Nonlinear Cross-View Sample Enrichment for Action Recognition

Ling Wang Hichem Sahbi

Department of Signal and Image Processing
Télécom ParisTech
Paris, France
{ling.wang,hichem.sahbi}@telecom-paristech.fr

September 12, 2014

Table of contents

Action recognition problem

Categorize the actions of human beings in a given video.

- Broad applications
  - Robotic vision, autonomous driving, surveillance, video indexing and retrieval, ...
- Great challenges
  - Cluttered background, outdoor environment, viewpoint change, insufficient training data, ...

Related work

- Build view invariant representations

  e.g. Huang et al. (ECCV 2012 Workshop), Junejo et al. (ECCV 2008), Le et al. (CVPR 2011), Zheng et al. (BMVC 2012)

- Combine models from different views

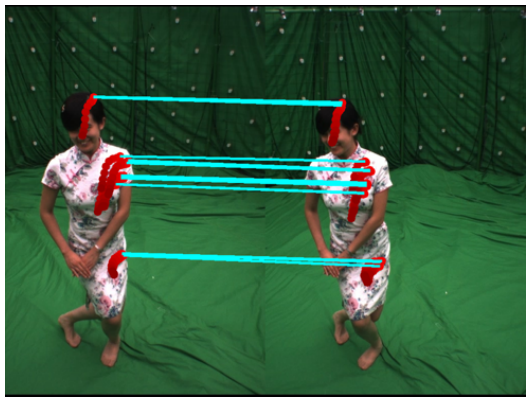
  e.g. Farhadi and Tabrizi (ECCV 2008), Weinland et al. (ECCV 2010)

- Transfer knowledge between viewpoints

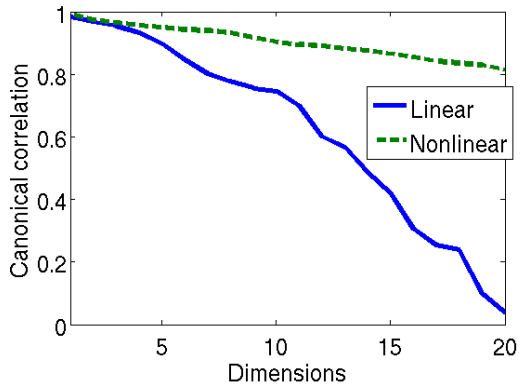
  e.g. Li and Zickler (CVPR 2012), Wu et al. (ICCV 2013), Zhang et al. (CVPR 2013)

Motivation

- Viewpoint changes cause large intra-class variations
  ✓ Features through different viewpoints are shown to be very correlated



(a) Aligned trajectories



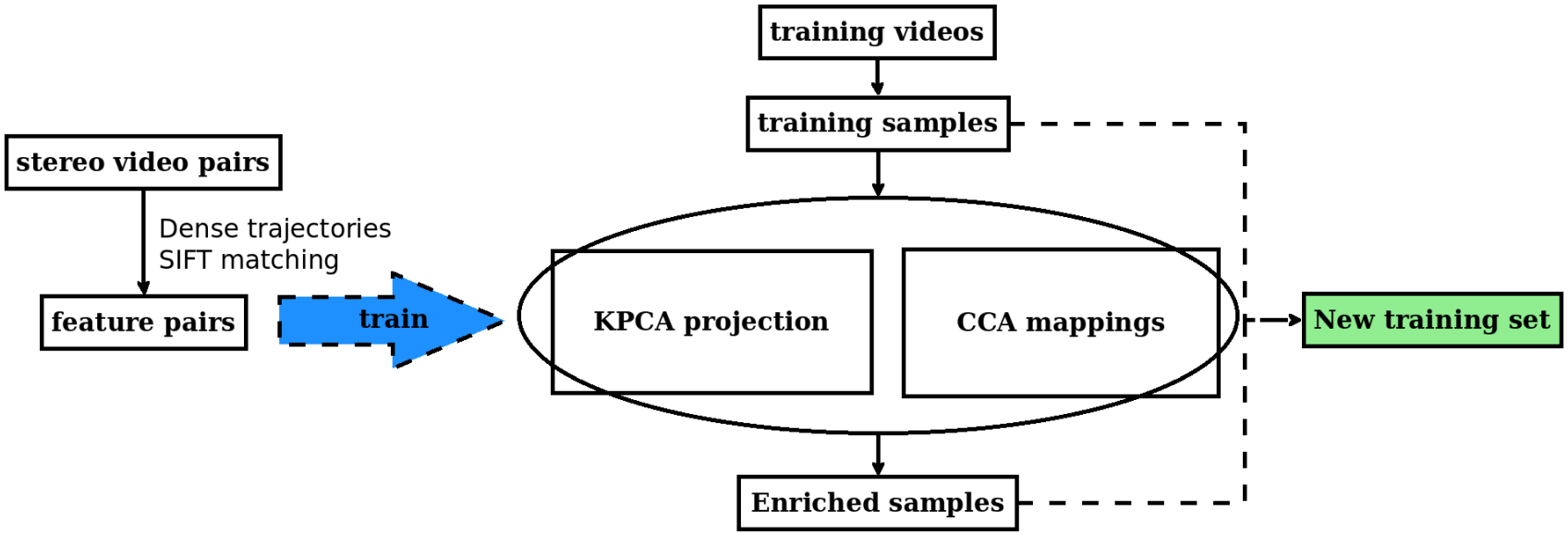
(b) Canonical correlations

- Labeled data are scarce and expensive to collect
  ✓ Training data can be enriched through transferring

Contribution

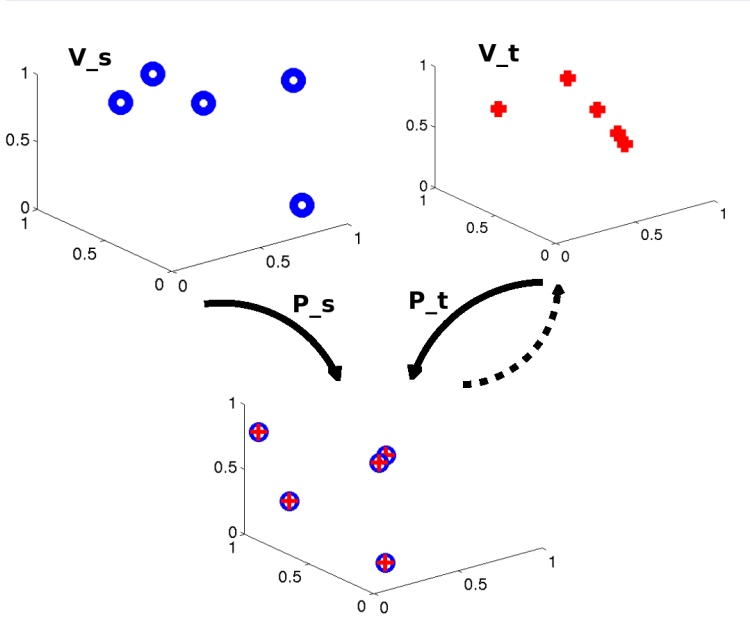
- We propose to enrich video data by transferring their features from few existing training videos (taken from source views) to other views.
- We study the impact of several factors including kernel choices as well as the dimensionality of the latent spaces.

Method overview



Training with canonical correlation analysis

Canonical correlation analysis (CCA) [Hardoon et al., 2004]



$$
\begin{aligned}
(\mathbf{P}_s, \mathbf{P}_t) &= \arg\max_{\mathbf{A},\mathbf{B}} \ \mathbf{A}' C_{st} \mathbf{B} \\
\text{s.t.} \quad \mathbf{A}' C_{ss} \mathbf{A} &= 1 \\
\mathbf{B}' C_{tt} \mathbf{B} &= 1
\end{aligned} \tag{1}
$$

Cross-view feature transfer using CCA

- Projection matrices $\mathbf{P}_s$, $\mathbf{P}_t$ define a common latent space (denoted by $\mathcal{L} \subset \mathbb{R}^d$) in which the correlation between $(\mathbf{P}_s' \, x_i^s, \mathbf{P}_t' \, x_i^t) \in \mathcal{L} \times \mathcal{L}$ is maximized $(i = 1, \ldots, n)$.
- Assuming that mappings $\mathbf{P}_s$, $\mathbf{P}_t$ are invertible (or utilizing Moore-Penrose pseudoinverse), we transfer features $\{\psi(x^s)\}$ (from the source view) to features $\{\psi(x^t)\}$ (in the target view) by

$$\psi(x^t) := (\mathbf{P}_s \mathbf{P}_t^{-1})'(\psi(x^s) - \bar{\psi}^s) + \bar{\psi}^t, \qquad (2)$$

Training dataset

Stereo video dataset [Liu et al., 2010]

- Transformation matrices $\mathbf{P}_s$, $\mathbf{P}_t$ are built from stereo videos that correspond to the same moving actors

(a) (b) (c) (d) (e) (f)

Figure: (b, c, d, e, f) correspond to the source views while (a, b, c, d, e) correspond to the target view

Evaluation dataset

UCF Sport dataset [Lan et al., 2011, Rodriguez et al., 2008]

- 150 videos from TV channels for sport events
- 10 categories (e.g. Diving, Golf, Kicking, Running)
- Split to training (103 videos) and test sets (47 videos)

Influence of KPCA mapping on action recognition

KPCA mapping + linear SVMs (baselines)

Kernels for KPCA \ KPCA dim (p)	64	128	256	512	1024	2048
Linear (baseline)	53.2	**57.4**	–	–	–	–
Polynomial	59.6	**61.7**	61.7	61.7	61.7	61.7
NegDist	61.7	66.0	**68.1**	68.1	68.1	68.1
GHI	68.1	68.1	**72.3**	72.3	70.2	70.2
Gaussian RBF ($\gamma = 0.01$)	66.0	68.1	**72.3**	70.2	70.2	72.3
Gaussian RBF ($\gamma = 1$)	**59.6**	59.6	59.6	59.6	59.6	59.6
Gaussian RBF ($\gamma = 100$)	**59.6**	59.6	59.6	–	–	–
Laplacian RBF ($\gamma = 0.1$)	66.0	68.1	70.2	70.2	**72.3**	72.3
Laplacian RBF ($\gamma = 1$)	66.0	68.1	68.1	68.1	68.1	**70.2**
Laplacian RBF ($\gamma = 10$)	63.8	66.0	**68.1**	68.1	68.1	68.1

Influence of CCA mapping on action recognition

Dimension and kernel choice

	Linear SVMs (LCK)		
	noenrich	enrich perfs w.r.t d	
$p \backslash d$		$d = 64$	$d = 128$
$p = 64$	53.2	**59.6**	-
$p = 128$	57.4	55.3	**61.7**

Table: This table shows action recognition performances (%) with and without the enrichment process for different values of p (related to linear KPCA mapping) and d (related to CCA).

Influence of CCA mapping on action recognition

Dimension and kernel choice

	Linear SVMs (LCK)						
	noenrich	enrich perfs w.r.t d					
$p \backslash d$		64	128	256	512	1024	2048
64	66.0	59.6	–	–	–	–	–
128	68.1	68.1	63.8	–	–	–	–
256	72.3	72.3	70.2	66.0	–	–	–
512	70.2	70.2	**72.3**	70.2	72.3	–	–
1024	70.2	**76.6**	76.6	74.5	72.3	70.2	–
2048	72.3	72.3	74.5	74.5	74.5	**76.6**	72.3

Table: This table shows action recognition performances (%) with and without the enrichment process for different values of p (related to Gaussian RBF KPCA mapping, with $\gamma = 0.01$) and d (related to CCA). Note that $d \leq p$ as the dimension of CCA cannot exceed that of KPCA.

Influence of CCA mapping on action recognition

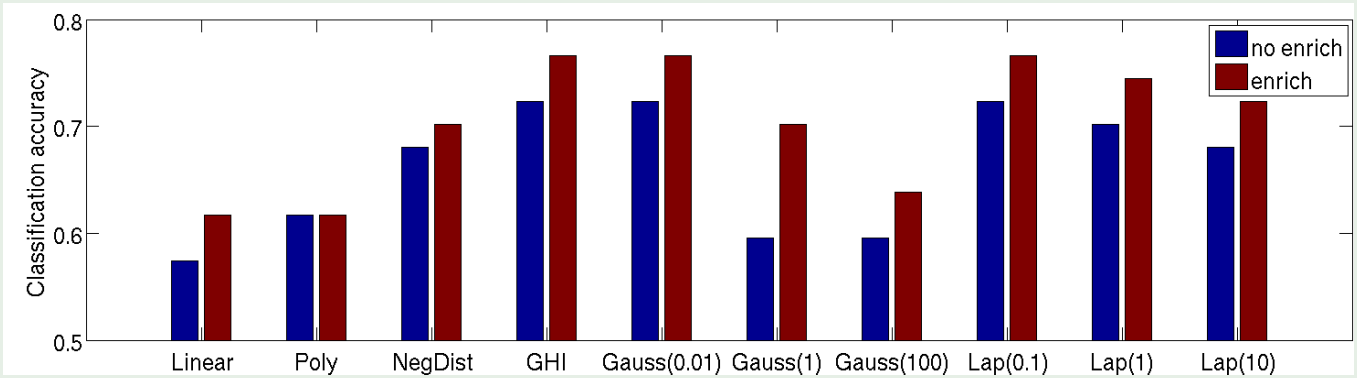
Motion (HOF etc.) v.s. Appearance (HOG) features

	Linear SVMs (LCK)		Nonlinear SVMs (RCK)	
$p \backslash d$	64	128	64	128
64	59.6/**61.7**	-	**80.9**/78.7	-
128	55.3/**61.7**	**61.7**/57.4	**80.9**/78.7	**80.9**/**80.9**

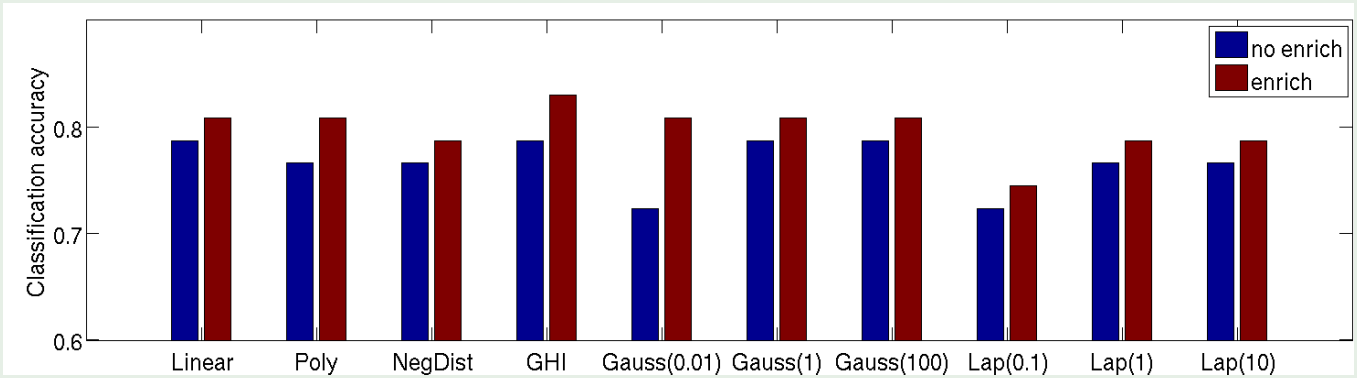
Table: This table shows a comparison between "motion and appearance transfer" vs. "motion transfer only" for different values of p, d. In these results linear kernel is used for KPCA. Note that $d \leq p$ as the dimension of CCA cannot exceed that of KPCA.

Overall performance

Impact of the enrichment process for different KPCA kernels



(a) Linear SVMs (LCK)



(b) Nonlinear SVMs (RCK)

Conclusion

- Inspired from the observation that cross-view features are highly and nonlinearly correlated, we used kernel-based canonical correlation analysis in order to map features across views.
- Experiments conducted show the positive impact of this enrichment process on action recognition and the influence of different (mainly nonlinear) kernels on the performances.

Influence of CCA mapping on action recognition

Dimension and kernel choice

$p \backslash d$	Linear SVMs (LCK)			Nonlinear SVMs (RCK)		
	noenrich	enrich perfs w.r.t d		noenrich	enrich perfs w.r.t d	
		$d = 64$	$d = 128$		$d = 64$	$d = 128$
$p = 64$	53.2	**59.6**	-	76.6	**80.9**	-
$p = 128$	57.4	55.3	**61.7**	78.7	**80.9**	80.9

Table: This table shows action recognition performances (%) with and without the enrichment process for different values of p (related to linear KPCA mapping) and d (related to CCA).

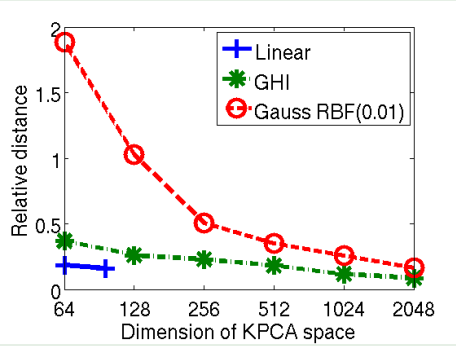
Influence of CCA mapping on action recognition

Dimension and kernel choice

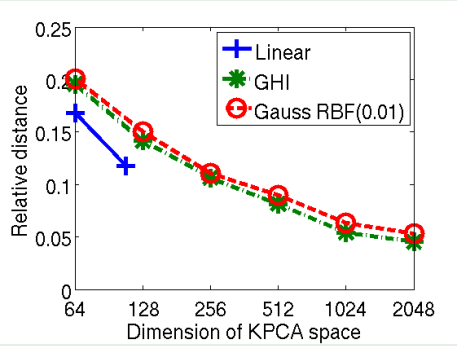
	Linear SVMs (LCK)							Nonlinear SVMs (RCK)						
	noenrich	enrich perfs w.r.t d						noenrich	enrich perfs w.r.t d					
$p \backslash d$		64	128	256	512	1024	2048		64	128	256	512	1024	2048
64	66.0	59.6	–	–	–	–	–	70.2	**72.3**	–	–	–	–	–
128	68.1	68.1	63.8	–	–	–	–	72.3	72.3	72.3	–	–	–	–
256	72.3	72.3	70.2	66.0	–	–	–	72.3	**76.6**	74.5	76.6	–	–	–
512	70.2	70.2	**72.3**	70.2	72.3	–	–	72.3	74.5	76.6	**80.9**	72.3	–	–
1024	70.2	**76.6**	76.6	74.5	72.3	70.2	–	72.3	**74.5**	74.5	74.5	72.3	72.3	–
2048	72.3	72.3	74.5	74.5	74.5	**76.6**	72.3	72.3	74.5	**80.9**	68.1	70.2	70.2	70.2

Table: This table shows action recognition performances (%) with and without the enrichment process for different values of p (related to Gaussian RBF KPCA mapping, with $\gamma = 0.01$) and d (related to CCA). Note that $d \leq p$ as the dimension of CCA cannot exceed that of KPCA.

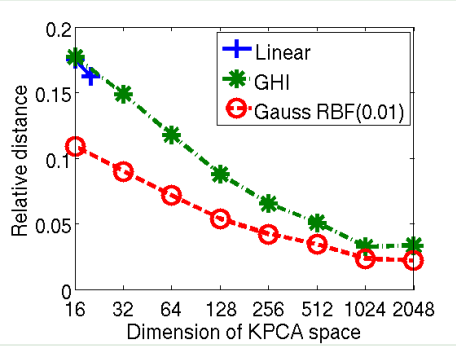
Transfer error



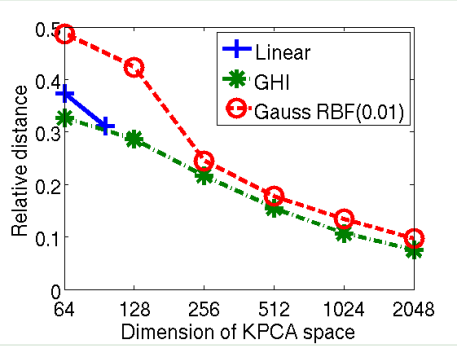
(a) HOG feature



(b) HOF feature



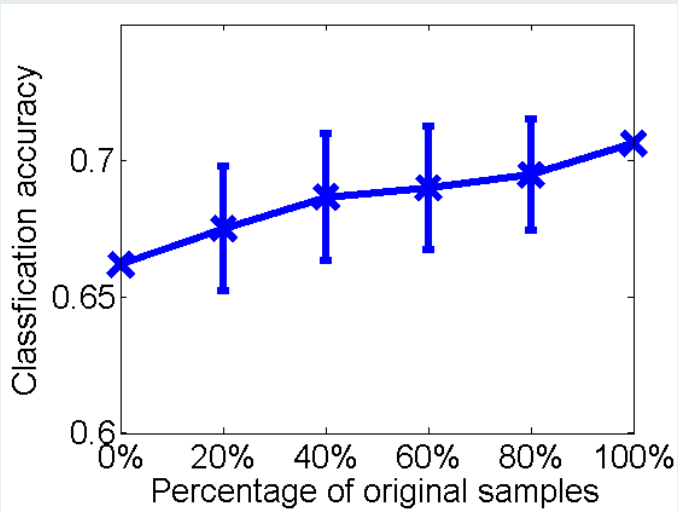
(c) Velocity feature



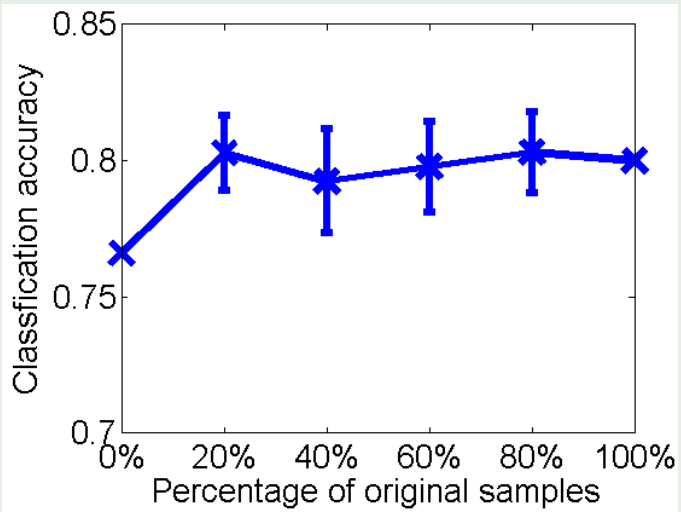
(d) MBH feature

This figure shows the trend of transfer error between generated and ground truth features in target views when increasing the dimension p of KPCA mapping; for fixed p, dim d is set to obtain the full rank p. This transfer error is measured using the average relative distance defined as $dist(x, z) := \frac{1}{n}\sum_{i=1}^{n} ||x_i - z_i||/||z_i||$.

Impact of the amount of enriched data



(a) Linear SVMs (LCK)



(b) Nonlinear SVMs (RCK)

Figure: This figure shows the evolution of action recognition performances w.r.t the fraction k of original training data involved in the enrichment. We report the average classification accuracy of 100 runs.