Multi-View Super Vector for Action Recognition

Zhuowei Cai[1], Limin Wang[1,2], Xiaojiang Peng[1], Yu Qiao[1,2]

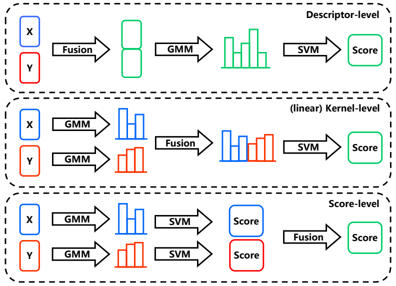
[1]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

[2]The Chinese University of Hong Kong

Motivation

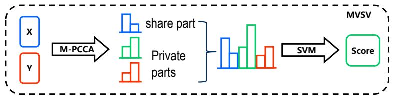
- Action videos have different types of local descriptors (HoG. HoF. MBHx/y), which capture different aspects of object feature.
- Fusing different types of descriptors can boost recognition performance.
- Current fusion pipelines presume different properties between local descriptors.
  - Direct concatenation presumes strong correlation between fusion descriptors.
  - Kernel average and score fusion prefer mutual independence between fusion descriptors.

Typical fusion pipelines:



Our Approach

- We develop the **M**ixture of **P**robabilistic **C**anonical **C**orrelation **A**nalyzers to model a pair of feature descriptors.
  - M-PCCA factorize descriptors from two sources into a share part between them and two private parts specific to each of them.
- These parts can be used to construct the super vector representation for subsequent classification.



Mixture model of Probabilistic CCA

- Non-linear Distribution of Descriptors
  - Descriptors are non-linearly distributed and their correlation relation is complex.
  - Observed descriptor pair share common information.
- Mixture model of Probabilistic CCA
  - Tackle non-linearity with mixture model
  - Encode shared information by specifying latent variable

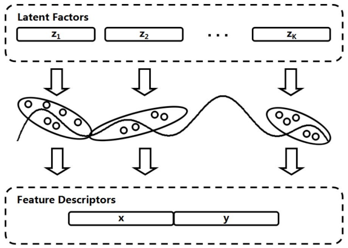


Figure 2. A graphical interpretation of M-PCCA. Each ellipse represents a PCCA submodel of M-PCCA. Observed descriptor pair is jointly generated by the mixture of K submodels.

- Mathematical Formulation
  - Let v = (x, y),

$$p(v) = \sum_k w_k p(v|k)$$

  - x, y satisfy

$$p(x|k, z_k) = \mathcal{N}(W_x^k z_k + \mu_x^k, \Psi_x^k)$$
$$p(y|k, z_k) = \mathcal{N}(W_y^k z_k + \mu_y^k, \Psi_y^k)$$

  - p(v | k) is Gaussian with

$$\mu_k = \begin{pmatrix} \mu_x^k \\ \mu_y^k \end{pmatrix},$$
$$\Sigma_k = \begin{bmatrix} W_x^k {W_x^k}^\top + \Psi_x^k, & W_x^k {W_y^k}^\top \\ W_y^k {W_x^k}^\top, & W_y^k {W_y^k}^\top + \Psi_y^k \end{bmatrix}.$$

Multi-View Super Vector

- In each local Probabilistic CCA, shared information Z and private information λ jointly generate the observed descriptors

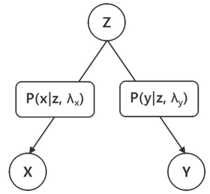


Figure 3. An interpretation of the MVSV representation. λ denotes parameters $\{\mu, \Psi, W\}$. Given z, descriptors x and y are mutually independent. z can thus be utilized as the shared information between them, with parameters λ_x and λ_y as their private information.

- Shared information Z is extracted via concatenating latent information z_k for each local PCCA

$$z_k = \left[\, {W_x^k}^\top, {W_y^k}^\top \,\right] \Sigma_k^{-1} \sum_i \gamma_{i,k}(v_i - \mu_k)$$

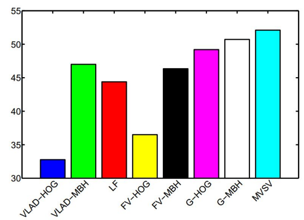
- Private information G_x is extracted by taking the gradients of the expected log likelihood with respect to λ_x (and λ_y)

$$\mathcal{G}_x = \left\{ \frac{\partial E(\mathcal{L})}{\partial \mu_x^k}, \frac{\partial E(\mathcal{L})}{\partial \sigma_x^k} \right\}_{k=1,\cdots,K}$$

- Z, G_x and G_y form the final MVSV.

Performance of Different Components

- Performance of differnet parts from MVSV on the HMDB51 dataset.



Experimental Results

- Comparison of performance on HMDB51.

HMDB51	FV		VLAD		MVSV
Fusion	d-level	k-level	d-level	k-level	k-level
HOG+MBH	50.9%	50.4%	47.0%	48.5%	**52.1%**
HOG+HOF	47.0%	48.3%	44.4%	47.7%	**48.9%**
MBH(x+y)	49.2%	49.1%	45.2%	47.0%	**51.1%**
Combine	52.4%	53.2%	51.5%	52.6%	**55.9%**

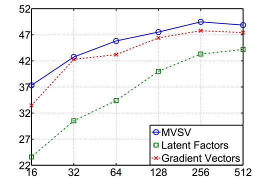
Table 1. Performance of MVSV on HMDB51 database. d-level refers to direct concatenation of descriptors. k-level refers to kernel average.

- Comparison of performance on UCF101.

UCF101	FV		VLAD		MVSV
Fusion	d-level	k-level	d-level	k-level	k-level
HOG+HOF	76.1%	77.7%	75.7%	77.5%	**78.9%**
MBH(x+y)	78.9%	78.7%	75.6%	76.3%	**80.9%**
Combine	81.1%	81.9%	80.6%	81.0%	**83.5%**

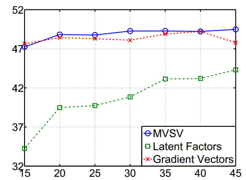
Table 2. Performance of MVSV on UCF101 database.

- Influence of Parameters



#components

latent dimension



- Comparison to the state-of-the-art

HMDB51		UCF101	
STIP+BoVW [16]	23.0%	STIP+BoVW [27]	43.9%
Motionlets [34]	42.1%	DT+VLAD	79.9%
DT+BoVW [32]	46.6%	DT+FV	81.4%
FWOT [37]	48.9%		
w-traj+VLAD [12]	52.1%		
DT+FV+SPM [21]	54.8%		
MVSV	**55.9%**	MVSV	**83.5%**

Table 3. Comparison of MVSV to the state-of-the-art methods.