Knowledge Adaptation for Ad Hoc Multimedia Event Detection with Few Exemplars

Z. Ma, Y. Yang, Y. Cai, N. Sebe and A. Hauptmann-ACM MM2012

Yunfei Wang

Department of Computer Science & Technology
Huazhong University of Science & Technology

December 6, 2012

Basic Concepts

Concept An abstract or general idea inferred from specific instances, e.g. fish,sky.

Event An observable occurrence, e.g. making a cake,landing a fish.

Recognition Associate objects that is already known with one or more labels.

Detection Detect the existence of concepts or events coming from an infinite semantic space through pre-trained detectors.

Knowledge Adaptation Also known as transfer learning, propagate knowledge from an auxiliary domain to a target domain.

In Ad Hoc MED,events are more generic and the events are unknown before conducting the detection task.Besides, there are few positive examples for training.

Illustration of framework

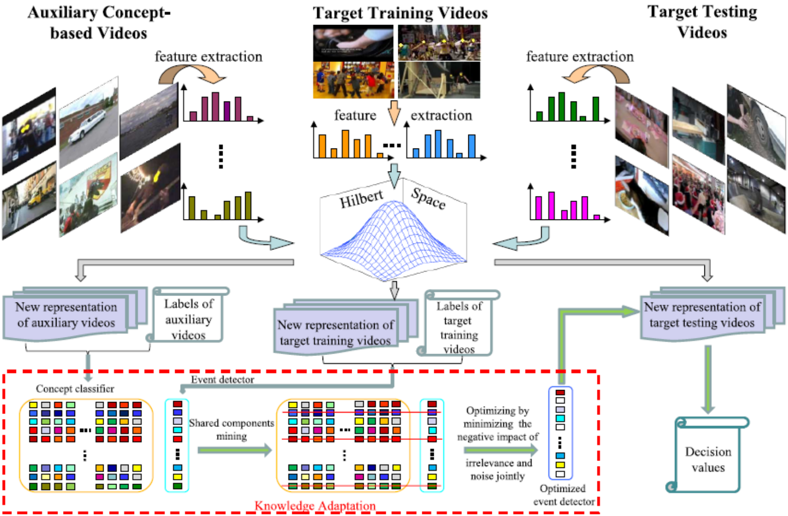

Figure: Framework

Preprocessing each video

Procedures

1. Extract Key Frames using shot boundary detection algorithm
2. Detect Interest Points utilizing Harris-Laplace interest point detector
3. Obtaining SIFT/CIFT features
4. Generate Bag-of-Words feature though clustering SIFT/CSIFT features
5. Map BoW feature into Hilbert Space with kernel trick
6. Perform full rank Principal Component Analysis in another Hilbert Space.

Explore knowledge from target training videos

$\tilde{X}_t = \{\tilde{x}_t^1, \tilde{x}_t^2, \cdots, \tilde{x}_t^{n_t}\} \in \mathbb{R}^{d_h \times n_t}$:training videos in Hilbert Space
$y_t = \{y_t^1, y_t^2, \cdots, y_t^{n_t}\}^T \in \{0,1\}^{n_t \times 1}$:corresponding labels.

Associate low-level representations and high-level semantics of videos by a decision function f:

$$f_t(\tilde{X}_t) = \tilde{X}_t^T W_t + 1_t b_t \tag{1}$$

where $W_t \in \mathbb{R}^{d_h \times 1}$ is an event detector which correlates $\tilde{X}_t$ with labels y_t.

f_t is decided by minimizing the following objective:

$$\min_{f_t} loss(f_t(X_t), y_t) + \mu\Omega(f_t) \tag{2}$$

Using $l_{2,1}$-norm based loss function because it's robust to outliers.Reformulate Eq.(3):

$$\min_{W_t, b_t} \left\| X_t^T W_t + 1_t b_t - y_t \right\|_{2,1} + \mu\Omega(W_t) \tag{3}$$

Adapt knowledge from auxiliary videos

$\tilde{X}_a = \{\tilde{x}_a^1, \tilde{x}_a^2, \cdots, \tilde{x}_a^{n_a}\} \in \mathbb{R}^{d_h \times n_a}$:auxiliary videos.
$Y_a = \{y_a^1, y_a^2, \cdots, y_a^{n_a}\}^T \in \{0,1\}^{n_a \times c_a}$:label matrix.

Mine the correlation between low-level representations and high-level semantics of the auxiliary concepts-based videos.

$$\min_{W_a, b_a} \left\| \tilde{X}_a^T W_a + 1_a b_a - Y_a \right\|_{2,1} + \gamma \Omega(W_a) \tag{4}$$

where $W_a \in \mathbb{R}^{d_h \times c_a}$ is a concept classifier.

Bridge the gap between concepts and event

1. Knowledge adaptation is based on the assumption that there are shared structures between the source and the target.
2. The shared noisy and irrelevant components in video representation weaken the performance of event detector.So they must be removed.
3. Concepts of $\tilde{X}_a$ and events of $\tilde{X}_t$ are related and grounded on similar low-level representations by W_a and W_t respectively.So the irrelevant or noisy components is similar in W_a and W_t,which can be uncovered by learning W_a and W_t jointly.

Objective function

Joint information

event detector: $W_t = \left[w_t^1, w_t^2, \cdots, w_t^{d_h}\right]$
concept classifier: $W_a = \left[w_a^1, w_a^2, \cdots, w_a^{d_h}\right]$
joint analyzer: $W = \left[w^1, w^2, \cdots, w^{d_h}\right]$, reflecting joint information from auxiliary videos and training videos, where $w^i = [w_a^i, w_t^i]$.

Remove shared irrelevant and noisy components using sparse model

$$\min \|W\|_{2,p} = \left(\sum_{i=1}^{d_h}\left(\sum_{j=1}^{c_a+1} W_{ij}^2\right)^{\frac{1}{2}}\right)^{2-p}$$

Final objective function

$$\min_{W_a, W_t, b_a, b_t} \left\|\tilde{X}_a^T W_a + 1_a b_a - Y_a\right\|_{2,1} + \left\|\tilde{X}_t^T W_t + 1_t b_t - y_t\right\|_{2,1} + \alpha\left(\sum_{i=1}^{d_h}\left(\sum_{j=1}^{c_a+1} |W_{ij}|\right)^{\frac{p}{2}}\right)^{\frac{1}{p}} + \beta\left(\|W_a\|_F^2 + \|W_t\|_F^2\right)$$

Algorithm

Input: Auxiliary data $\tilde{X}_a \in \mathbb{R}^{d_h \times n_a}, Y_a \in \mathbb{R}^{n_a \times c_a}$;
Training data $\tilde{X}_t \in \mathbb{R}^{d_h \times n_t}, y_t \in \mathbb{R}^{n_t \times 1}$; Parameters α,β.
Output: Optimized $W_t \in \mathbb{R}^{d_h \times 1}$ and $b_t \in \mathbb{R}^1$.

Set $t = 0$,initialize $W_a \in \mathbb{R}^{d_h \times c_a}$ and $W_t \in \mathbb{R}^{d_h \times 1}$ randomly;
repeat

- Compute $\tilde{X}_a^T W_a + 1_a b_a - Y_a = \left[u^1, \cdots, u^{n_a}\right]^T, \tilde{X}_t^T W_t + 1_t b_t - y_t = \left[v^1, \cdots, v^{n_t}\right]^T$,and $W = \left[w^1, \cdots, w^d\right]^T$;
- $D_a^{ii} = \frac{1}{2\|u^i\|_2}$,$D_t^{ii} = \frac{1}{2\|v^i\|_2}$,and $D^{ii} = \frac{1}{\frac{2}{p}\|w^i\|_2^{2-p}}$;
- $W_a^{t+1} = (\tilde{X}_a H_a D_a H_a \tilde{X}_a^T + \alpha D + \beta I_d)^{-1} \tilde{X}_a H_a D_a H_a Y_a$;
- $b_a^{t+1} = \frac{1}{n_a} 1_a^T Y_a - \frac{1}{n_a} 1_a^T \tilde{X}_a^T W_a^{t+1}$;
- $W_t^{t+1} = (\tilde{X}_t H_t D_t H_t \tilde{X}_t^T + \alpha D + \beta I_d)^{-1} \tilde{X}_t H_t D_t H_t y_t$;
- $b_t^{t+1} = \frac{1}{n_t} 1_t^T y_t - \frac{1}{n_t} 1_t^T \tilde{X}_t^T W_t^{t+1}; t = t + 1$;

until *Convergence*;
return W_t and b_t.

Algorithm 1: Optimizing the event detector