Available online at www.sciencedirect.com

ScienceDirect

AASRI Procedia 6 (2014) 111 – 117

AASRI Procedia

www.elsevier.com/locate/procedia

2013 2nd AASRI Conference on Computational Intelligence and Bioinformatics

Action-Scene Model for Human Action Recognition from Videos*

Yifei Zhang+, Wen Qu, Daling Wang

Northeastern University, Shenyang 110819, China

Abstract

Human action recognition from realistic videos attracts more attention in many practical applications such as on-line video surveillance and content-based video management. Single action recognition always fails to distinguish similar action categories due to the complex background settings in realistic videos. In this paper, a novel action-scene model is explored to learn contextual relationship between actions and scenes in realistic videos. With little prior knowledge on scene categories, a generative probabilistic framework is used for action inference from background directly based on visual words. Experimental results on a realistic video dataset validate the effectiveness of the action-scene model for action recognition from background settings. Extensive experiments were conducted on different feature extracted methods, and the results show the learned model has good robustness when the features are noisy.

Peer-review under responsibility of Scientific Committee of American Applied Science Research Institute

Keywords: action recognition, contextual cue, video processing, action-scene model

* Supported by the Key Program of National Natural Science Foundation of China under Grant No.61033007 and the Fundamental Research Funds for the Central Universities of China under Grant No.N100304004.

+ Corresponding author. Tel.: +86-(0)24-83687776
E-mail address: zhangyifei@mail.neu.edu.cn

Peer-review under responsibility of Scientific Committee of American Applied Science Research Institute
doi:10.1016/j.aasri.2014.05.016

1. Introduction

As more and more video surveillance systems appear in every walk of life, automatic human action recognition from realistic videos has become a major concern for both academic researchers and commercial companies. In action recognition, one fundamental problem is to recognize whether an action is belong to a given action category. Although a great amount of impressive results have been achieved on datasets collected from controlled environments, such as KTH [1], WEIZMEN [2], much less progresses have been made on realistic videos due to complex background settings and action alternations from realistic videos. However, the intricate background and varied actions equally give us a clue to explore the contextual relationship between actions and scenes for action recognition on realistic videos.

As human actions always occur under particular scenes where a rich source of contextual cues will present, the contextual relationships of background settings could be learned as a complement for action recognition. Although recognition based on scene detectors have become commonplace in the related literatures [3,4], a main problem of detector based methods is how to select general scene and action categories for all videos. Otherwise, training detectors is time-consuming and performances of recognition will be affected by the learned detectors.

In this paper, we intend to learn contextual cues using a generative framework with little prior knowledge on scene categories. The contextual relationship between actions and scenes could be used to infer actions from background settings. An action-scene model is proposed to automatically model the relationship between actions, scenes and background features, where the number of scenes only need to be given instead of the categories of scenes. Firstly, a video will be segmented into person regions and background regions. Then the relationship of a given action category and scenes will be modeled using an action-scene model. Finally a factor is computed based on the proposed model to measure the importance of learned context cue and the responding probability of action category is given from this model.

The rest of this paper is organized as follows. Section 2 reviews related work. Section 3 describes the process of feature extraction and video representation. Action-scene model is proposed in Section 4 and experimental results are shown in section 5. Finally we present concluding remarks and future work in Section 6.

2. Related Work

Human action recognition has been an important area of research in the field of computer vision. From single scenario to complicated scenario, various approaches are presented for action recognition to meet different application demands.

Early works were focus on action recognition with single scene [5]. However, this scenario is too ideal to appear in realistic circumstances. As more attention paid to complex scenarios, some recent approaches [3,6,7,8] tried to deal with “wild” videos. Laptev et al automatically annotated movie videos using multi-channel non-linear SVMs [6]. Liu et al proposed a framework to recognize actions which combined both motion and static features [7]. These appear-based learning approaches work hard to select general scene types and object categories suitable for all videos. To handle the complex realistic videos, contextual semantics are used in lots of works. Ikizler-Cinbis et al combined the features of object, scene and actions in a multiple instance learning framework [9]. However, they didn’t learn the relationship between objects, actions and scenes for action identification. Marszalek et al exploited the context of movie scripts and developed a joint scene-action SVM-based classifier by training several scene detectors [4]. However, due to a lack of training data for all the potential objects and scenes, it is time-consuming to collect annotated data for each category. In addition, a semantic gap would be generated between the learned contextual cues from texts and

visual features.

In this paper, we propose a novel approach different from the detector-based methods in learning contextual cues that has no use for action and scene categories. Inspired by the recent work of action recognition in static images [10] and unsupervised learning method in [2], we use a generative model to represent action and scene in videos. The underlying action-scene dependency is captured on the action-scene model directly from visual features.

3. Visual Feature Extraction and Video Representation

3.1. Body detection

The person body detection in videos is based on Felzenszwalb's object detector [11] and mean shift tracking [12]. We first use the object detector to find candidate person regions from each frame image. Since there are many false alarms and miss-detections, a sliding window based method is used to discard false alarms and track person area for missing detections. Two detections are viewed as similar if the bounding boxes overlap exceeds 40%. The window size is empirically set as 15 frames and the threshold is half of the windows size. Then, the miss detections are filled with tracking from previous frames using mean shift algorithm. The person detection provides the bonding boxes of person location at each frame.

3.2. Features extraction

- **Person-centric feature extraction.** We use the spatio-temporal interest point detector proposed by Dollar [13] to capture the spatially and temporally interesting movements. After extracting the person-centric features from a video clip, we describe each interest point with HOG (Histograms of oriented Gradients) [14] descriptor.
- **Background feature extraction**. The non-person region is seen as background region. In order to capture the color, shape and local features of the background, we extract color histograms for color features, Gist descriptors [15] for shape features and SIFT descriptors [16] from key frames for static features.

3.3. Bag-of-features video representation

Bag-of-feature based video representation has approved to be effective for action recognition in unstrained videos [7]. The bag-of-feature model represents a video clip as a vector over a feature vocabulary or "visual codebook". The "visual codebook" is constructed by clustering visual features detected from all videos. The center of each cluster is defined to be a codeword or visual word in the codebook. Thus, each detected feature in a video clip can be assigned to a unique cluster having the nearest distance. The *i*th bin of bag-of-feature is the number of features that be assigned to the *i*th cluster in the video.

4. Learning of Action-scene Relationship

The detector based context learning needs to know the prior knowledge about scene categories which is usually dependent on the dataset. Therefore, the detector-based method is usually inflexible to the general datasets, as well as the unstable performance of detectors is prone to affect the accuracy. In this paper, we aim to identify the action class in a video from scene features with less prior knowledge. Compared with scene categories occur in the dataset, the approximate number of them is easier to get.

According to the relation between videos, actions, scenes and visual words, we naturally deduce a

generative probabilistic model to model the contextual cues between them. In the action-scene model, each video can be seen as a mixture of action categories, where each action is a probability distribution over scenes and each scene is associated with a distribution over visual words. Each video clip is looked as a distribution of actions, and each action is a distribution of visual word. The action-scene model discovers not only which action happens in a video, but also which scene is associated with the action.

Suppose we have a collection of V video clips, each video v is represented as a set of visual words. Each word belongs to the visual codebook including W unique visual words. Assume we have K action categories, S scene categories. Now we can describe the process generating each video v as:

1) For each clip v, there is a multinomial distribution over K action categories $\theta_v \sim Dirichlet(\alpha)$.

2) For each visual word vi of the clip v:

a) Choose an action category z_{vi} *according to Multinomial distribution* (θ_v).

b) Choose a scene s_{vi} *from* $p(s_{vi}|z_{vi}, \psi_z)$*, a multinomial probability conditioned on the action* z_{vi} *and* ψ_z*. Here,* ψ_z *is a multinomial distribution of action over scene* s_{vi} *with* $\psi_z \sim Dirichlet(\beta)$.

c) Choose a visual word w_{vi} *from* $p(w_{vi}|s_{vi}, \varphi_s)$*, a multinomial probability conditioned on the scene* s_{vi} *and* φ_s*. And* φ_s *is the distribution of scene over visual words with* $\varphi_s \sim Dirichlet(\xi)$.

Here, θ, φ and ψ are multinomial distribution associated with video clip v, scene s and action z. Under this generative processing, the probability of video corpus on visual word set w, conditioned on α, β and ξ is:

$$\begin{aligned} \boldsymbol{P}(\boldsymbol{w},\boldsymbol{z},\boldsymbol{s},\theta,\psi \mid \alpha,\beta,\xi) = & \prod_{i=1}^{N_v} \boldsymbol{P}(\boldsymbol{w}_{vi} \mid \boldsymbol{S}_{vi},\varphi_{s_{vi}})\boldsymbol{p}(\varphi_{s_{vi}} \mid \beta) \\ & \prod_{i=1}^{N_v} \boldsymbol{P}(\boldsymbol{w}_{vi} \mid \boldsymbol{z}_{vi},\psi_{z_{vi}})\boldsymbol{p}(\psi_{z_{vi}} \mid \xi) \\ & \prod_{v=1}^{V}\prod_{i=1}^{N_v} \boldsymbol{P}(\boldsymbol{z}_{vi} \mid \theta_v)\boldsymbol{p}(\theta_v \mid \alpha) \end{aligned} \tag{1}$$

This action-scene model includes the following parameters: the action distribution of video clip θ, action-scene distribution ψ, the scene distribution φ and the assignments of individual words to action z and scene s. In this paper, we use Gibbs sampling [17] to evaluate the posterior distribution on z and scene s. Then θ, φ and ψ can be inferred from the results of z and s.

Given z, s, w, α, β and ξ, computing the posterior distributions on θ, φ and ψ is straightforward. Using the hypothesis that the *Dirichlet* is conjugate to the ***multinomial***, we can estimate θ, φ and ψ with:

$$\theta_{zv} = \frac{\boldsymbol{C}_{zv}^{\boldsymbol{ZV}} + \alpha}{\sum \boldsymbol{C}_{zv}^{\boldsymbol{ZV}} + \boldsymbol{K}\alpha},\ \varphi_{ws} = \frac{\boldsymbol{C}_{ws}^{\boldsymbol{WS}} + \beta}{\sum \boldsymbol{C}_{ws}^{\boldsymbol{WS}} + \boldsymbol{W}\beta},\ \psi_{sz} = \frac{\boldsymbol{C}_{sz}^{\boldsymbol{SZ}} + \xi}{\sum \boldsymbol{C}_{sz}^{\boldsymbol{SZ}} + \boldsymbol{S}\xi} \tag{2}$$

where $\boldsymbol{C}_{zv}^{\boldsymbol{ZV}}$ is number of the words from video v been assigned to an action z, $\boldsymbol{C}_{ws}^{\boldsymbol{WS}}$ is the time that a word w been assigned to a scene s, *and* $\boldsymbol{C}_{sz}^{\boldsymbol{SZ}}$ is the time that a scene s been assigned to an action z, and $\sum \boldsymbol{C}_{zv}^{\boldsymbol{ZV}}$, $\sum \boldsymbol{C}_{ws}^{\boldsymbol{WS}}$ and $\sum \boldsymbol{C}_{sz}^{\boldsymbol{SZ}}$ respectively denote the total number of words in video v, the total number of words assigned to scene s, and the total number of scenes assigned to action z. Then for each word, the basic equation for the Gibbs sampler is:

$$P(z_{vi} = x, s_{vi} = j \mid w_{vi} = m, s_{-vi}, z_{-vi}, w_{vi}, \alpha, \beta, \xi) = \theta_{zv}\varphi_{ws}\psi_{sz} \quad (3)$$

where z_{vi}=z, s_{vi}=s represent the assignments of the ith word in a video clip to action z and scene s respectively. w_{vi}=w represents the observation that the ith word is the wth word in the codebook, and z_{-vi}, s_{-vi} represent all action and scene assignment not including the ith word. Since the assignment of visual words to actions and scenes will be given quickly from equation 3, the recognition processing will be carried out by tracking three count matrixes $\boldsymbol{C}^{ZV}$, $\boldsymbol{C}^{WS}$ and $\boldsymbol{C}^{SZ}$.

In the training stage, the Gibbs sampling is applied over all the visual words in the training set while the count matrixes will be computed and saved. When classifying a new video clip, the Gibbs sampling is run on the visual word in this clip. The assignments of visual words to actions and scenes will be evaluated from equation 3. The count matrices are updated from each assignment. The video clip is classified with a category index with the largest probability in θ_v. Since θ_v only draw the index of action category without knowing the actual action class label for the index, an action class label is assigned to the index by using ground truth labels in the training dataset. For each action index, it is named with the most popular action class label within videos belongs to this index.

5. Experimental Results

We evaluate our approach on the challenging YouTube dataset provided by Liu et al [7] which consists of 1168 videos collected by Liu et al. The dataset covers 11 action classes: basketball shooting, biking, diving, golf swing, horse riding, soccer juggling, swing, tennis swing, trampoline jumping, volleyball spiking, and walking with a dog. Each category of action is performed under several different scenes and divided into 25 subsets.

The hidden parameter θ, φ and ψ were estimated based on training set. For all models we fixed the number of actions at K=11 and S is set to be the actual number of scene categories in dataset. The other parameters are set as α=50/K, β = 0.3 and ξ = 0.01 from experimental analysis.

We compare our action-scene model with LDA[18]. The average performances with two models are shown in Table 1. We see that the proposed method achieve an improvement over all the action categories. The improvement is especially remarkable for the actions that have strong relation with scenes: basketball (15.13%) and trampoline jump (10.72%). It demonstrates that the scene cues are informative and complementary for identifying the actions in realistic videos.

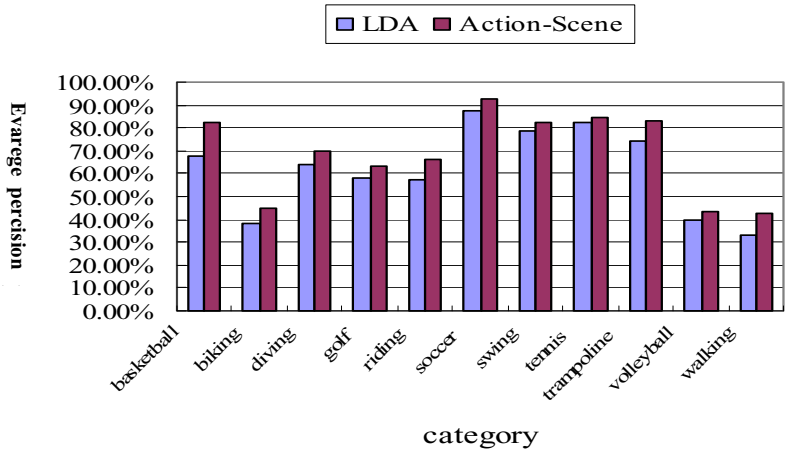

Fig. 1. The Recognition performance of action recognize for Youtube Dataset

To evaluate the effect of person detector on the recognition performance, we compare the results using manually annotated person areas and the areas from person detectors in Section 3.1, and the results in Table 1 show that the performance of LDA can receive a better performance than the action-scene model when the person areas are annotated precisely, and a worse performance when the person areas are automatically detected. The action-scene model has better robustness than LDA when the features are noisy.

Table 1. The Recognition performance with different feature extraction methods

Recognition Model	Annotation	Detector
LDA	78.3%	62.4%
Action-Scene	73.5%	76.8%

6. Conclusion

In this paper, we explore to recognize human actions from background settings of realistic videos. An action-scene model is proposed to model and learn the relationship between actions and scenes with little prior knowledge on scene categories. A generative learning method is used to infer actions from scenes according to a certain distribution. Experimental results validate that the addition of the context cues indeed improve the recognition precision. Besides, our proposed action-scene model has good robustness when applied in realistic video datasets.

There are a wide variety of potential applications on the research of action recognition. For further work, we intend to combine our method with known methods to learn application specified contextual cues from the massive realistic videos.

References

[1] C.Schuldt, I.Laptev, and B.Caputo, “Recognizing human actions: a local SVM approach,” in ICPR, 2004.
[2] J.Niebles, H.Wang, and L.Fei-Fei, “Unsupervised learning of human action categories using spatial-

temporal words," International Journal of Computer Vision, vol. 79, issue 3, pp. 299-318, 2008.
[3] T.Motwani, R.Mooney, "Improving video activity recognition using object recognition and text mining," in ECAI, 2012.
[4] M.Marszalek, I.Laptev, C.Schmid, "Actions in context," in CVPR, 2009.
[5] M. Blank, L. Gorelick, E. Shechtman, "Actions as space-time shapes," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529-551, April 1955.
[6] Laptev, M.Marszalek, C.Schmid, B.Rozenfeld, "Learning realistic human actions from movies," in CVPR, 2008.
[7] J.Liu, J.Luo, M.Shah, "Recognizing realistic actions from videos "in the wild" in CVPR, 2009.
[8] D.Han, L.Bo, C.Sminchisescu, "Selection and context for action recognition," in ICCV, 2009.
[9] N.Ikizler-Cinbis, S.Sclaroff, "Object, scene and actions: combining multiple features for human action recognition," in ECCV, 2010.
[10] B.Yao, L.Fei-Fei, "Modeling mutual context of object and human pose in human-object interaction activities," in CVPR, 2010.
[11] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained partbased models, " IEEE Trans. Pattern Anal. Mach. Intell., 2010.
[12] D. Comaniciu, V. Ramesh, P. Meer, "Real-time tracking of non-rigid objects using mean shift, " in CVPR, 2000.
[13] P.Dollar, V.Rabaud, G.Cottrell, S. Belongie. "Behavior recognition via sparse spatio-temporal features," in 2nd joint IEEE International workshop on visual surveillance and performance evaluation of tracking and surveillance, 2005.
[14] N. Dalal, B. Triggs, "Histograms of oriented gradients for human detection," in CVPR, 2005.
[15] A.Oliv, A.Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope," International Journal of Computer Vision, vol. 42, issue 3, pp. 142-175, 2001.
[16] D.Lowe. "Distinctive image features form scale-invariant keypoints," International Journal of Computer Vision, vol. 60, issue 2, pp. 91-110, 2004.
[17] B.Schölkopf, A.Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond, MIT Press, 2002.
[18] D.Blei, A.Ng, M.Jordan, "Latent dirichlet allocation," Journal of Machine Learning Research, issue 3, pp. 993-1022, 2003.