Lixin Duan Nanyang Technological University
Dong Xu Nanyang Technological University
Shih-Fu Chang Columbia University

CVPR '12

Exploiting Web Images for Event Recognition in Consumer Videos: A Multiple Source Domain Adaptation Approach

Contributions

- We present a new method called Domain Selection Machine (DSM) to take advantage of abundant freely available web images for event recognition in consumer videos.
- DSM automatically selects the most relevant source domains with our newly introduced data-dependent regularizer.
- We integrate different types of features (i.e., SIFT features from images and space-time features from videos) from different domains by using our proposed target decision function.

Background

- Event recognition in consumer videos is important in video indexing and retrieval, but it is also very challenging due to unconstrained camera motion and large intra-class variations.
- The recent work [Ref 1] developed an event recognition approach by using web videos from YouTube. We also observe that there are much more web images from different sources.
- We only have few or even no labeled consumer videos for training. Data distributions from the consumer video domain and web image domain are different.

Domain Selection Machine

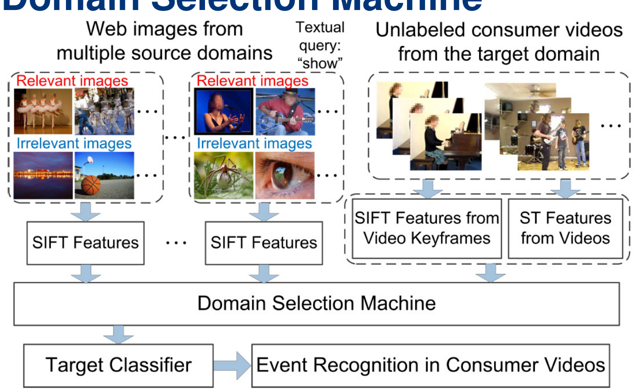

Figure 1. Illustration of our proposed method Domain Selection Machine (DSM) for event recognition in consumer videos.

Regularizer for source domain selection

$$\Omega(f) = \frac{1}{2}\sum_{s=1}^{S} d_s \sum_{i=1}^{m}\left(f^T(\mathbf{x}_i^T) - f^s(\mathbf{x}_i^T)\right)^2$$

- $d_s \in \{0,1\}$ is a domain selection indicator for the s-th source domain
- f^T: target classifier
- f^s: pre-learned source classifier

Integrating SIFT and ST features in the target classifier

$$f(\mathbf{x}) = f_{2\mathrm{D}}(\mathbf{x}) + f_{3\mathrm{D}}(\mathbf{x})$$
$$= \sum_{s=1}^{S} d_s\beta_s f^s(\mathbf{x}) + \mathbf{w}'\varphi(\mathbf{x}) + b$$

- $f_{2\mathrm{D}}(\mathbf{x}) = \sum_{s=1}^{S} d_s\beta_s f^s(\mathbf{x})$ based on SIFT features
- $f_{2\mathrm{D}}(\mathbf{x}) = \mathbf{w}'\varphi(\mathbf{x}) + b$ based on ST features

Formulation of DSM

$$\min_{\mathbf{d},\mathbf{w},b,\boldsymbol{\beta},\mathbf{f}^T} \frac{1}{2}(\|\mathbf{w}\|^2 + \|\boldsymbol{\beta}\|^2) + C\sum_{i=1}^{m}\ell\left(f^T(\mathbf{x}_i^T) - f(\mathbf{x}_i^T)\right) + \theta\cdot\Omega(f)$$

$$\text{s.t.} \sum_{s=1}^{S} d_s \geq 1, \quad d_s \in \{0,1\}$$

- We solve the optimization problem by iteratively updating $\{\mathbf{w}, b, \boldsymbol{\beta}, \mathbf{f}^T\}$ and $\mathbf{d}$.

Experiments

Datasets

- Kodak [Ref 1]: 195 consumer videos
- YouTube: 561 consumer videos
- CCV [Ref 2]: 2726 consumer videos

Results

Table 1. Mean Average Precisions (MAPs) of all methods on the Kodak dataset.

	SVM_A	DASVM	Multi-KMM	DAM	CP-MDA	DSM_{sim}	DSM
MAP	27.95%	25.68%	24.22%	27.66%	24.41%	33.67%	**35.46%**

Table 2. Mean Average Precisions (MAPs) of all methods on the YouTube dataset.

	SVM_A	DASVM	Multi-KMM	DAM	CP-MDA	DSM_{sim}	DSM
MAP	31.17%	29.40%	31.98%	32.58%	30.27%	33.75%	**35.26%**

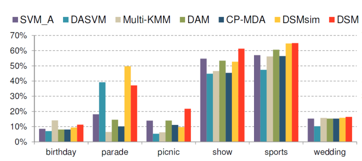

Figure 2. Per-event Average Precisions (APs) of all methods on the Kodak dataset.

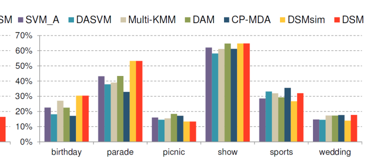

Figure 3. Per-event Average Precisions (APs) of all methods on the YouTube dataset.

Table 3. Mean Average Precisions (MAPs) of all methods on the CCV dataset.

	SVM_A	DASVM	Multi-KMM	DAM	CP-MDA	DSM_{sim}	DSM
MAP	17.14%	18.38%	19.77%	17.01%	17.49%	17.80%	**21.76%**

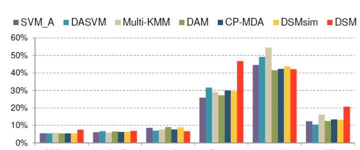

Figure 4. Per-event Average Precisions (APs) of all methods on the CCV dataset.

References

[Ref 1] L. Duan, X. Dong, I. W. Tsang, and J. Luo. Visual Event Recognition in Videos by Learning from Web Data. In CVPR, 2010.

[Ref 2]: Y.-G. Jiang, G. Ye, S.-F. Chang, D. Ellis, and A. C. Loui. Consumer Video Understanding: A Benchmark Database and An Evaluation of Human and Machine Performance. In ICMR, 2011.