COMPLEX EVENT RECOGNITION BY LATENT TEMPORAL MODELS OF CONCEPTS

Ehsan Zare Borzeshi[1], *Afshin Dehghan*[2], *Massimo Piccardi*[1], *and Mubarak Shah*[2]

School of Computing and Communications, University of Technology, Sydney (UTS)[1],
Centre for Research in Computer Vision, University of Central Florida (UCF)[2]

ABSTRACT

Complex event recognition is an expanding research area aiming to recognize entities of high-level semantics in videos. Typical approaches exploit the so-called "bags" of spatio-temporal features such as STIP, ISA and DTF-HOG; yet, more recently, the notion of concept has emerged as an alternative, intermediate representation with greater descriptive power, and "bags of concepts" have been used for recognition. In this paper we argue that concepts in an event tend to articulate over a discernible temporal structure and we exploit a temporal model using the scores of concept detectors as measurements. In addition, we propose several heuristics to improve the initialization of the model's latent states and take advantage of the time-sparsity of the concepts. Experimental results on videos from the challenging TRECVID MED 2012 dataset show that the proposed approach achieves an improvement in average precision of 8.92% over comparable bags of concepts, thus validating the use of temporal structure over concepts for complex event recognition.

1. INTRODUCTION AND RELATED WORK

Recognition of complex events in video is a current research focus with potential application, amongst others, to Web search, multimedia indexing, retrieval and annotation, and real-time monitoring of public premises. In this paper, we refer to events of high semantic complexity such as "renovating a home", "proposing to marry", "meeting at the town hall" and the like. Large samples of these events have increasingly become available to researchers via public repositories such as YouTube and Vimeo or organized collections such as the TRECVID datasets [1].

For recognition, approaches based on *bags of low-level features* such as ISA [2], DTF-HOG [3], STIP [4] and MBH [5], which had already proved effective for recognition of simple actions and gestures, have also proved effective for the recognition of complex events. This result is very important as it shows that, despite their complex nature, many events can be well characterized by features of low-level semantics [6, 7, 8]. However, hierarchical approaches have also become increasingly popular in recent years where more general "concepts" are first identified and then used as atoms for the characterization and recognition of complex events [9, 10, 6, 7]. Concept detectors are typically trained in a supervised manner and can be used to build "bags of concepts" for detecting events of interest in a given video. For instance, in [10] and [9], a large dataset was collected and used to train concept detectors for a task of video annotation. However, such concepts were trained in constrained conditions and are not suitable for general videos. Loui *et al.* in [10] collected a benchmark dataset containing 25 general concepts: however, they are based on static images, not videos. Concepts have also been deployed in the form of attributes [11], which can be considered as concepts with small granularity [6]. [7] used deep learning to find data-driven concepts. Data-driven concepts are an interesting idea and have shown promising performance: however, they are harder to link to a conceptual description of the videos.

Events are occurrences in time and as such they are likely to exhibit some degree of internal dynamics and/or temporal structure. Recently, works such as [12], [13] have demonstrated the importance of temporal structure in complex event recognition. In this paper, we propose to combine the use of trained concept detectors with a latent temporal model. We divide an event video into time slices and use the scores of concept detectors as measurements in a hidden conditional random field (HCRF) [14], learning its parameters with latent structural large margin [15]. The hidden state chain in the HCRF allows joint decoding of all the concepts in the event and forms the basis for event recognition. Moreover, moving from the empirical observation that concepts in an event may be sparse in time, we enforce an equivalent sparsity in the latent states. The main contributions of this paper are: i) using concept detector scores as measurements for a latent temporal model with the aim of leveraging on both trained concept detectors and the properties of latent structural models; ii) enforcing sparsity in the decoded chain of latent states in order to mirror the time-sparse distribution of concepts in an event, iii) exploring various state initializations to improve the quality of the latent large margin solution, and iv) providing a comparative evaluation against several types of bag-of-features including various low-level features, concepts, and combinations of low-level features and concepts.

As dataset, we have utilized the NIST's TRECVID MED 2012 toolkit dataset [1] that is very probing in terms of event complexity [16]. The experimental results presented later in the paper show that the combined use of concept detectors and latent temporal models significantly improves recognition performance at a parity of features and concepts.

2. LATENT STRUCTURAL SVM FOR HIDDEN CONDITIONAL RANDOM FIELDS

In this section, we refer to the graphical model used in this work as hidden conditional random field (HCRF) even though we approach its learning by a maximum-margin method. The graphical model is displayed in Fig. 1. The learning objective for training the HCRF with maximum margin is defined as:

$$\begin{aligned} &\underset{w,\xi_i}{\operatorname{argmin}} \left(\|w\|^2 + C\sum_{i=1}^{N}\xi_i \right) \quad s.t. \\ &w^T\Psi(y_i, h^*_{1:T_i}, x_{1:T_i}) - w^T\Psi(y, h_{1:T_i}, x_{1:T_i}) \\ &\qquad \geq 1-\xi_i \qquad \forall\{y,h_{1:T_i}\} \neq \{y_i, h^*_{1:T_i}\} \end{aligned} \qquad (1)$$

where y is an event label, y_i is the ground-truth label of event sample i, $x_{1:T_i}$ is its sequence of measurements and $h_{1:T_i}$ is an assignment for its latent states. The event label is a binary variable taking value 1 for the given event and 0 otherwise. Each latent state, $h_t, t = 1 \ldots T_i$, takes values over a discrete range of indices, $\{1 \ldots H\}$, representing the internal dynamical state of the HCRF. Each measurement, $x_t, t = 1 \ldots T_i$, is an F-dimensional feature vector extracted from the image (in our case, it is the output of $F = 93$ concept detectors). The parameter vector, w, contains three types of parameters, or weights: i) *transition weights*, w^{tr}, scoring the transitions between consecutive states, indexed by the current and previous state values; ii) *emission weights*, w^{em}, indexed by the current state value and the index of the dimension in the measurement; and iii) *compatibility weights*, w^{cmp}, indexed by the current state value and the event value (positive or negative class).

Notation $w^T\Psi(y, h_{1:T}, x_{1:T})$ is a compound notation for the HCRF score:

$$\begin{aligned} w^T\Psi(y, h_{1:T}, x_{1:T}) = \sum_{t=2}^{T} w^{tr}_{ij}\delta\left[h_{t-1} = i, h_t = j\right] + \\ + \sum_{t=1}^{T}\sum_{f=1}^{F} w^{em}_{if} x_{tf}\delta\left[h_t = i\right] + \sum_{t=1}^{T} w^{cmp}_{ib}\delta\left[h_t = i, y = b\right] \end{aligned} \qquad (2)$$

Given that the states are unsupervised in the training set, their best assignment for sample i must be inferred as

$$h^*_{1:T_i} = \underset{h_{1:T_i}}{\operatorname{argmax}}\, w^T\Psi(y_i, h_{1:T_i}, x_{1:T_i}) \qquad (3)$$

This problem can be resolved by an appropriately weighted Viterbi decoder in $O(T)$ time and the solution replaced in the constraints in (1) as estimated ground truth. Variable ξ_i is the slack variable for sample i, allowed to take non-negative values so as to let the inequality constraints be violated. The sum of the slack variables over the training set, $\sum_i^N \xi_i$, is an upper bound over the total classification error [17]. One can then see that the objective function in (1) pursues a minimization of the empirical error, while regularizing the solution by enforcing the largest possible class margin. Learning of the HCRF is obtained by alternating the solution of (1) and (3) until convergence.

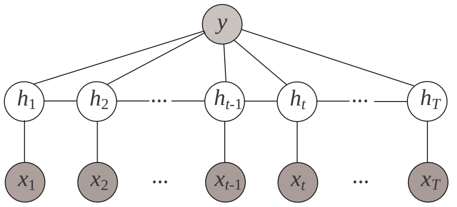

Fig. 1. The graphical model of the hidden conditional random field.

Due to the exponential number of possible combinations of $y, h_{1:T_i}$ in (1), exhaustive verification of the constraints would not be feasible. However, [17] and [15] have shown that it is possible to find ϵ-correct solutions in polynomial time by using only the "most violated" constraints, i.e. the configuration of class and states with the highest sum of score and loss:

$$\bar{y}_i, \bar{h}_{1:T_i} = \underset{y, h_{1:T_i} \neq y_i, h^*_{1:T_i}}{\operatorname{argmax}} \left(w^T\Psi(y, h_{1:T_i}, x_{1:T_i}) + 1 \right) \qquad (4)$$

For the HCRF detector, such a configuration can still be efficiently determined in $O(T)$ time by a 2-best Viterbi decoder.

2.1. Latent State Initialization

Due to the presence of the latent variables, learning the HCRF is overall a non-convex problem, whereas the solution of (1) is convex in isolation. Learning can be initialized by either an arbitrary vector w in (3) or an arbitrary $h^*_{1:T_i,i}$ state sequence in (1). Choosing a state sequence could be preferable since it is more confined than selecting a continuous vector, yet learning proves very sensitive to the states' initialization. [18] uses the states returned by an equivalent graphical model trained generatively by expectation-maximization (EM). However, EM requires an arbitrary initialization at its turn. In [12], the initial states are first assigned with a unique label, and then the number of labels is reduced by agglomerative clustering. In this work, we propose initialization strategies inspired by the assumed semantics for the states:

1. *Non-informative assignment (NInf)*: the initial states of each positive sample are all assigned with label 1, while those of negative samples are all assigned with label 2.

2. *Non-informative assignment with overlapping state (NInfOv)*: the initial states of each positive sample are assigned with alternate labels 1 and 2 every other frame. The states of the negative samples are assigned with alternate labels 2 and 3 likewise. This is to enforce an overlapping state across the two classes.

3. *Asymmetric assignment (Asymm)*: given that the negative class is expected to be more spread out (from being the combination of many classes), its states are assigned randomly over a small range of integers, $\{2 \dots H\}$.

4. *Asymmetric assignment with neutral state (Sparsity)*: this assignment is similar to the previous, with the addition of a further state meant to represent "no concept". This neutral state is not included in any initial assignments, rather only reserved in anticipation of the learning stage.

2.2. Time-Sparsity of Concepts

Fig. 2, top, shows the output of 93 concept detectors for a "Dog show" event: most detectors never activate significantly during the sequence (we use a threshold of 0.4 for visualization), and the few that do typically activate for only a few frames at a time. Fig. 2, bottom, shows a corresponding state trellis: state 1 is the "no concept" state, and state 2 activates in loose correspondence with the highest responses from the detectors. This behavior supports the idea that the number of utilized concepts per event is relatively small, and that they tend to be time-sparse. To leverage this property, we chose to encourage sparsity in the decoded state sequence of the HCRF by favoring transitions towards the neutral state. We obtain this by multiplying the weights for the transitions towards the neutral state by a positive coefficient, S, (as $S * w_{1j}^{tr}$) during the computation of both (3) and (4).

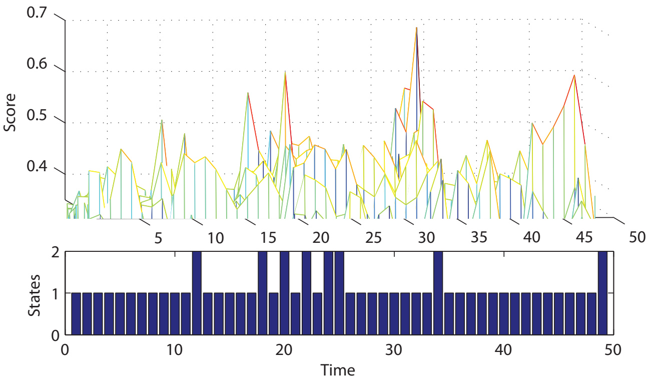

Fig. 2. Time-sparsity of concepts and corresponding states.

3. EXPERIMENTS

We experimented the proposed method on a subset (10 events) of TRECVID MED 2012 event collection multimedia dataset (EC12 hereafter). Its events are challenging due to their heavily variable duration (ranging from 30 seconds to 30 minutes), frame rate (from 12 to 30 fps), and resolution (from 320 to 1280). The dataset used is consisting of 2000 videos, that we have split as 70% for training and 30% for testing. Both the concept detectors and the HCRF have been trained on the training set alone, and the test videos have been used blindly for testing without any further adjustment of the parameters. As evaluation metric, we have adopted the *average precision* which is an average of the precision at various levels of recall (equivalent to the area under the precision-recall curve) [8, 6]. A total number of 93 concepts were annotated over a portion of the training data. These concepts were selected based on the description in the TRECVID competition kit and by viewing sample videos. For each concept, an SVM model was trained using STIP as features for detection [4]. In order to compute the concepts' scores in a video, we first divide it into overlapping clips, with a clip length of 180 frames and a step size of 60 frames. Subsequently, the score of each detector is computed for every clip in the video leading to an intermediate representation as a multivariate time series. For training our model, we have set the sparsifying coefficient, S, to vary over $[2, 5000]$ in logarithmic steps; ϵ was set over $[10^{-2}, 10^{-6}]$; the number of states, H, was made vary between 3 and 15; C was set to 100; and the linear kernel used as kernel.

We compare our approach with the following methods:

- Bag-of-Concepts (BoConcept): we first apply max-pooling on the time series representation of each video, leading to a 93-dimensional vector containing the maximum score of each concept detector in that video. We use an SVM directly on such obtained high-level features, and we refer to this setup in the tables as BoConcept.

- Bag-of-Words (BoW): in this case, we cluster various low-level features (STIP, ISA, and DTF-HOG) to obtain a dictionary. Subsequently, we compute a histogram of word frequency for each feature. We use a codebook size of 10000 for all the features, and min-max normalization for the histograms. We refer to this approach in the tables by the name of the features used in the BoW framework (i.e STIP, ISA, and DTF-HOG).

- Combinations of the various low-level features, and of features and concepts: we use early fusion to combine a) all the low-level features (All-LL); b) STIPs and concepts (since the concepts were trained over STIPs; referring to this combination as STIP + Concepts) and c) all low-level features and concepts (All-LL + Concepts). When fusing low-level features and concepts, we preprocess the low-level features with PCA to reduce their dimensionality from 10000 to 200 to make it comparable with that of the Bag-of-Concepts features (93).

The TRECVID MED 2012 event collection consists of the following 10 complex events: *Bike trick (BiT), Cleaning appliance (CA), Dog show (DS), Giving direction (GD), Marriage proposal (MaP), Renovating a home (RH), Rock climbing (RC), Town hall meeting (TM), Race winning (RW)*, and *Metal craft project (MeP)*. The performance results for the EC12 dataset are reported in Table 1 as average precision for each class and overall mean value. In the first four columns

Table 1. The average precision for the EC12 dataset using both concepts and low-level features.

Event	STIP	BoConcept	STIP+Concepts	**Ours**	ISA	DTF-HOG	All-LL	All-LL+Concepts
BiT	61.78	69.59	67.83	*70.68*	66.48	65.09	72.94	74.22
CA	69.68	67.27	71.27	*73.76*	62.95	64.22	71.11	74.15
DS	47.88	60.09	62.36	*68.18*	62.25	66.87	66.72	68.80
GD	54.27	56.83	48.31	*75.05*	58.77	66.26	62.48	60.98
MaP	77.47	66.61	73.76	*73.86*	65.97	64.74	79.28	79.56
RH	73.06	57.48	68.03	*68.81*	76.27	66.86	73.74	72.70
RC	65.60	65.41	72.83	*76.13*	72.09	80.99	75.41	79.60
TM	69.06	60.16	72.09	*74.36*	69.20	76.90	67.48	71.95
RW	74.90	72.42	77.54	*79.65*	76.97	71.64	75.22	75.74
MeP	81.58	73.09	82.09	*77.58*	65.68	67.82	69.80	79.35
Mean	67.53	64.89	69.61	*73.81*	67.66	69.14	71.42	73.69

Table 2. Comparing initializations for EC12.

Event	NInf	NInfOv	Asymm	Sparsity
BiT	63.25	67.72	**70.68**	70.39
CA	70.91	59.91	71.83	**73.76**
DS	59.74	62.56	58.41	**68.18**
GD	65.51	**75.05**	71.59	72.30
MaP	55.29	68.23	65.45	**73.86**
RH	65.68	65.50	67.44	**68.81**
RC	73.80	67.78	74.73	**76.13**
TM	66.91	72.94	69.37	**74.36**
RW	69.79	69.80	**75.36**	70.12
MeP	71.23	74.07	74.97	**79.65**
Mean	66.22	68.36	69.99	**72.76**

we report the performance of all methods that use STIP as low-level feature. Comparing mean values, one can see that the proposed method reports a remarkable improvement of 8.92% over Bag-of-Concepts, of 6.28% over STIP, and of 4.20% over the fusion of STIP and concepts. This result gives evidence to the benefit of exploiting temporal structure over the concept detector scores. The remaining columns in Table 1 show the performance of the other single low-level features (ISA and DTF-HOG) and the fusion methods. DTF-HOG proves the best single low-level feature. The proposed method outperforms all single features, their fusion (All-LL), and even achieves a mean precision slightly higher than that of the fusion of all low-level features and concepts (All-LL + Concepts; 73.81% vs. 73.69%).

Table 2 shows a comparison of the average precision obtained with the different state initialization methods. For most classes (7 out of 10) and on average, the *Sparsity* approach outperforms the other initializations. This results gives evidence that enforcing sparsity during state decoding is generally beneficial. In addition, Table 2 shows that the different initializations have a major impact on performance.

4. CONCLUSION

In this paper, we have presented an approach to complex event recognition combining a latent temporal model and trained concept detectors. Since learning the temporal model proves heavily sensitive to state initialization, we have proposed several heuristics for effective initialization. In addition, we have suggested exploiting the time-sparsity of the concept detector scores by a corresponding sparsity in the decoded states. Experimental results over the challenging TRECVID MED 2012 Event Kit Collection show that the mean average precision of the proposed method is 8.92% higher than that of a Bag-of-Concepts methods using the same concepts. In addition, it is 4.20% higher than that of the best method using the same low-level features (STIP), and even higher than that achieved by combining various low-level features and concept scores. These results give strong evidence to the benefit of exploiting temporal structure over the concepts and to the effectiveness of the proposed approach.

5. ACKNOWLEDGMENT

This work was partially supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20066. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S.Government.

6. REFERENCES

[1] "TRECVID multimedia event detection track. http://www-nlpir.nist.gov/projects/tv2011/tv2012.html," 2012.

[2] Q.V. Le, W.Y. Zou, S.Y. Yeung, and A.Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *CVPR*, 2011.

[3] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *CVPR*, 2011.

[4] I. Laptev and T. Lindeberg, "Space-time interest points.," in *ICCV*, 2003.

[5] H. Wang, A. Klaser, C. Schmid, and C. L. Liu, "Action recognition by dense trajectories," in *CVPR*, 2011.

[6] H. Izadinia and M. Shah, "Recognizing complex events using large margin joint low-level event model," *ECCV*, 2012.

[7] Y. Yang and M. Shah, "Complex events detection using data-driven concepts," *ECCV*, 2012.

[8] A. Tamrakar, S. Ali, Q. Yu, J. Liu, O.Javed, A. Divakaran, H. Cheng, and H. Sawhney, "Evaluation of low-level features and their combinations for complex event detection in open source videos," in *CVPR*, 2012.

[9] X. Liu and B. Huet, "Automatic concept detector refinement for large-scale video semantic annotation," in *IEEE Fourth International Conference on semantic computing*, 2010.

[10] A. Loui, J. Luo, S. Chang, D. Ellis, W. Jiang, L. Kennedy, K. Lee, and A. Yanagawa, "Kodak's consumer video benchmark data set: concept definition and annotation," in *Multimedia Information Retrieval*, 2007.

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

[12] Kevin Tang, Li Fei-Fei, and Daphne Koller, "Learning latent temporal structure for complex event detection," in *CVPR*, 2012.

[13] Weixin Li, Qian Yu, Harpreet Sawhney, and Nuno Vasconcelos, "Recognizing activities via bag of words for attribute dynamics," in *CVPR*, 2013.

[14] A. Quattoni, S. Wang, L. Morency, M. Collins, and T. Darrell, "Hidden conditional random fields," *PAMI*, 2007.

[15] Chun nam Yu and Thorsten Joachims, "Learning structural SVMs with latent variables," in *ICML*, 2009.

[16] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *IJCV*, 2010.

[17] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," *JMLR*, 2005.

[18] S.N. Parizi, J.G. Oberlin, and P.F. Felzenszwalb, "Reconfigurable models for scene recognition," in *CVPR*, 2012.