Modelling Temporal Structures in Video Event Retrieval using an AND-OR Graph

Maaike H.T. de Boer
TNO and Radboud University
The Hague and Nijmegen, The Netherlands
Email: maaike.deboer@tno.nl

Camille Escher
Institut Superieur d'Electronique de Paris
Paris, France
Email: escherCamille@gmail.com

Klamer Schutte
TNO
The Hague, The Netherlands
Email: klamer.schutte@tno.nl

***Abstract*—One of the challenges in Video Event Retrieval, the field in which (a sequence of frames with) high-level events are retrieved from a set of videos, is to model the temporal structure. One way to incorporate this information is using AND-OR graphs, which is a type of graphical model consisting of layers with AND nodes and OR nodes. We introduce new nodes, such as the BEFORE and WHILE node, for AND-OR graphs to explicitly model temporal information. The advantage of these nodes is that the graph is insightful and transparent for a user. Additionally, the graph can both be created by a user or with the use of training examples. We perform initial experiments on a video surveillance dataset named VIRAT, which contains temporally inverse events with the same concepts, such as entering and exiting a building. We compare performance to state of the art Support Vector Machine and Hidden Markov Model methods. We show that our proposed graph with WHILE and BEFORE nodes outperforms the state of the art methods.**

Keywords–AND-OR graph; Temporal Information; Event Retrieval.

I. INTRODUCTION

Nowadays, the most common way to search for a video is to type a textual query in a search engine. Most general search engines, such as Youtube, contain videos with added textual information or *metadata*. In the security domain, this information is often not available. The content of the video should be analyzed to be able to search through those videos. This field of research is named *content-based visual information retrieval*. Within content-based visual information retrieval, we focus on Video Event Retrieval. A complex or high-level event is defined as 'long-term spatially and temporally dynamic object interactions that happen under certain scene settings' [1]. An open challenge in Video Event Retrieval is to model the temporal structure. The difference between videos and images is the temporal structure. It is, however, not directly clear how this temporal structure should be incorporated in image retrieval systems.

Current state of the art methods use Convolutional Neural Networks (CNNs) to train concept detectors [2][3]. Implicitly the temporal structure can be modelled by for example a 3D CNN model [4]. The drawback of the CNN models is that a huge amount of training examples should be available, training of the detectors takes a lot of time and the results are not insightful in why a detector did select a certain action.

Instead of training a neural network for each event, other state of the art methods often use pre-trained concept detectors on images and combine them temporally to represent an event. This combination can be done using some kind of pooling, such as average or max pooling or the more sophisticated Fisher vector or Vector of Locally Aggregated Descriptors (VLAD) pooling [5], and a classifier such as an Support Vector Machine (SVM) [6][7]. This method works well when certain objects or actions are highly indicative for a certain event, but temporally distinctive events, such as the difference between *entering a building* and *exiting a building*, are hard for this type of methods.

Another branch in classification is that of graphical models. Graphical models use probability and graph theory to find structure in sequential data [1]. Examples of such models are Hidden Markov Models (HMM), Conditional Random Fields (CRFs) and AND-OR graphs. The main contribution of this paper is the introduction of BEFORE and WHILE nodes, which support the explicit modelling of the temporal information in an AND-OR graph. The advantage of these nodes is that the graph is insightful and can easily be created by a user or by training examples.

In Section 2, we provide some related work on graphical models in the field of content-based visual information retrieval. Section 3 explains the details of our proposed model with the BEFORE and WHILE nodes. Section 4 contains the experiments on the Video and Image Retrieval and Analysis Tool (VIRAT) 2.0 dataset in which we compare our proposed model with an SVM and HMM model. Section 5 consists of the discussion, conclusion and future work.

II. RELATED WORK ON GRAPHICAL MODELS

The simplest case of graphical models are HMMs. HMMs are often used in human action recognition and event retrieval. An overview is provided by Jiang et al. [1]. For example, Li et al. [8] use salient poses as hidden states to form a model for an action. Tang et al. [9] use a latent structural SVM to learn the feature vectors to feed an HMM. Chen et al. [10] present a framework for video event classification using probabilistic HMM event classification. An advantage of these models is that temporal information can be modelled by these types of models and the models are transparent, but a disadvantage is that causality cannot be modelled and the probability of an event being present is based on a final state. When multiple events have the same end state, these cannot easily be distinguished.

Other types of graphical models are Conditional Random Fields (CRFs) and Dynamic Bayesian Networks (DBNs). Although Vail et al. [11] have shown that CRFs can outperform HMMs in action recognition, these models are disadvantageous in situations where the dependency between events and subevents needs to be modelled [1]. DBNs are a solution to the causality problem of HMMs, but they assume that states are conditionally independent. This makes temporal structure harder to model.

The final type of graphical model is the AND-OR graph. These graphs are often used in the context of grammars. In

 ISBN: 978-1-61208-548-7

our previous work, we have proposed a system with a more extended grammar [12], a stochastic grammar to model a temporal sequence [13] and a grammar model that is robust to noisy inputs [14]. In these grammars, AND-OR graphs can represent hierarchical components by using alternating layers of AND and OR nodes. The AND nodes represent entities that should occur together. An example is the *Part-Of* relation with a person and its parts, such as arms and legs. An example in the event retrieval is the co-occurrence of several objects, such as a car and a person that have to be present at the same time. The OR nodes represent alternative configurations of a certain entity. An example is the skin color, the gender or the type of hair of a person. Each graph has LEAF nodes at the bottom of the graph, which represent the smallest components and one ROOT node, which represents the whole entity that is modelled. Commonly, the AND-OR graphs are formalized by $G = (V, E)$ in which V represents the set of vertices or nodes, and E is the set of undirected edges expressing the relation between two nodes of consecutive layers. During inference, the LEAF nodes are filled with their values. In video retrieval, these values are often binary or a value between zero and one. The values travel bottom up to the ROOT node. The AND nodes take the (normalized) sum of the values and the OR node takes the maximum value of its decendants. The value at the ROOT node represents the score for that modelled entity, such as an event.

Within event retrieval, Tang et al. [15] use the AND-OR graph to fuse multi-modal features. A special type of AND-OR graphs, named Spatial-Temporal AND-OR graphs (ST-AOGs) are previously used to recognize cars [16] and to combine image information with textual information [17]. A very related work is presented by Pei et al. [18]. They use a stochastic context sensitive grammar to present a hierarchical composition of events and temporal relations. They use an AND of temporally related ORs. They represent all events in one model. The disadvantage of this model is that the graph should be re-trained in cases of new events. In general, the advantage of AND-OR graphs is that they are insightful and they can be used on top of grammar models, but a disadvantage is the computational cost of the large number of possible configurations. As a solution structural constraints are often chosen to limit the computational complexity of the learning process.

III. Model Representation

Our model is represented by an undirected graph G, of which an example is shown in Figure 1. We propose a graph that contains a BEFORE node, followed by WHILE nodes. These WHILE nodes are connected to ID(entity) and/or NOT nodes. The LEAF nodes represent the objects at certain time points. This graph can be created by a user that can visualize the query in the graph, or the graph can be created using positive and negative training examples.

A. Inference

To infer whether the event presented by the graph is present in a certain sequence, we use a bottom-up approach to calculate the value at the root node. The value of the leaf nodes is the concept classifier score at a certain time point. The root value can be interpreted as a probability or a score.

The formulas for the nodes are formalized as:

$$v_{ID}(l_{a,t}) = l_{a,t} \tag{1}$$

where $l_{a,t}$ is the value of leaf node a at time point t and v_{ID} is the ID node connected to one of the objects at time t.

$$v_{NOT}(l_{a,t}) = 1 - l_{a,t} \tag{2}$$

where v_{NOT} is the NOT node connected to one of the objects at time t.

$$v_{WHILE}(v_{1,t}, .., v_{m,t}) = \frac{\sum_{i=1,..,m} v_{i,t}}{m} \tag{3}$$

where $v_{1,t}$ to $v_{m,t}$ are the nodes connected to the v_{WHILE} node at time t.

$$v_{BEFORE}(v_1, .., v_n) = \prod_{i=1,..,n} v_i \tag{4}$$

where v_1 to v_n are the nodes connected to the v_{BEFORE} node.

The WHILE node is, thus, an OR node. The BEFORE node is different from the AND node, because the BEFORE node takes the product and the AND node takes one of the values. Although we do not state that the subevents connected by the BEFORE node are independent, our formula equals a joint probability of the subevents assuming independency.

When the length of the test sequence is not comparable to the expected sequence length of the graph, a simple dynamic time warping algorithm is applied. In this algorithm, we delete all redundancies in the consecutive frames and create subsequences of the proper length. These subsequences are all subsequences that can be created with length t, in which t is the amount of time points in the created graph. The highest root node score is used to present the event.

B. Training

In training, we initialize the ID/NOT layer with ID(entity) nodes. The amount of BEFORE nodes is based on the amount of time warped time points of the positive instances for the event. Our current model only has one BEFORE node. The amount of WHILE and ID nodes is the amount of objects that are relevant for this event. Currently, each BEFORE node is connected to two WHILE nodes. Each WHILE node is connected to half the amount of objects in the bag of objects. The ID nodes are randomly pointing to one of the LEAF nodes at their time point.

The graph is trained using the ratio R between the root score of the positive examples ($\overrightarrow{P}_{c,r}$) for class c (in our case an event) on root node r and the negative examples ($\overrightarrow{N}_{c,r}$):

$$R = \frac{1 - ||\overrightarrow{N}_{c,r}||^2 + \overrightarrow{P}_{c,r}}{2} \tag{5}$$

During training, three types of moves are possible:

- Pivot: change the object ($l_{a,t}$) in the bag of objects that the ID/NOT node is pointing to.
- Polarity inversion: replace the ID node by a NOT node or vice versa.

 ISBN: 978-1-61208-548-7

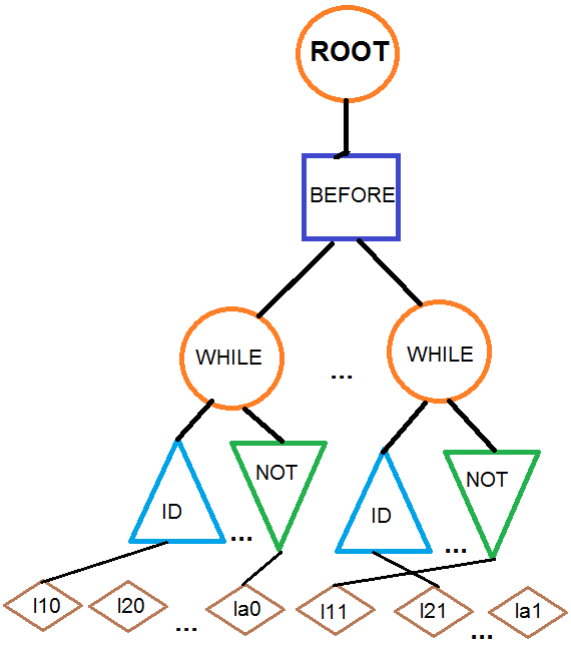

Figure 1. Proposed BEFORE-WHILE Graph model

- Pivot + polarity inversion: first apply a pivot and then a polarity inversion before processing the bottom-up inference.

In a more generalized system, some moves can be added, such as addition of an ID/NOT, WHILE or BEFORE node, as well as removing parts of the graph. During the training process, the following procedure is repeated until convergence of R.

- start with graph G, which has ratio R
- a vertex v is randomly selected in the ID-NOT layer
- one move is randomly selected among the 3 type of moves and new values for $\overrightarrow{P}_{c,v}$ and $\overrightarrow{N}_{c,v}$ are assigned to v.
- the bottom-up inference is applied to propagate the new values to $\overrightarrow{P}_{c,i}$ and $\overrightarrow{N}_{c,i}$ of each node of the graph
- the ratio R is calculated
- if the ratio R' of new G' is higher than R of G, G' becomes the new G, otherwise continue with the old G

IV. Experiments

We use the VIRAT 2.0 dataset [19] to perform our experiments. This dataset contains videos in the surveillance domain with temporally inverse events with the same concepts. These concepts are *person*, *car*, *other vehicle*, *object* and *bike*. The values of these concepts can be represented as (p, c, v, o, b), in which the variables are the values for each of the concepts. Instead of extracting the visual features and applying concept classifiers on the videos, we use the ground truth information of these concepts. For each (predefined) time point, we have binary values in each video for each concept. As explained in the previous section, our method can also handle concept classifier values between zero and one.

We focus on eight events, which can be temporally represented as:

- *person loading an object to a vehicle*: (1, 1, 0, **1**, 0) - (1, 1, 0, **0**, 0)
- *person unloading an object from a vehicle*: (1, 1, 0, **0**, 0) - (1, 1, 0, **1**, 0)
- *person opening a vehicle trunk*: (1, 1, 0, 0, 0) - (1, 1, 0, 0, 0)
- *person closing a vehicle trunk*: (1, 1, 0, 0, 0) - (1, 1, 0, 0, 0)
- *person getting into a vehicle*: (**1**, 1, 0, 0, 0) - (**0**, 1, 0, 0, 0)
- *person getting out of a vehicle*: (**0**, 1, 0, 0, 0) - (**1**, 1, 0, 0, 0)
- *person entering a facility*: (**1**, 0, 0, 0, 0) - (**0**, 0, 0, 0, 0)
- *person exiting a facility*: (**0**, 0, 0, 0, 0) - (**1**, 0, 0, 0, 0)

The bold digit indicates the temporal difference for each event. In two events, which are *person opening a vehicle trunk* and *person closing a vehicle trunk* no difference is present using these five concepts. We, therefore, cannot distinguish these events in this experiment.

For the training, we compared a manually created graph with the trained graph and no difference was found. To distinguish one event from another event (which are the negative training examples), only three concepts are relevant: *person*, *car* and *object*. Each graph consists of one ROOT node, one BEFORE node connected to two WHILE nodes. Each WHILE node is connected to three LEAF nodes, which are the relevant concepts.

We used the standard scene independent process presented by Oh et al. [19], so that the training and testing sets are

 ISBN: 978-1-61208-548-7

composed by videos extracted from multiple scenes. For each video, we calculate the root node score and compare that score to the score for each of the other events. We use the Mean Accuracy, based on the confusion matrix among the events, to report performance. This is the standard approach for this dataset [19] and calculated by taking the amount of correctly classified videos divided by the total amount of videos per class and averaging over all classes / events.

We compare the results of our model with an (RBF) SVM trained on the mean pooled keyframes, an (RBF) SVM with the feature vectors of two time sequences concatenated and an HMM. The results for the methods are shown in Table I.

TABLE I. MEAN ACCURACY SCORES ON VIRAT 2.0 DATASET

Method	Mean Accuracy
SVM_{mean}	0.39
HMM	0.45
SVM_{concat}	0.60
Graph	0.61

The SVM with mean pooling has the lowest performance. This is an expected result, because the temporally opposite events cannot be represented by this type of SVM. Creating a longer feature with the time information increases performance. The HMM has slightly worse results compared to the proposed graph model and the SVM with concatenated time sequences. This is due to the fact that three events have the same end state (11000). These events are, thus, confused using the HMM.

V. Dicussion, Conclusion and Future Work

This work explores a graphical model that makes directly clear which concepts and which relations play a role in a certain event. We propose a model in which BEFORE and WHILE nodes are used as well as ID and NOT nodes. The temporal nodes show the temporal relation and the ID and NOT nodes show which concepts are important and in which polarity (present or not present). Initial experiments on a simple surveillance dataset using ground truth annotations show that our model seems slightly, but not significantly, better than state of the art methods. In future work, it is important to create an improved training process, upgrade the model in a way that it can handle multiple BEFORE nodes and test our model on a difficult dataset with noisy concept detectors. Our model should also be compared to other AND-OR graph based models in the event retrieval field. We, however, provided a solid base for an insightful graph model that can model temporal relations and is transparent, which makes it easy for users to create temporal queries in graphical format.

References

[1] Y.-G. Jiang, S. Bhattacharya, S.-F. Chang, and M. I. Shah, "High-level event recognition in unconstrained videos," Int. J. of Multimedia Information Retrieval, 2012, pp. 1–29.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, 2012, pp. 1097–1105.

[3] Y.-G. Jiang, Z. Wu, J. Wang, X. Xue, and S.-F. Chang, "Exploiting feature and class relationships in video categorization with regularized deep neural networks," in arXiv preprint arXiv:1502.07209, 2015.

[4] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in ICCV. IEEE, 2015, pp. 4489–4497.

[5] Z. Xu, Y. Yang, and A. G. Hauptmann, "A discriminative CNN video representation for event detection," in Proc. of CVPR, 2015, pp. 1798–1807.

[6] H. Zhang et al., "VIREO-TNO @ TRECVID 2015: Multimedia Event Detection," in Proc. of TRECVID 2015, 2015.

[7] Y. Kim, J. Chen, M.-C. Chang, X. Wang, E. M. Provost, and S. Lyu, "Modeling transition patterns between events for temporal human action segmentation and classification," in Int. Conf. on Automatic Face and Gesture Recognition, vol. 1. IEEE, 2015, pp. 1–8.

[8] W. Li, Z. Zhang, and Z. Liu, "Expandable data-driven graphical modeling of human actions based on salient postures," IEEE Transactions on Circuits and Systems for Video Technology, vol. 18, no. 11, 2008, pp. 1499–1510.

[9] K. Tang, L. Fei-Fei, and D. Koller, "Learning latent temporal structure for complex event detection," in Proc. on CVPR. IEEE, 2012, pp. 1250–1257.

[10] H.-S. Chen and W.-J. Tsai, "A framework for video event classification by modeling temporal context of multimodal features using HMM," J. of Visual Communication and Image Representation, vol. 25, no. 2, 2014, pp. 285–295.

[11] D. L. Vail, M. M. Veloso, and J. D. Lafferty, "Conditional random fields for activity recognition," in Proc. Int. Conf. on Autonomous agents and multiagent systems. ACM, 2007, p. 235.

[12] P. Hanckmann, K. Schutte, and G. J. Burghouts, "Automated textual descriptions for a wide range of video events with 48 human actions," in European Conference on Computer Vision. Springer, 2012, pp. 372–380.

[13] G. Sanromà, L. Patino, G. Burghouts, K. Schutte, and J. Ferryman, "A unified approach to the recognition of complex actions from sequences of zone-crossings," Image and Vision Computing, vol. 32, no. 5, 2014, pp. 363–378.

[14] K. Schutte et al., "Long-term behavior understanding based on the expert-based combination of short-term observations in high-resolution cctv," in SPIE, vol. 9995. International Society for Optics and Photonics, 2016.

[15] K. Tang, B. Yao, L. Fei-Fei, and D. Koller, "Combining the right features for complex event recognition," in Proc. of the Int. Conf. on Computer Vision, 2013, pp. 2696–2703.

[16] B. Li, T. Wu, C. Xiong, and S.-C. Zhu, "Recognizing car fluents from video," arXiv preprint arXiv:1603.08067, 2016.

[17] K. Tu, M. Meng, M. W. Lee, T. E. Choe, and S.-C. Zhu, "Joint video and text parsing for understanding events and answering queries," MultiMedia, IEEE, vol. 21, no. 2, 2014, pp. 42–70.

[18] M. Pei, Y. Jia, and S.-C. Zhu, "Parsing video events with goal inference and intent prediction," in ICCV. IEEE, 2011, pp. 487–494.

[19] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J. T. Lee, S. Mukherjee, J. Aggarwal, H. Lee, L. Davis et al., "A large-scale benchmark dataset for event recognition in surveillance video," in Conf. on CVPR. IEEE, 2011, pp. 3153–3160.

 ISBN: 978-1-61208-548-7