VISUAL EVENT DETECTION USING MULTI-DIMENSIONAL CONCEPT DYNAMICS

Shahram Ebadollahi[*], *Lexing Xie*[*], *Shih-Fu Chang*[†*] *John R. Smith*[*]

[*]IBM T. J. Watson Research Center, Hawthone, NY
[†] Dept. of Electrical Engineering, Columbia University, New York, NY

ABSTRACT

A novel framework is introduced for visual event detection. Visual events are viewed as stochastic temporal processes in the semantic concept space. In this concept-centered approach to visual event modeling, the dynamic pattern of an event is modeled through the collective evolution patterns of the individual semantic concepts in the course of the visual event. Video clips containing different events are classified by employing information about how well their dynamics in the direction of each semantic concept matches those of a given event. Results indicate that such a data-driven statistical approach is in fact effective in detecting different visual events such as *exiting car*, *riot*, and *airplane flying*.

1. INTRODUCTION

Providing semantic access to video repositories has always been a major goal for the multimedia community. In recent years, a good amount of effort has been put into methods for modeling the visual semantic concepts, i.e. general categories of objects, scene, their coexistence and interactions. Acceptable results have been achieved for the case where enough annotated training data exist for concepts in a lexicon, as evidenced by the annual TRECVID [1] benchmark. However, the majority of the concepts that have been reported are of *static* nature, such as *indoors*, *outdoors*, *greenery*, *etc*. For events, or concepts that are distinct in the action of objects and the evolving interaction among objects and the scene, such as *"airplane takeoff"* and *"riot"* , automatic detection still remains a challenging problem.

The prior literature on the problem of visual event detection could be divided into two main categories. The first category, which we refer to as *object-centered*, regards an event as a spatial, temporal, and logical interaction of multiple objects (agents, actors). The primary focus of the works in this category is to track the objects and analyze their activity patterns [2, 3, 4]. The object-centered approach has a decomposition view of the events [4] in space-time and tries to extract constituent elements of an event and analyze their characteristics. This approach, which is rooted in computer vision, although has been successfully applied to certain problems, for example in surveillance applications, quickly becomes infeasible for videos with unconstrained content and minimally controlled context, such as news footages. The works in the second category, on the other hand, have less of a computer vision flavor and analyze an event from a pure statistical point of view. In [5], for example, statistical models of feature dynamics were learned for audio and video channels, and their combination was used to detect events such as *explosion*. Xie *et al.* [6] detected and segmented the *play* and *break* events in soccer videos by learning the dynamics of the color and motion features for each event. However, these approaches rely directly on low-level features, which are often not as intuitive as other event components, such as objects or visual concepts.

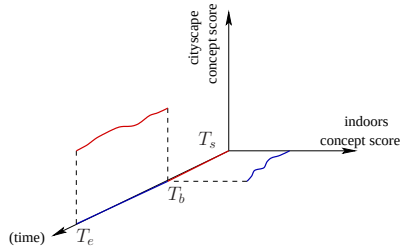

Fig. 1. Evolution patterns of concepts *indoors* and *cityscape* in the event *"exiting building"*. T_s =start of event, T_b =switch from indoors to outdoors, where *cityscape* concept will become significant, T_e =end of event.

We propose a novel approach to the problem of event modeling and detection. Events in our approach are regarded as stochastic temporal processes in the semantic concept space [7]. An available pool of semantic concept detectors form the basis of this space. Each concept detector provides its view of the world as depicted in a video clip.

The central assumption in our approach is that during the progression of a visual event, several concurrent concepts evolve in a pattern specific to that event. Figure 1 illustrates this idea in its simplified form. During an event such as *exiting a building*, one expects to observe the concept *indoors* in the initial stage of the event and then as the event progresses switch to the concept *cityscape* and stay in that state for some time (Fig. 1). We submit such a framework is powerful and can be used to model a large number of events, as will be confirmed in our experiments later (Section 3).

[*]Work performed while visiting IBM T.J. Watson Research Center.

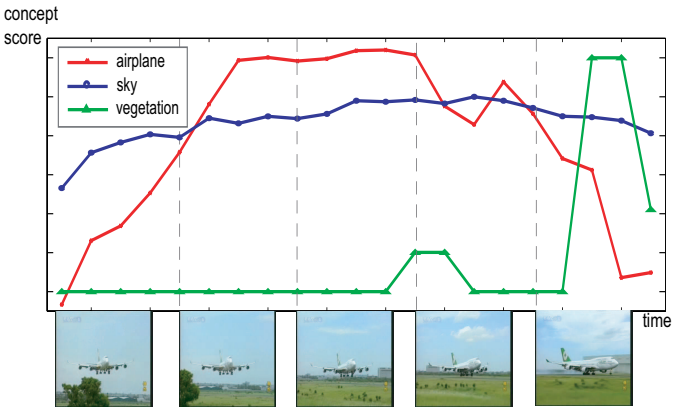

Fig. 2. Example of projecting a video segment of the event airplane landing onto a three-dimensional concept space (airplane, sky, vegetation)

This approach, which is concept-centered as opposed to object-centered, aims at learning the dynamics of concurrent concepts from exemplars of an event in a pure data-driven fashion. In this work, we develop novel use of this space by modeling the temporal dynamics within. The concept-space used in the present paper, was formed by 39 LSCOM-lite [8] concept detectors, which were obtained by training Support Vector Machine (SVM) classifiers [9] over visual features such as grid local color, texture, motion, and edge. The proposed approach is tested on several different event categories. Results indicate the effectiveness of the concept-centered approach for certain events, such as *riot*, *helicopter hovering*, with a clear performance gain from 28% to 58%.

In the rest of this paper, we provide the details of our approach and modeling scheme, followed by the experiments and discussions.

2. MODELING EVENTS IN THE SEMANTIC CONCEPT-SPACE

2.1. Semantic Concepts and Events

The proposed method in this paper relies on the availability of a pool of semantic concept detectors. We employ the 39 semantic concepts of LSCOM-Lite [8], which are the interim results of the effort in developing a Large-Scale Concept Ontology for Multimedia (LSCOM) [10].

The concepts were selected based on semi-automatic mapping of 26377 noun search terms from BBC query logs in late 1998 to Wordnet senses, division of semantic concept space into a small number of orthogonal dimensions, and evaluation of 2003 and 2004 TRECVID search topics [1]. The dimensions consist of program category, setting/scene/site, people, object, activity, event, and graphics. A collaborative effort was completed in 2005 to produce annotations of the 39 concepts over the entire development set of TRECVID 2005 videos. Human subjects judge the presence or absence of each concept in the keyframe of each shot. Statistical concept detectors were built for each of the 39 LSCOM-Lite concepts from the available news video corpus distributed by NIST.

The events targeted for modeling in this work were selected from the LSCOM lexicon. LSCOM is an ARDA sponsored [10] effort for developing an expanded multimedia concept lexicon on the order of 1000.

2.2. Modeling

Let's assume that a concept detector set $\Delta = \{\delta_1, \ldots, \delta_N\}$ is available. Let $\mathcal{V}$ be the collection of videos, where each video $V \in \mathcal{V}$ is represented as a sequence of its frames $V = \{f_1, \ldots, f_T\}$.

Applying each available concept detector to the frames of a video clip results in an array of semantic concept confidence scores $C^n_{1:T} = \{c^n_1, \ldots, c^n_T\}$, where $c^n_t = \delta_n(f_t)$ is the score assigned to the $t-th$ frame by the $n-th$ concept detector. The matrix $\Phi(V) = \Phi(.) \otimes V = [C^1_{1:T} \left| C^2_{1:T} \right| \ldots \left| C^N_{1:T} \right]$ maps the video clip V into a trajectory $\alpha(t)$ in the semantic concept space $(\mathcal{C})$as depicted in figure 2, through the projection operator $\Phi(.) = [\delta_1(.), \ldots, \delta_N(.)]'$.

After projecting the video clips into $(\mathcal{C})$ we model the evolution pattern of the concepts in this space. We assume that the concept detectors are independent from each other, *i.e.* they have been independently trained, possibly using different data sets. The basis for this assumption is because these scores can be independently obtained (using separate sources of data, labeling, classifier etc.), rather than asserting the statistical independence of the score values on any dataset. Due to the independence assumption, we can decompose the trajectory $\alpha(t)$ in $\mathcal{C}$ into its projections onto the N semantic concept axis.

We then proceed to model the evolution pattern of $\alpha_n(t)$, which is the shadow of the trajectory $\alpha(t)$ on the $n-th$ concept axis. A Hidden Markov Model (HMM) [11] with the structure as shown in figure 3 is used to model the pattern of the dynamics of the concept score on each axis. One of the states captures the *"on time state"* of the concept during the event and the other state captures its *"off time state"*.

The application of the modeling scheme to all concurrent concept threads is depicted in figure 3, where the HMM model has been unrolled for each of the threads. As shown in this figure, we do not take into account the interdependency between the hidden states of the different concept threads, which can increase the number of the parameters of the model exponentially.

The multi-thread model of the event is learned from a set of previously annotated exemplars of the event of interest. After training the thread models, a set of test sequences $V_{\text{validation}} = \{V_1, \ldots, V_K\}$ are passed through the semantic projection operator $\Phi(.)$ to obtain their corresponding trajectories in the concept-space $\{\alpha_1(t), \ldots, \alpha_K(t)\}$. Each trajectory is then passed through the array of HMM models, and is assigned a score by them. Three different types of scores are tried in this paper, 1-*Log-Likelihood (LL)*, 2-*State Histogram (SH)*, 3-*Fisher Score (FS)* [12]. *Log-likelihood* is a natural

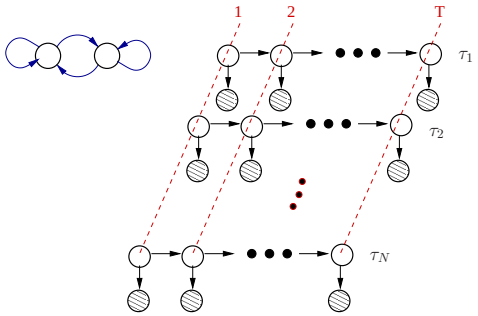

Fig. 3. On the right-hand side the concurrent thread models is shown. The top level for each thread τ_n is the sequence of hidden states, and the bottom row is the sequence of length T of observation for that thread. The 2-state HMM used for each thread is shown on the top left corner.

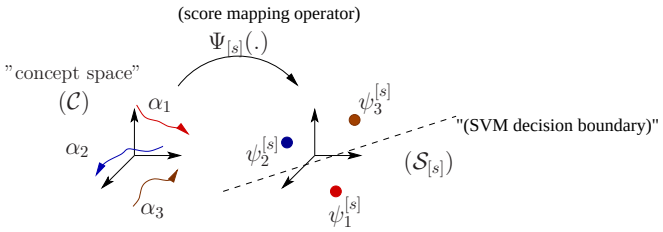

Fig. 4. Mapping activities from concept-space to score-space and finding the discriminating hyper-plane between instances of an event and negative examples.

scoring of a sequence with respect to a HMM. *State histogram* is employed to capture the *on-off* pattern of concepts during an event and if any of the two states in each thread is dominant. Fisher score [12], is used to extract features from a sequence with respect to a generative model. It is a mapping from the observation sequence to the gradient space of the generative model. The partial derivative of the generative model with respect to each parameter provides a description of the way that particular parameter contributes to the process of generating an observed sequence, therefore is an indicator for how well the observed sequence could be generated by the model.

Based on the type of score used, trajectory $\alpha_k(t)$ gets mapped to point $\psi_k^{[s]}$ in score-space $\mathcal{S}_{[s]}$ by score operator $\Psi_{[s]}(.)$, where $s \in \{LL, SH, FS\}$ determines the type of score. The score operators are the following:

$$\begin{aligned}\Psi_{[LL]}(.) &= (log(P(.|H_{\tau_n})),\\ \Psi_{[SH]}(.) &= ([h_j(.|H_{\tau_n}]),\\ \Psi_{[FS]}(.) &= ([\nabla_\theta log(P(.|H_{\tau_n})])\end{aligned} \quad (1)$$

, where H_{τ_n} is the HMM model for thread τ_n (Figure 3), (θ) is the set of parameters of the HMM model, and h_j is the fraction of the length of input sequence that it spends in state j.

After projection of the input sequences into the score-space of the multi-thread model, a discriminant classifier is used in this space for classification of sequences into different event categories. SVM classifiers are used for the categorization of sequences in the concept-space due to their well proved discrimination performance. The SVM classifier essentially fuses the output scores of the multi-thread HMM models.

3. EXPERIMENTS AND RESULTS

A set of seven different events was selected for the experiments, based on the number of sequences available for each event and the length of the sequences. The events are: *airplane flying*, *exiting car*, *ground combat*, *handshaking*, *helicopter hovering*, *parade*, and *riot*, all part of the LSCOM lexicon [10]. The minimum duration of the sequences for all seven events is 5 I-Frames long and the number of sequences available for each is more than 52. We assume that an event is manifested in a single shot. Therefore, we are not concerned about cross-shot dynamics in this paper.

Available sequences for each event were decomposed into 40%, 40%, 20% ratios for the experiments. The I-Frames of each sequence (2.5 I frames per second on average) in the data set were passed through the available 39 LSCOM-lite concept detectors [1] to form the basis of the semantic concept space $(\mathcal{C})$ in our experiments. These detectors are SVM classifiers and were previously trained from a news video corpus, using raw color and texture features. Concept detectors map all sequences to the semantic concept space $(\mathcal{C})$.

For each event $m = \{1, \ldots, M\}$, using the first portion of the data, an array of HMM models were trained, one per concept thread. The second 40% of the sequence from all events were then evaluated by the models for event m (Figure 3) and mapped to three different score-spaces $\mathcal{S}_{[LL]}, \mathcal{S}_{[SH]}, \mathcal{S}_{[FS]}$ (Figure 4). In each space a linear SVM classifier was trained, which resulted in three different classifiers for event m: $\sigma^m_{[LL]}$, $\sigma^m_{[SH]}$, $\sigma^m_{[FS]}$. These classifiers were then used to distinguish between sequences of event m and those from other events in different score-spaces using the remaining 20% of the data set.

To form a baseline to compare the effectiveness of modeling the evolution pattern of an event in the semantic concept-space, we trained SVM classifiers on the key-frames of each sequence using the array of concept scores for the key-frames. For each sequence $\alpha_k(t)$, a key-frame $\hat{f}_k$ was selected (we chose the middle I-Frame of the sequence).

The key-frames of the different sequences were then mapped to the semantic concept-space using the same mapping operator $\Phi(.)$. For each event category m, we then trained a SVM classifier to distinguish between its key-frames and those of other events. The same decomposition of the data set used in

[1] *airplane, animal, boat ship, building, bus, car, charts, computer tv, corporate leader, court, crowd, desert, entertainment, explosion fire, face, flag us, government leader, map, meeting, military, mountain, natural disaster, office, outdoor, people marching, person, police security, prisoner, road, sky, snow, sports, studio, truck, urban, vegetation, walking running, waterscape waterfront, weather.*

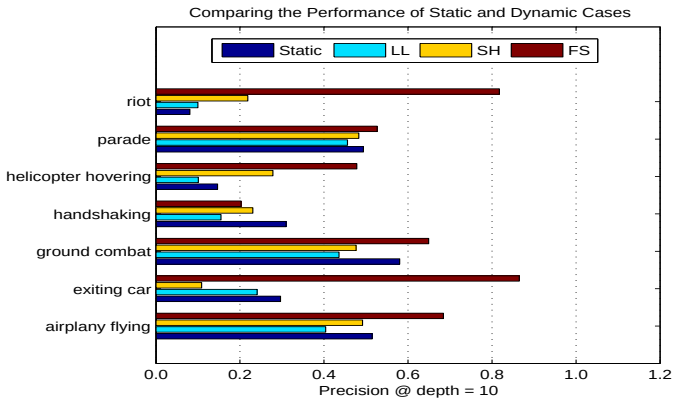

Fig. 5. Comparison between the performance of the proposed modeling approach (using 3 different scores), and the classification of static key-frames in concept-space. The results of applying an event model to test sequences were ranked and the precision at top 10 was obtained. Comparable results could be obtained with different settings for the measure.

the dynamic case was also employed in this case to make the two experiments (dynamic vs. static) comparable. Note that experiments for both the dynamic (evolving models) and static (key-frames) cases were done in 5 rounds of randomized sequences. Figure 5 shows the *precision at depth 10*. From this chart we see that for event *handshaking*, the dynamic model does worse (-10%) than static (key-frame based) one, meaning that mappings of the sequences obtained from this event into the 39-dimensional concept space spanned by the LSCOM-lite detectors lack discriminability. However, for events *riot* $(+58\%)$, *exiting car* $(+46\%)$ and *helicopter hovering* $(+28\%)$, there is signifcant gain when the evolution pattern of the events in concept space is used, employing the *FS* score. Events *airplane flying*, *ground combat*, and *parade* both methods essentially are the same, indicating that given the 39-dimensional concept space there is no dynamics involved in those events. The other two scores, *SH*, and *LL* do not perform as well as *FS*. This could be attributed to the fact that Fisher score captures how well an observed sequence could have been generated by the model, rather than being a single score such as log-likelihood.

4. DISCUSSION AND CONCLUSION

We proposed a concept-centered as opposed to object-centered approach for visual event modeling and detection. This is a novel event modeling approach, which aims at learning the evolution pattern of an event in the semantic concept-space. Results verify that the approach is effective for detecting certain kinds of events. This is the first attempt in exploring the use of semantic concept-space for modeling events that are of dynamic nature.

There are many issues to be addressed in the future in using the concept-centered approach for event modeling and detection. The first and foremost, is the sufficiency of the semantic concept-space for modeling a certain event. As shown in the results, the approach wasn't as effective for certain event categories. This could be due to the fact that the 39 LSCOM-lite concepts used for making the semantic concept-space are not adequate in the context of those events. The second issue is how to select the most informative concept threads and discard nuisance threads to better model the evolution pattern of the events in concept-space. This basically is equivalent to selecting the most informative sub-space of the concept-space.

5. ACKNOWLEDGEMENTS

This material is based upon work funded in whole by the U.S. Government. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. Government. This work was partially supported by DARPA under contract NBCHC050097.

6. REFERENCES

[1] The National Institute of Standards and Technology (NIST), "TREC video retrieval evaluation," 2001–2005, http://www-nlpir.nist.gov/projects/trecvid/.

[2] Y. A. Ivanov and A. F. Bobick, "Recognition of visual activities and interactions by stochastic parsing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000.

[3] M. Brand, N. Oliver, and A. Pentland, "Coupled hidden markov models for complex action recognition," in *CVPR '97: Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*, Washington, DC, USA, 1997, p. 994, IEEE Computer Society.

[4] R. Nevatia, T. Zhao, and S. Hongeng, "Hierarchical language-based representation of events in video streams," in *In the Proceedings of IEEE Workshop on Event Mining (EVENT'03)*, 2003.

[5] M. R. Naphade, T. Kristjansson, B. Frey, and T. S. Huang, "Probabilistic multimedia objects (multijects): A novel approach to video indexing and retrieval in multimedia systems," in *Proceedings of IEEE International Conference on Image Processing (ICIP'98)*, October 1998, vol. 3, pp. 536–540.

[6] L. Xie, S.F. Chang, A. Divakaran, and H. Sun, "Structure analysis of soccer video with hidden markov models," in *Proceedings of International Conference on Accoustic, Speech and Signal Processing (ICASSP'02)*, May 13–17 2002.

[7] A. Natsev, M. R. Naphade, and J. R. Smith, "Semantic representation: search and mining of multimedia content," in *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, 2004, pp. 641–646, ACM Press.

[8] M. R. Naphade, L. Kennedy, J. Kender, S. F. Chang, J. R. Smith, P. Over, and A. Hauptmann, "A light scale concept ontology for multimedia understanding for trecvid 2005," Tech. Rep., IBM Research Technical Report, 2005.

[9] V. N. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998.

[10] "LSCOM lexicon definitions and annotations version 1.0, DTO challenge workshop on large scale concept ontology for multimedia," Tech. Rep., Technical Report 217–2006–3, Columbia University, March 2006.

[11] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, February 1989.

[12] T. S. Jaakkola, M. Diekhans, and D. Haussler, "Using the Fisher Kernel Method to Detect Remote Protein Homologies," in *Seventh International Conference on Intelligent Systems for Molecular Biology*, 1999, pp. 149–158.