | 11.32% | 0.00% | 3.77% | 24.53% | 24.53% | 33.96% | 1.89% |
| Surprise | 11.54% | 0.00% | 19.23% | 7.69% | 34.62% | 19.23% | 7.69% |

# Outline

- Introduction
- Data and protocols
- Proposed method
- Experiments
- Conclusion

中国科学院 Chinese Academy of Sciences

# Conclusion (1/1)

- Conclusion
  - Video representation
    - 4 types of image features
    - 3 types of image set models
  - Video classification
    - multiple (6 types) kernel methods fusion on Riemannian manifold
    - 2 modality, i.e. audio and video, score-level fusion
- Future work
  - Few difficult categories
  - Effective fusion strategy

中国科学院

Chinese Academy of Sciences

# Thanks!

**Source code available at:**

**http://vipl.ict.ac.cn/members/myliu**