A Novel Probabilistic Approach Utilizing Clip Attribute as Hidden Knowledge for Event Recognition

Xiaoyang Wang and Qiang Ji
Dept. of ECSE, Rensselaer Polytechnic Institute, USA
{*wangx16, jiq*}*@rpi.edu*

Abstract

This paper proposes a novel probabilistic approach to utilize clip attributes as hidden knowledge for event recognition. Event recognition in surveillance videos is very challenging due to its large intra-class variations and relative low image resolution. The clip attributes, that are available only during training, provide auxiliary hidden information about the variation of the event appearance. Utilizing such hidden knowledge can help better model the joint probability distribution between event and its observations, and thus improve the recognition performance. We propose a probabilistic model to systematically incorporate the clip attributes into the event recognition. Experiments on real surveillance data show improved event recognition performance with the use of the clip attributes.

1 Introduction

The event recognition for real-world surveillance applications has become an emerging topic [13]. Different from the traditional action recognition tasks with KTH [16] or Weizmann [4] datasets consisting of actions performed by single person in clean backgrounds, or the tasks developed for movies [9] and sports [12], the recognition of complex multi-object interactive events in realistic scene surveillance videos is facing great challenges.

A major challenge for event recognition in these scenarios lies on the large intra class variation. For instance, the event "person getting out of vehicle" may look very different, depending on if the event happens on the side of vehicle facing the camera or happens on the other side of the vehicle away from the camera. Similarly, the event "person unloading a vehicle" may also look different, depending on if the unloading happens from the seats of vehicle or from the trunk. Other realistic factors can also cause the event appearance variation like the variation of target size, illumination change, and shadows, etc.

In order to handle such challenges, we propose to exploit additional properties about video clips to help better model the intra-event variations. For this, we introduce clip attributes that provide further information about the context under which an event happens. Such information is called hidden knowledge [18, 17] since they are only available during training. Hidden knowledge can facilitate the training. Since the proposed clip attributes can affect the image observation of event, they can help better model the joint distribution between event and observations, and thus improve event recognition performance. In this paper, we propose to construct a Bayesian network (BN) [6] to capture the relationships between event label, its observations, and the clip attributes.

The recent work by V. Vapnik in [18, 17] proposed the "SVM+" where hidden information is used to predict the optimal slack variables in the support vector machine (SVM) objective function. In these works, the hidden information is additional features for predicting the class labels. In contrast, the proposed clip attributes capture the factors that contribute to intra-event variations.

Utilizing visual attributes to enhance the object and action recognition has drawn great academic attentions in recent years. In works [2, 8, 14, 20, 15], visual attributes behave as an intermediate representation between low-level image features and high-level categories. In particular, Lampert et al [8] introduce a probabilistic approach to infer the visual attributes, and then classify new categories by zero-shot learning with the inferred visual attributes. Research works in [19, 11, 7] treat visual attributes as latent variables, and formulate the classification problem using a latent SVM framework [3]. Moreover, in [5], a multi-task learning approach is proposed to regularize the models for both visual attributes and the object categories. However, these proposed visual attributes reflect directly the natural properties of the classes/categories and the attributes need to be estimated during testing. For this, different

types of attribute classification models are constructed to estimate attribute values. In contrast, our proposed clip attributes characterize the variation factors for event appearance, and we utilize such clip attributes during training only to help improve event modeling and subsequent recognition.

In summary, our contributions are in two folds: 1) we propose to use clip attributes to capture the factors that contribute to the intra-class event variation ; 2) we develop a BN model to probabilistically incorporate the clip attributes to help improve event modeling.

2 Clip Attributes as Hidden Knowledge

The concept of hidden knowledge was first proposed by [18, 17]. Different from Vapnik's work, we propose to probabilistically capture and utilize the hidden knowledge to benefit event recognition.

2.1 Hidden Knowledge in Probabilistic View

Hidden knowledge, or hidden information [18, 17], is some additional information a given at the training stage about training example x with class label c. Such a information will not be available at the test stage.

From the probabilistic view, the training stage of a classical pattern recognition problem can be described as: given a set of training data pairs

$$(x_1, c_1), \ldots, (x_N, c_N), \quad x_i \in \mathbf{X}, \quad c_i \in \mathbf{C}$$

generated according to a fixed but unknown probabilistic distribution $P(X, C)$ with parameter set ϕ, find an estimate of $\hat{\phi}$ from the training data pairs that can maximize the joint likelihood $P(\mathbf{X}, \mathbf{C}|\hat{\phi})$.

With hidden knowledge, given the training triplets

$$(x_1, a_1, c_1), \ldots, (x_N, a_N, c_N), x_i \in \mathbf{X}, a_i \in \mathbf{A}, c_i \in \mathbf{C}$$

generated from joint probabilistic distribution $P(X, A, C)$ with parameter set θ, we find its estimation $\hat{\theta}$ that can maximize the joint likelihood $P(\mathbf{X}, \mathbf{A}, \mathbf{C}|\hat{\theta})$ instead.

During testing, the goal is to find the class label c^* that

$$c^* = \arg\max_C P(C|X; \hat{\phi})$$

with test example X and estimated parameter $\hat{\phi}$.

With the model learned using hidden knowledge, we marginalize over the hidden knowledge term A, i.e.,

$$c^* = \arg\max_C \sum_A P(C, A|X; \hat{\theta})$$

with test example X and estimated parameter $\hat{\theta}$. The probability $P(C, A|X; \hat{\theta})$ can be readily derived from joint probability $P(X, A, C|\hat{\theta})$.

2.2 Clip Attributes

Clip attribute is an extra description of video clips that is only available during training. Unlike the attributes proposed in [2, 8, 19, 11, 14, 5, 20, 15, 7], clip attributes provide auxiliary information about the context under which an event occurs. Specifically, for this research, clip attributes identify some factors that contribute to intra-class event variation. As shown in Figure 1, the clip attributes we consider include "occlusion by vehicle", "target in shadow", "target at side of vehicle", "target at vehicle tail", and "target size". Each such clip attribute is labeled with a discrete value "1" or "0", where "1" stands for "true" and "0" stands for "false". Depending on an attribute's value, the same event may look very different.

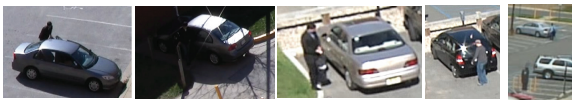
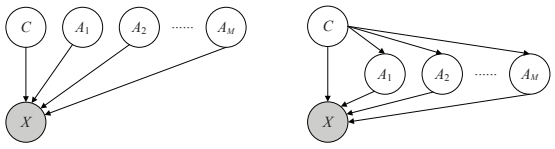


Figure 1: Clip attribute examples. (a) "occlusion by vehicle"; (b) 'target in shadow"; (c) "target at side of vehicle"; (d) "target at vehicle tail"; (e) "target size".

Unlike the features that are extracted on both training and testing data for classification, such clip attributes are only obtained during training. We utilize these clip attributes to help better model the joint probability distribution between event and observations.

3 Probabilistically Modeling Clip Attributes as Hidden Knowledge

Suppose we have M types of clip attributes labeled in total, and denote them as $A_{1\sim M}$ respectively. To capture the clip attributes, we need to model the joint probability $P(X, A_{1\sim M}, C)$, where X is the sample observation, and C is the event label. In the following, we use the Bayesian network (BN) [10]. A BN can efficiently represent a joint probability distribution among a set of variables, where the nodes denote random variables and the links denote the conditional dependencies among variables.



(a) Model I (b) Model II

Figure 2: BNs incorporating attributes.

In our BN models shown in Fig. 2, node C is the discrete root node representing the event label. Node X denotes the observation vector whose elements can be either discrete or continuous. In our implementation,

continuous feature vector is used as observation. The binary nodes $A_1, \ldots, A_M$ describes the corresponding clip attribute $A_{1\sim M}$.

For both Model I and Model II shown in Fig. 2a and Fig. 2b respectively, event nodes C and attribute nodes $A_1, \ldots, A_M$ are the parent nodes of node X. In model II, node C is also the parent node for nodes $A_1, \ldots, A_M$. Comparatively, in Model I, all attribute nodes $A_1, \ldots, A_M$ are root nodes. These two models are formed with different assumptions. For Model I, we assume the attribute is class independent; while for Model II, we assume attributes are class dependent.

For Model I, the joint probability can be factorized as

$$P(X, A_1, \ldots, A_M, C) = P(C)P(A_1) \cdots P(A_M)P(X|C, A_1, \ldots, A_M) \quad (1)$$

through conditional independency assumptions of BNs.

And for Model II, the joint probability distribution is

$$P(X, A_1, \ldots, A_M, A_x, C) = P(C)P(A_1|C) \cdots P(A_M|C)P(X|C, A_1, \ldots, A_M) \quad (2)$$

Suppose the estimated model parameter set is $\hat{\theta}$ (either for Model I or Model II), the model inference for event classification is to find the event label c^*, i.e.,

$$c^* = \arg\max_C \sum_{A_{1\sim M}} P(C, A_1, \ldots, A_M|X; \hat{\theta}) = \arg\max_C \sum_{A_{1\sim M}} \frac{P(X, A_1, \ldots, A_M, C|\hat{\theta})}{\sum_C \sum_{A_{1\sim M}} P(X, A_1, \ldots, A_M, C|\hat{\theta})} \quad (3)$$

where we marginalize over random variables A_1,..., A_M on probability $P(C, A_1, \ldots, A_M|X; \hat{\theta})$ through Equation 3 using the joint probabilities obtained by Equation 1 or 2 for Model I and Model II respectively.

4 Learning of Model Parameters

Given a set of N training samples with observations $\mathbf{X} = \{x_1, \ldots, x_N\}$, class labels $\mathbf{C} = \{c_1, \ldots, c_N\}$, and M clip attributes $\mathbf{A}_m = \{a_{m1}, \ldots, a_{mN}\}$ where $m \in [1, M]$, and supposing these N training samples are i.i.d., we can learn the model parameters for Model I and Model II with the maximum likelihood (ML) estimation. We define the log likelihood as

$$\log P(\mathbf{X}, \mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_M, \mathbf{C}|\theta) = \log \prod_{i=1}^{N} P(x_i, a_{1i}, a_{2i}, \ldots, a_{Mi}, c_i|\theta) \quad (4)$$

The ML would then give the learned parameter $\hat{\theta}$ using Equation 5 with probability normalization constraints for discrete node parameters.

$$\max_\theta \log P(\mathbf{X}, \mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_M, \mathbf{C}|\theta) \quad (5)$$

5 Experiments

We use the VIRAT public 1.0 data set [13] for experiments. It is a realistic natural dataset for video surveillance applications with large intra-class variations. There are six types of events to recognize in this dataset. They are: *Loading a Vehicle* (LAV), *Unloading a Vehicle* (UAV), *Opening a Trunk* (OAT), *Closing a Trunk* (CAT), *Getting into a Vehicle* (GIV), *Getting out of a Vehicle* (GOV) respectively. The VIRAT public dataset is very challenging compared to other event/activity/action datasets like KTH [16], Weizmann [4], HOHA [9] and UCF Sports [12] datasets. It is a natural scene surveillance dataset with low image resolutions on event targets. Also, the VIRAT public dataset focuses on complex events which include the interactions between persons and vehicles. These complex events are more difficult to recognize than the simple events like walking or running.

Currently, the event category labels of the VIRAT public 1.0 testing dataset are not released. Therefore, in our experiments, we use its training dataset with cross validations. There are in total 188 annotated event clips in this training dataset. And we labeled 5 attributes for each clip. These clip attributes are: "occlusion by vehicle", "target in shadow", "target at side of vehicle", "target at vehicle tail", and "target size" respectively.

In our models, we assume Gaussian distributions for continuous observations. In such case, if we do not use any attributes, our models would degrade to the Naive Bayes (NB) model. Thus, we use NB as a baseline model to be compared with. The histogram of oriented gradients (HOG) features [1] extracted from bounding boxes are used as observations.

5.1 Attributes and Their Combination

We first compare the performances for our models incorporating one best or combined attributes. In this experiment, six-category event recognition is performed and each event clip is classified into one of the six categories. The random chance would be 16.77%. We compare the recognition rate of baseline NB, Model I with one best clip attribute "occlusion by vehicle" ("Occlusion"), Model I with combined clip attributes, Model II with one best clip attribute "Occlusion", and Model II with combined clip attributes. Results are show in Fig. 3.

From Fig. 3, we see a single attribute "Occlusion" greatly improves the recognition accuracy. For Model I, incorporating this attribute would improve the performance of NB by over 10%. The Model I with combined clip attributes further improves the result to 36.17%. We observe that Model II also improves results by incorporating attributes. But it performs not as well as

Model I. We think our labeled attributes are generally class-independent, and assume such dependency would cause biased estimations of distributions. However, we believe Model II would be powerful to model class-dependent attributes.

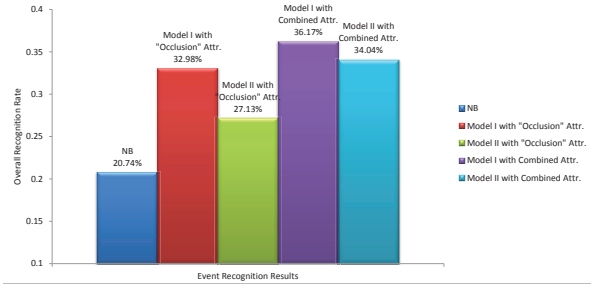


Figure 3: Event recognition accuracy comparison.

For comparison, we also tried the RBF kernel SVM and received 30.32% recognition rate.

5.2 Areas Under ROC Curve

We also compare the event recognition receiver operating characteristic (ROC) curves for each event with different models. In our ROC curve, the vertical axis stands for the event recall rate, and the horizontal axis stands for the event false alarm rate. Due to space limit, we do not present the ROC curve figures but provide the area under ROC curve, which is a more direct evaluation of ROC performance. With a larger area under ROC curve, the model performance is better. Table 1 compares the areas under ROC curve for each event and their average with different models. The combined attributes are used for both Model I and Model II. We can see both Model I and Model II outperform the NB and SVM classifiers on average area under ROC curves.

Table 1: Area under ROC curve comparison

Events	NB	SVM	Model I	Model II
LAV	0.634	0.610	**0.735**	0.675
UAV	0.491	0.614	**0.624**	0.568
OAT	**0.593**	0.574	0.518	0.591
CAT	0.654	0.495	0.617	**0.670**
GIV	0.560	**0.631**	0.539	0.512
GOV	0.595	**0.679**	0.644	0.631
Average	0.588	0.600	**0.613**	0.608

6 Conclusion

In this work, we propose a probabilistic model to incorporate clip attributes as hidden knowledge to improve event recognition from surveillance videos. The events in surveillance videos are usually challenging due to large intra-class variations. The proposed clip attributes can identify factors that contribute to event appearance variations. We propose to exploit such factors to help better model the joint distribution between event and observations. Experiments on realistic surveillance videos demonstrate that the proposed model for incorporating clip attributes as hidden knowledge can effectively improve the event recognition performance.

References

[1] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[2] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009.

[3] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008.

[4] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *IEEE Trans. on PAMI*, 29(12):2247 –2253, 2007.

[5] S. J. Hwang, F. Sha, and K. Grauman. Sharing features between objects and their attributes. In *CVPR*, 2011.

[6] D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT Press, 2009.

[7] A. Kovashka, S. Vijayanarasimhan, and K. Grauman. Actively selecting annotations among objects and attributes. In *ICCV*, 2011.

[8] C. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009.

[9] I. Laptev, M. Marszalek, and etc. Learning realistic human actions from movies. In *CVPR*, 2008.

[10] W. Liao and Q. Ji. Learning bayesian network parameters under incomplete data with domain knowledge. *Pattern Recognition*, 42(11):3046 – 3056, 2009.

[11] J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *CVPR*, 2011.

[12] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos in the wild. In *CVPR*, 2009.

[13] S. Oh, A. Hoogs, and etc. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR*, 2011.

[14] D. Parikh and K. Grauman. Interactively building a discriminative vocabulary of nameable attributes. In *CVPR*, 2011.

[15] D. Parikh and K. Grauman. Relative attributes. In *ICCV*, 2011.

[16] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *ICPR*, 2004.

[17] V. Vapnik and A. Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5-6):544–557, 2009.

[18] V. Vapnik, A. Vashist, and N. Pavlovitch. Learning using hidden information (learning with teacher). In *Inter. Joint Conf. on Neural Networks*, 2009.

[19] Y. Wang and G. Mori. A discriminative latent model of object classes and attributes. In *ECCV*. 2010.

[20] B. Yao, X. Jiang, A. Khosla, A. Lin, L. Guibas, and L. Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *ICCV*, 2011.

[21] Z. Zeng and Q. Ji. Knowledge based activity recognition with dynamic bayesian network. In *ECCV*, 2010.