Action Recognition Using Instance-Specific and Class-Consistent Cues

Chin-An Lin[1,2], Yen-Yu Lin[1], Hong-Yuan Mark Liao[3], and Shyh-Kang Jeng[2]

[1]Research Center for Information Technology Innovation, Academia Sinica, Taiwan

[2]Graduate Institute of Communication Engineering, National Taiwan University, Taiwan; [3]Institute of Information Science, Academia Sinica, Taiwan

1. Summary

- We aim to resolve the difficulties of action recognition arising from the large intra-class variations. These unfavorable variations make it infeasible to represent one action instance by other ones of the same action. We hence propose to extract both instance-specific and class-consistent features to facilitate action recognition.
- Contributions:
  - *Instance-specific features*: Self-similarities among frames of an action sequence. Multivariate linear prediction (MLP) is adopted to aggregate all the causalities among frames.
  - *Class-consistent features*: Characteristics shared by instances of the same action. Support vector machines (SVMs) are used to discover these features based on the bag-of-words model.
  - We propose a generative formulation to integrate the two complementary types of features, and boost the performance.

2. Background

- We view actions as multivariate time signals. For signal processing in our approach, several essential techniques are demonstrated here.

2.1 Wide-Sense Stationary Process

- A discrete stochastic process $\{x(t): t \in T\}$, where T is a countable set, is a collection of random variables.
- The process is said to be wide-sense stationary if $\mathrm{E}[x(t)] = c$, and $\mathrm{Cov}(x(t), x(t+\tau)) = \mathrm{R}(\tau)$ for all t.
- In this paper, we assume actions can be perfectly modeled by wide-sense stationary processes.

2.2 Linear Prediction

- Assume $x(t)$ is a wide-sense stationary process with zero mean, the basic formulation is

$$x(t) = \sum_{k>0} a_k x(t-k) + e(t),$$

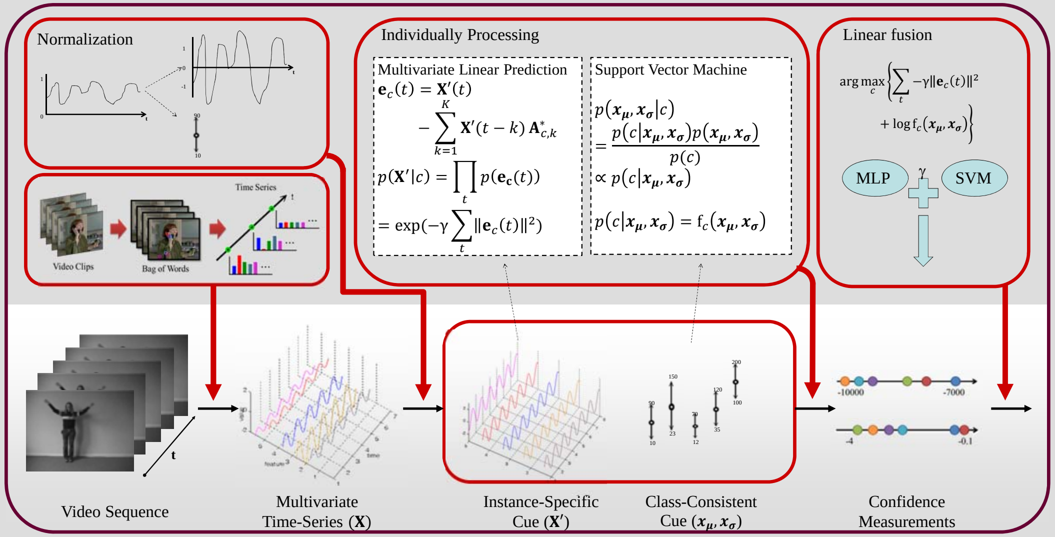
where a_k are coefficients, and $e(t)$ is the reconstruction error.
- We often choose coefficients a_k that minimize the expected value of squared error, $\mathrm{E}[e^2(t)]$.
- The above problem have an optimal solution, and it can be simplified into a linear system.

2.3 Support Vector Machine

- An powerful algorithm for classification problems.
- Given training data $D = \left\{(x_i, y_i) | x_i \in \mathbb{R}^d, y_i \in \{-1, 1\}\right\}_{i=1}^{n}$, we want to find the maximum hyper-plane that divides the points having $y_i = 1$ from those having $y_i = -1$.
- The primal form of the problem is

$$\min_{w,b}\max_{\alpha > 0}\left\{\frac{1}{2}\|w\|^2 - \sum_{i=1}^{n}\alpha_i[y_i(w \cdot x_i - b) - 1]\right\}$$

- We use LIBSVM to solve this problem



3. Generative Model

- The main idea is to consider the static and dynamic information of a multivariate time signal separately. This is based on the assumption that these two information are independent for action recognition.
- The generative formulation for a observed time signal $\mathbf{X}(t), t = 1, \dots, T$ is

$$p(\mathbf{X}, c) = p(\mathbf{X}'|c)p(\boldsymbol{x}_\mu, \boldsymbol{x}_\sigma|c)p(c),$$

where $\mathbf{X}'$is dynamic information containing instance-specific cue, and $(\boldsymbol{x}_\mu, \boldsymbol{x}_\sigma)$ is static information containing class-consistent cue.
- The problem is

$$\arg\max_c\{p(\mathbf{X}'|\boldsymbol{\theta}_c)p(\boldsymbol{x}_\mu, \boldsymbol{x}_\sigma|\boldsymbol{\theta}_c)p(c)\}$$

where $\boldsymbol{\theta}_c$ is the set of model parameters for specific action c.

4.1 Instance-Specific Cue via MLP

- The basic formulation of MLP for a wide-sense stationary process $\mathbf{X}'(t)$ with zero-mean and unit variance is

$$\mathbf{X}'(t) = \sum_{k=1}^{K}\mathbf{X}'(t-k)\mathbf{A}_k + \mathbf{e}(t) = \dot{\mathbf{X}}_t\mathbf{A} + \mathbf{e}(t),$$

$$\mathbf{X}'(t) \in \mathbb{R}^{1XD}, \mathbf{A}_k \in \mathbb{R}^{DXD}, \mathbf{e}(\mathbf{t}) \in \mathbb{R}^{1XD}.$$

- For the action process $\mathbf{X}'_c(t)$ with label c, it can be written as $\mathbf{X}'_c(t) = \dot{\mathbf{X}}_{t,c}\mathbf{A}^*_c + \mathbf{e}_c(t)$, where

$$\mathbf{A}^*_c = \arg\min_{\mathbf{A}} \mathrm{E}[\|\mathbf{e}_c(t)\|^2 + \lambda\|\mathbf{A}\|_F],$$

where$\|\mathbf{A}\|_F$ is *Frobenius norm*, and λ is determined by using *cross-validation.*

In the probability form we can write

$$p(\mathbf{X}'_c(t)|\mathbf{X}'_c(t-1), \dots, \mathbf{X}'_c(t-K), \mathbf{A}^*_c) = p(\mathbf{e}_c(t))$$

- Thus conditional probability is

$$p(\mathbf{X}'_c) = p(\mathbf{X}'|c) = \prod_t p(\mathbf{X}'(t)|\mathbf{X}'(t-1), \dots, \mathbf{X}'(1), \mathbf{A}^*_c)$$

$$= \prod_t p(\mathbf{e}_c(t))$$

In our experiments we choose

$$p(\mathbf{e}_c(t)) = \exp(-\gamma\|\mathbf{e}_c(t)\|^2).$$

4.2 Class-Consistent Cue via SVM

- For $p(\boldsymbol{x}_\mu, \boldsymbol{x}_\sigma|c)$, from Bayes rule

$$p(\boldsymbol{x}_\mu, \boldsymbol{x}_\sigma|c) = \frac{p(c|\boldsymbol{x}_\mu, \boldsymbol{x}_\sigma)p(\boldsymbol{x}_\mu, \boldsymbol{x}_\sigma)}{p(c)} \propto p(c|\boldsymbol{x}_\mu, \boldsymbol{x}_\sigma),$$

where $p(c)$ is the prior knowledge, and we assume $p(\boldsymbol{x}_\mu, \boldsymbol{x}_\sigma)$ is uniformly distributed. Therefore, we use SVM to learn a function $\mathrm{f}_c(\boldsymbol{x}_\mu, \boldsymbol{x}_\sigma)$ to approximate $p(\boldsymbol{x}_\mu, \boldsymbol{x}_\sigma|c)$.

4.3 Linear Fusion

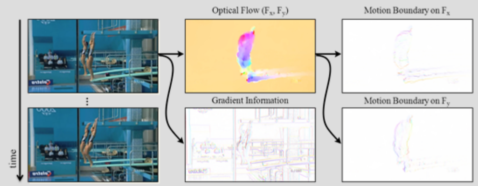
- Conclude the above results, the decision function become

$$\arg\max_c\left\{\sum_t -\gamma\|\mathbf{e}_c(t)\|^2 + \log \mathrm{f}_c(\boldsymbol{x}_\mu, \boldsymbol{x}_\sigma)\right\},$$

where γ weights the importance between dynamic and static information.

5. Video description

- We use three types of features: HOG; HOF; MBH.



- Bag-of-word representation to multivariate time-series
  - We first use Wang's method to describe video into bag-of-word representation.
  - For each time interval, we count words inside it to compile the histogram of words.
  - These histograms are temporally ordered to form a multivariate time-series.

6. Experimental Results

KTH action dataset (6 actions/25 human subjects)

- We follow the author's evaluation protocol
  - 2/3 human subjects for training; 1/3 human subjects for testing
  - Report average accuracy over all classes.

5.1 Results

- The average accuracy of the proposed method

Method	Recognition rate
Wang et al. [6]	94.2%
Chen and Aggarwal [12]	90.9%
Le et al. [15]	93.9%
Ours (instance-specific)	93.9%
Ours (class-consistent)	93.6%
Ours (combined)	**95.0%**

- The confusion matrix on the KTH dataset
  - We get significantly improvement on recognizing between running and jogging.

	walk	run	jog	box	clap	wave
walk	**100%**	0%	0%	0%	0%	0%
run	0%	**82.6%**	*17.4%*	0%	0%	0%
jog	0%	*9.0%*	**91.0%**	0%	0%	0%
box	0.7%	0%	0%	**99.3%**	0%	0%
clap	0%	0%	0%	2.8%	**97.2%**	0%
wave	0%	0%	0%	0%	0%	**100%**