EXPLOITING PROBABILISTIC RELATIONSHIPS BETWEEN ACTION CONCEPTS FOR COMPLEX EVENT CLASSIFICATION

Somayeh Keshavarz *Imran Saleemi* *George Atia*

College of Engineering and Computer Science, University of Central Florida
Emails: somayehkeshavarz@knights.ucf.edu, imransaleemi@gmail.com, george.atia@ucf.edu

ABSTRACT

Videos of complex events are difficult to represent solely as bags of low level features. Increasingly, supervised concepts or attributes are being employed as the intermediate representation of such videos. We propose a probabilistic framework that models the conditional relationships between the concepts and events and devise an approximate yet tractable solution to infer the posterior distribution to perform event classification. Using noisy outputs of pre-trained concept detectors, we learn semantic and visual dependencies between event and concept pairs. The co-occurrence between concept pairs is also learned as a marginal over training samples. The proposed method then employs the learned prior, as well as the probabilities of occurrence of specific concepts in a test video to infer the probability of each event using weighted average one-dependence estimation. The evaluation shows that our method improves event classification compared to recent literature on the TRECVID data set.

***Index Terms*—** Bayesian Network, Probabilistic model, Statistical learning, Event classification

1. INTRODUCTION

Audio-visual data, specifically open-source web videos, are a massive source of information in a rapidly changing digital world. Videos are uploaded by amateur consumers under unconstrained camera motion, varying illumination conditions and high degree of background clutter. In addition, the wide visual diversity in scene settings under which these videos are captured makes video classification an extremely challenging research problem.

A significant body of research has been reported in the context of video classification. To this end, researchers have proposed the use of raw audio-visual features [1, 2, 3] in different variants of bag-of-video-words framework [4, 5, 6]. Sophisticated feature weighting [7] and fusion [8] have also been proposed to improve event recognition performance. Although these approaches perform satisfactorily to some extent, they are incapable of providing any understanding of the semantic structure present in a complex event depicted in a video. Therefore, a logical and computationally tractable way is to decompose a video depicting a complex event into a sequence of spatio-temporal actions called concepts. These concepts are expected to provide a meaningful intermediate level of abstraction towards representing a complex event. For instance, “changing a vehicle tire” is a complex event where the objects human, vehicle, tire and tools interact.

While various techniques were proposed in this context, we only describe related work on Bayesian Networks (BN), which is the statistical model that we use. A BN [9] is a directed acyclic graph that represents a joint probability distribution over a set of random variables. A large body of work on classification using BN has been developed. For example, Kafai et al. [10] used a BN for vehicle classification. However, their proposed structure was constructed manually and only works with a very small number of variables. In fact, learning an optimal BN is generally NP-hard [11]. As such, Silander and Myllymaki [12] introduced a simple method to learn the structure of a BN, albeit their method is only applicable for small structures.

We aim to learn an improved approximation of the joint distribution of concepts and events. The main idea behind this is to perform a weighted averaging of multiple pairwise conditional distributions to model the full conditional. Our approach scales well with the number of concepts.

The rest of the paper is organized as follows. We introduce our proposed approach in Section 2. The solution for event classification using a BN is presented in Section 3. Section 4 presents the experimental results on two different datasets. Finally, we conclude in Section 5.

2. PROPOSED APPROACH

We propose to formulate event detection as a Bayesian inference problem, where we have confidence $p\left(e \mid v\right)$ (probability of event e in video v). We use capital and small letters to denote random variables and their realizations, respectively. Let $E \in \mathcal{E}$ denote a random variable of a specific complex multimedia event, where $\mathcal{E} = \{1, 2, ..., m\}$. Let k denote the number of concepts. A confidence score can simply be obtained from m trained classifiers using typical low-level bag-of-feature representations [4, 6, 13]. Moreover,

978-1-5090-2175-8/17/$31.00 ©2017 IEEE

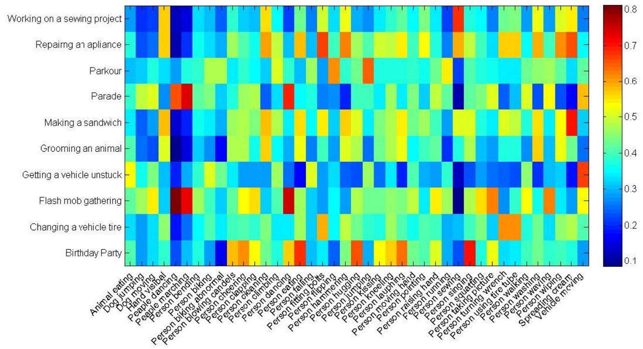

Fig. 1: Representing the frequency of various annotated concepts across several complex events. Each column represents a concept while each row represents an event.

we are going to benefit from the fact that concepts have relations with events. For instance, the event "changing vehicle tire" and the concept "turning wrench" are highly correlated while this event is not correlated for example with the concept "skateboarding". Figure 1 shows a typical event-concept co-occurrence matrix constructed using manual annotation. It can be observed that each complex event can be represented as a distinct distribution of mid-level concepts. Similar relationships can be learned for the joint occurrence of various mid-level concepts. In other words, videos depicting the complex event "changing vehicle tire" are expected to contain co-occurring concepts such as "turning wrench" and "opening trunk". The goal of our approach is to leverage the relationships between concepts and event classes to improve event classification. Assuming a uniform prior over all events, a Maximum Likelihood (ML) approach is used to find the class of a video v. In particular,

$$\hat{e}_{ML} = \arg\max_{e \in \mathcal{E}} p\left(v \mid e\right). \tag{1}$$

If all concepts were independent and v is considered as a vector of concept detector indicator functions $\mathbf{C} = [C_1, C_2, ..., C_k]$, where $C_k \in \{0, 1\}$ is the indicator function for the k-th concept, (1) could be simplified as

$$\hat{e}_{ML} = \arg\max_{e \in \mathcal{E}} \prod_{i=1}^{k} p\left(c_i \mid e\right), \tag{2}$$

which is the Naive Bayes classifier [14]. However, in real-world data mining applications, we cannot assume independence relationships between concept-concept or concept-event pairs as they are highly dependent. Also considering all relations make it an NP-hard problem. Hence, our approach leverages a tool from the statistics literature called Averaged One-Dependence Estimator (AODE) [15] to effectively approximate the joint distribution. Herein, AODEs are used to relax the concept independence assumption by aggregating all one-dependence models for an event class. Related techniques were introduced by Wainwright et al. in [16], where a tree reparameterization framework is used for approximate inference over graphs with cycles.

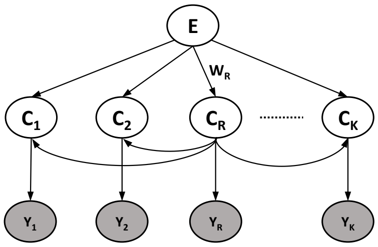

Fig. 2: A special TAN for the root concept C_R for the proposed concept predictions model.

For each concept, a special Tree Augmented Bayesian Network (TAN) in which a given concept is the parent of all other concepts is constructed. The concept is called the root concept C_R. As shown in Figure 2, the root concept is connected to all other concepts. By building this type of TAN for all existing concepts, we have all connections between the concepts in each tree. Then, we average the aggregate of these special TANs. If all TANs are equally weighted, each concept will have the same influence, which clearly is not the case for many real-world applications including complex event analysis. Thus, a natural way to extend AODEs is to assign each root concept different weight. More precisely, each TAN should be weighted differently. Under this assumption, using Weighted-Average One-Dependence Estimator (WAODE), the ML estimate can be written as

$$\hat{e}_{ML} = \arg\max_{e \in \mathcal{E}} \sum_{i=1}^{r} W_i \cdot p(c_i, e) \prod_{\substack{j=1 \\ j \neq i}}^{k} p\left(c_j \mid c_i, e\right), \tag{3}$$

where r is the number of root concepts and W_i is the weight for the TAN rooted at the i^{th} concept. Section 3.2 shows how to calculate r and W_i.

3. INFERENCE OF EVENT LABEL

We are going to infer the label of the event for each video using (3). To this end, we need to estimate the probabilities, the weights and realizations of the concept indicator functions, which are discussed in this section.

3.1. Estimation of Probabilities

We estimate the probability values as

$$p\left(c_i, e\right) = \frac{F\left(c_i, e\right) + \frac{1}{2n}}{n + 1}, \tag{4}$$

$$p\left(c_j \mid c_i, e\right) = \frac{F\left(c_j, c_i, e\right) + \frac{1}{2}}{F\left(c_j, c_i\right) + 1}. \tag{5}$$

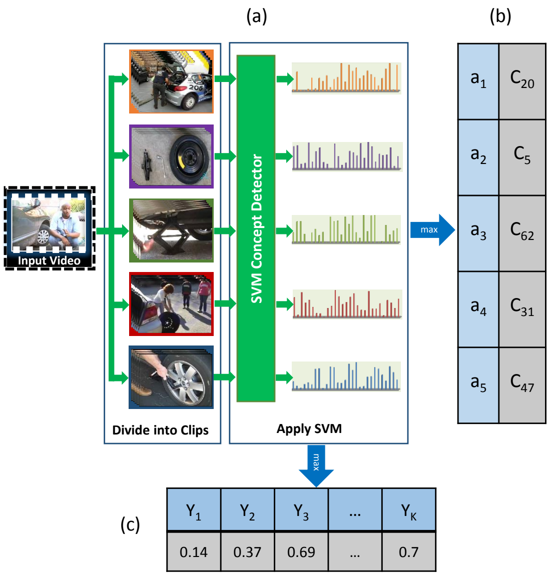

Fig. 3: (a) Result of k indivisual SVM concept detectors, (b) Separate concept with highest score in each clip, and (c) Compute the max score of each concept among all clips.

Here, $F\left(c_i, e\right)$ represents the number of training videos with event label e that contain concept i, $F\left(c_j, c_i, e\right)$ the number of training video with event label e that contain both concepts i and j, and $F\left(c_j, c_i\right)$ the number of training videos in which concepts i and j co-occur regardless of the event label. The factor of 2 in the numerators of (4) and (5) accounts for the number of possible values for each concept indicator function.

3.2. Weights of Root Concepts

In this section, we describe our methodology for computing the weights W_i of the root concepts in (3). Our approach is based on the method of information gains (IG) that was introduced in the context of regression trees [17]. IG is a good measure for deciding the relevance of an attribute.

Our goal is to capture the importance of a particular concept i in event classification. To do so, we set W_i equal to IG$(\mathcal{S}, C_i)$,

$$\mathrm{W}_i = IG(\mathcal{S}, C_i) = H(\mathcal{S}) - \frac{|\mathcal{S}_i|}{|\mathcal{S}|}H(\mathcal{S}_i) - \frac{|\bar{\mathcal{S}}_i|}{|\mathcal{S}|}H(\bar{\mathcal{S}}_i), \quad (6)$$

where $\mathcal{S}$ denotes the set of all training videos, $\mathcal{S}_i \subseteq \mathcal{S}$ the set of training videos for which $C_i = 1$ with complement $\bar{\mathcal{S}}_i$. Hence, $\mathcal{S}_i(e) \subseteq \mathcal{S}_i$ is the set of videos of event e where $C_i = 1$.

Adopting the same notation of [17] for entropy of sets, the entropy $H(\mathcal{S}_i)$ of the set $\mathcal{S}_i$ is defined by summing over all events as

$$H(\mathcal{S}_i) = -\sum_{e=1}^{m} \frac{|\mathcal{S}_i(e)|}{|\mathcal{S}_i|} \log \frac{|\mathcal{S}_i(e)|}{|\mathcal{S}_i|}. \quad (7)$$

The entropy of the sets $\mathcal{S}$ and $\bar{\mathcal{S}}_i$ are defined in a similar way. We note that a concept with larger weight, leads to larger reductions in average uncertainty about the events. Therefore, we retain $r = \frac{1}{2}k$.

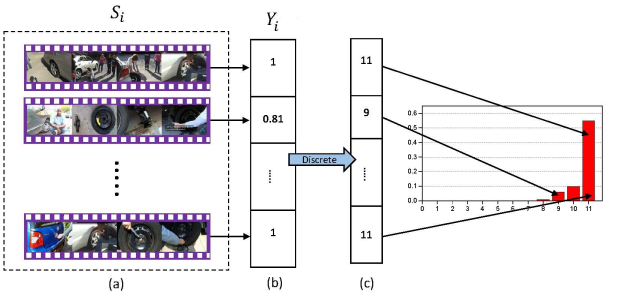

Fig. 4: Visual schematic for constructing the histogram corresponding to $C_i = 1$. (a) Set of training videos for which $C_i = 1$, (b) maximum confidence score of concept i in each video, and (c) quantized scores.

3.3. Inferring Concepts

Unlike the training phase, in testing we do not have access to concept annotations, thus the concepts need to be inferred. One possibility to infer the concept indicator functions is to apply k individual SVM concept detectors to each clip (Figure 3 (a)) and choose the concept with the highest score (Figure 3 (b)). However, the accuracy of this method is fairly low.

The structure we used for inferring concepts is shown in Figure 2. We refer to this method as calibrated concept detection. The concept score measurements $Y_1, Y_2, ..., Y_K$ are shown with shaded circles since they are observed evidence nodes. The value of Y_i (Figure 3 (c)) is equal to the maximum score of the SVM detectors for concept i among all clips. The event label node E and the concept nodes, $C_1, C_2, ..., C_K$, are shown as white circles since their states are to be inferred given the evidence.

3.3.1. Calibrated concept detection

Our method for concept detection combines two sources of information, namely, the score information, $Y_i, i = 1, \dots, K$, from SVM concept detectors and the information from annotations. The main objective is to calibrate our inference based on the ability of the SVM detectors to recognize different concepts. In this method, we first quantize each observed score Y_i to a number of discrete levels (in our experiments we use 11 levels). We modeled the observation nodes using multinomial distributions. Then, we construct two histograms for each concept using results from annotations. The first histogram corresponds to occurrence of a concept, while the second to no occurrence. To clarify, we define the set $\mathcal{S}_{i,j} \subset \mathcal{S}_i$, as the set of training videos for which C_i occurred (based on annotations), and for which the quantization level of Y_i is equal to j. So, $b_i(j)$, the frequency of the j^{th} level in the histogram related to $C_i = 1$, is

$$b_i(j) = \frac{|\mathcal{S}_{i,j}|}{|\mathcal{S}_i|} \tag{8}$$

and $\bar{b}_i(j)$ is defined similarly for $C_i = 0$. Figure 4 illustrates a visual schematic for constructing these histograms.

Having constructed a total of $2k$ histogram, we perform concept detection for each test video using the following rule. First, the SVM score Y_i, is quantized. Assuming the quantized level is equal to j, we decide $C_i = 1$ if and only if $b_i(j) > \bar{b}_i(j)$. In our experiments, this detection approach was shown to alleviate the uncertainly associated with the concept scores obtained by SVM.

4. EXPERIMENTAL RESULTS

Datasets. TRECVID11-MED and TRECVID12-MED are challenging datasets of complex events.We conducted experiments on EC11, EC12 and DEVT datasets. DEVT is a subset of TRECVID11-MED [18] that contains 15 complex events. EC11 and EC12 are subsets of TRECVID11-MED that are partially annotated and marked with 93 concepts. EC12 includes 10 more events from the TRECVID12-MED dataset.

Features. Motion Boundary Histogram (MBH) [1] features were extracted from each clip and a histogram of visual words was constructed for each of these clips.

Quantitative comparison. We compared our proposed approach to the method in [19], as well as to SVM with the RBF kernel using a bag of low level features. Bhattacharya et al. [19] introduced two methods, namely, SSID-S and H-S, to model the temporal relationship among spatiotemporal concepts of a video. While we compared our results to both methods, we only show comparisons to H-S since it has better performance on this dataset.

4.1. Experiments on DEVT

We have trained our model on the EC data set and have evaluated our approach on DEVT. For fair comparison with the method proposed in [19], which can only work for videos containing at least two clips, we do not show the result of videos with event label 9 since these videos have only one clip owing to the fact that we divide each video to clips of longer duration.

We compare the different classification methods based on Accuracy. Figure 5 shows the confusion matrices of event classification with 14 different classes. The Average Accuracy AA of our proposed method is 2.95% higher than H-S [19], and 5.72% higher than SVM.

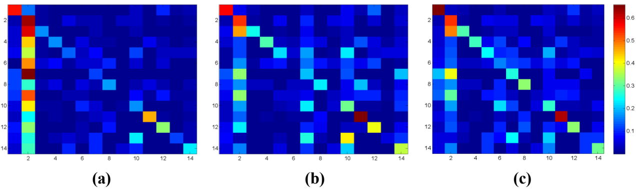

Fig. 5: Confusion matrices of 14 complex events in the DEVT video dataset using (a) SVM event classifiers $AA = 26.59$, (b) H-S [19] $AA = 29.36$ and (c) Proposed method $AA =$ **32.31%**

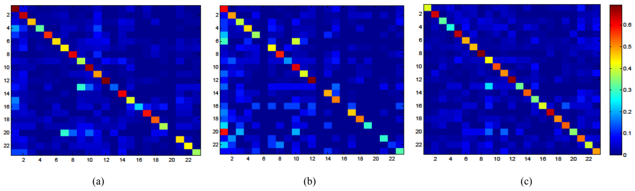

Fig. 6: Confusion matrices of 23 complex events in the EC video dataset (a) SVM method $AA = 46.23$, (b) H-S [19] $AA = 36.06$ and (c) Proposed method $AA =$ **48.5%**

4.2. Experiments on EC

We have trained and tested our proposed method on EC11 and EC12 using 3-fold cross validation. Videos with event labels 9 and 27 are not considered for the same reason explained earlier. The confusion matrices are displayed in Figure 6. The AA for our method is at least 2.27% higher than the other methods.

5. CONCLUSION

We have proposed a novel probabilistic inference framework for complex video event recognition using supervised action concepts. To the best of our knowledge, this is the first principled approach to attempt to model the conditional relationships between complex events and the exhaustive set of intermediate concepts by constraining dependencies to pairwise joint distributions while avoiding the need to manually re-encode new graph structures as the number of concepts increases. Our experiments conclusively demonstrate that this method outperforms conventional, as well as state-of-the-art techniques on multiple challenging data sets of complex event videos.

6. REFERENCES

[1] Heng Wang, Alexander Klaser, Cordelia Schmid, and Cheng-Lin Liu, "Action recognition by dense trajectories," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 3169–3176.

[2] Ivan Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2-3, pp. 107–123, 2005.

[3] Yang Yang and Mubarak Shah, "Complex events detection using data-driven concepts," in *Computer Vision–ECCV*, pp. 722–735. Springer, 2012.

[4] Hui Cheng, Amir Tamrakar, Saad Ali, Qian Yu, Omar Javed, Jingen Liu, Ajay Divakaran, Harpreet S Sawhney, Alex Hauptmann, Mubarak Shah, et al., "Team sri-sarnoffs aurora system@ trecvid 2011," in *Proceedings of NIST TRECVID, Workshop*, 2011, vol. 101.

[5] YG Jiang, X Zeng, G Ye, S Bhattacharya, D Ellis, M Shah, and SF Chang, "Combining multiple modalities, contextual concepts, and temporal matching," in *NIST TRECVID Workshop*, 2010, vol. 1.

[6] Pradeep Natarajan, Prem Natarajan, Vasant Manohar, Shuang Wu, Stavros Tsakalidis, Shiv N Vitaladevuni, Xiaodan Zhuang, Rohit Prasad, Guangnan Ye, Dong Liu, et al., "Bbn viser trecvid 2011 multimedia event detection system," in *NIST TRECVID Workshop*. Citeseer, 2011, vol. 62.

[7] Zhongwen Xu, Yi Yang, I. Tsang, N. Sebe, and A.G. Hauptmann, "Feature weighting via optimal thresholding for video analysis," in *IEEE International Conference on Computer Vision (ICCV)*, Dec 2013, pp. 3440–3447.

[8] P. Natarajan, Shuang Wu, S. Vitaladevuni, Xiaodan Zhuang, S. Tsakalidis, Unsang Park, R. Prasad, and P. Natarajan, "Multimodal feature fusion for robust event detection in web videos," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2012, pp. 1298–1305.

[9] Judea Pearl, "Bayesian networks," *Department of Statistics, UCLA*, 2011.

[10] Mehran Kafai and Bir Bhanu, "Dynamic Bayesian networks for vehicle classification in video," *IEEE Transactions on Industrial Informatics*, vol. 8, no. 1, pp. 100–109, 2012.

[11] Doug Fisher and Hans-J Lenz, *Learning from Data: Artificial Intelligence and Statistics V*, vol. 5, Springer Science & Business Media, 1996.

[12] Tomi Silander and Petri Myllymaki, "A simple approach for finding the globally optimal bayesian network structure," *arXiv preprint arXiv:1206.6875*, 2012.

[13] Amir Tamrakar, Saad Ali, Qian Yu, Jingen Liu, Omar Javed, Ajay Divakaran, Hui Cheng, and Harpreet Sawhney, "Evaluation of low-level features and their combinations for complex event detection in open source videos," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 3681–3688.

[14] Kevin P Murphy, "Naive Bayes classifiers," *University of British Columbia*, 2006.

[15] Geoffrey I Webb, Janice R Boughton, and Zhihai Wang, "Not so naive Bayes: aggregating one-dependence estimators," *Machine learning*, vol. 58, no. 1, pp. 5–24, 2005.

[16] Martin J Wainwright, Tommi S Jaakkola, and Alan S Willsky, "Tree-based reparameterization framework for analysis of sum-product and related algorithms," *IEEE Transactions on Information Theory*, vol. 49, no. 5, pp. 1120–1146, 2003.

[17] Tom M Mitchell, "Machine learning. wcb," 1997.

[18] Paul Over, Jon Fiscus, Greg Sanders, David Joy, Martial Michel, Alan F Smeaton-Alan, and Georges Quénot-Georges, "Trecvid 2013–an overview of the goals, tasks, data, evaluation mechanisms, and metrics," 2014.

[19] Subhabrata Bhattacharya, Mahdi M Kalayeh, Rahul Sukthankar, and Mubarak Shah, "Recognition of complex events: Exploiting temporal dynamics between underlying concepts," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 2243–2250.