Recognition of Complex Events: Exploiting Temporal Dynamics between Underlying Concepts

Subhabrata Bhattacharya[1] Mahdi M. Kalayeh[2] Rahul Sukthankar[3] Mubarak Shah[2]
subh@ee.columbia.edu mahdi@eecs.ucf.edu sukthankar@google.com shah@crcv.ucf.edu
[1]Columbia University [2]University of Central Florida [3]Google Research
http://cs.ucf.edu/~subh/ctr/

Abstract

While approaches based on bags of features excel at low-level action classification, they are ill-suited for recognizing complex events in video, where concept-based temporal representations currently dominate. This paper proposes a novel representation that captures the temporal dynamics of windowed mid-level concept detectors in order to improve complex event recognition. We first express each video as an ordered vector time series, where each time step consists of the vector formed from the concatenated confidences of the pre-trained concept detectors. We hypothesize that the dynamics of time series for different instances from the same event class, as captured by simple linear dynamical system (LDS) models, are likely to be similar even if the instances differ in terms of low-level visual features. We propose a two-part representation composed of fusing: (1) a singular value decomposition of block Hankel matrices (SSID-S) and (2) a harmonic signature (H-S) computed from the corresponding eigen-dynamics matrix. The proposed method offers several benefits over alternate approaches: our approach is straightforward to implement, directly employs existing concept detectors and can be plugged into linear classification frameworks. Results on standard datasets such as NIST's TRECVID Multimedia Event Detection task demonstrate the improved accuracy of the proposed method.

1. Introduction

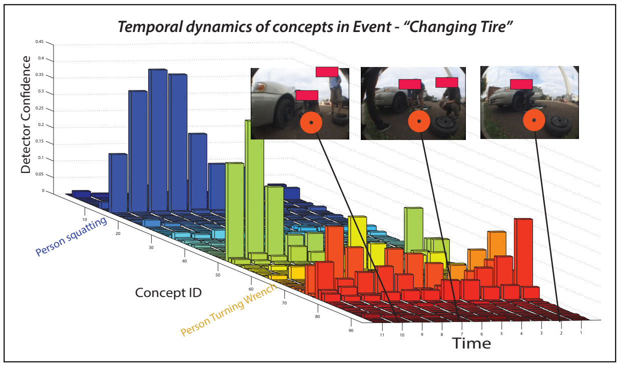

Figure 1. Temporal evolution of concept detector responses during, *changing a vehicle tire*. The proposed representation explicitly captures the joint dynamics of these detectors to better recognize complex events.

Recognition of complex events [2, 20] in unconstrained videos continues to be a challenging problem across the computer vision research community. Recent research in this direction emphasizes on concept based approaches [16] as they provide a richer semantic interpretation than bags of low-level features. Consider the complex event, *changing a vehicle tire* where the following objects: human, vehicle, tire, tools interact in a specific temporal order in a particular scene (typically in an outdoor environment). Thus, for *changing a vehicle tire*, the following sequence of human-object interactions can typically be observed in a consistent temporal order: *rolling tire*, *squatting*, *jacking-up vehicle*, *turning wrench* (Fig. 1). Such interactions can also be termed as spatiotemporal action concepts. Earlier research (e.g., [11]) has demonstrated the feasibility of repeatably detecting such concepts in unconstrained settings; concept-based event representations are expected to gain additional research attention in the future [10].

While concept detectors can provide reasonable estimates of the probability of a particular action being observed during a specified temporal interval, how best to integrate this information in order to accurately recognize complex events remains a challenge. Probabilistic graphical models such as Hidden Markov Models [19, 26], Conditional Random Fields (CRF) [5], Bayesian Networks (BNs) [8] and Dynamic Bayesian Networks (DBNs) [7] have been popular for similar tasks in a variety of domains. Although such methods have performed well in other applications, they seem to be sensitive to noise in the concept detection and require the application of significant domain knowledge in order to achieve computational tractability. This has recently led to the exploration of alternate ap-

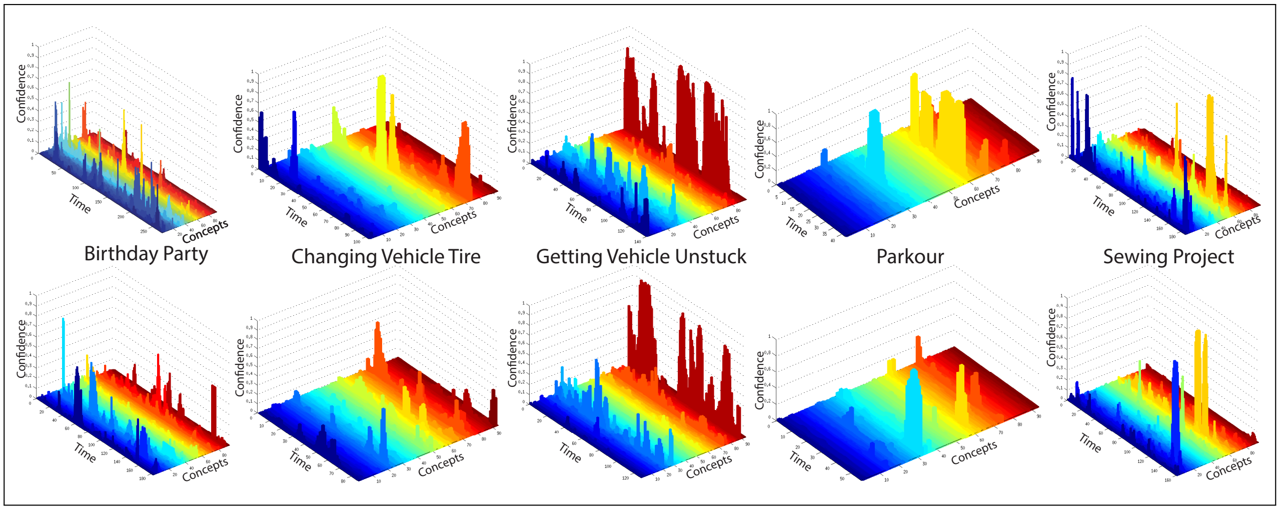

Figure 2. Similarity in evolution of concept sequences within same event categories. 2 sample vector time-series each obtained from a single event class is shown here. Each of the 5 columns represent a complex event.

proaches that serve as motivation for our work [15, 16].

Li et al. [16] use a dictionary of low-level spatiotemporal attributes (e.g., left arm moving up) that are modeled using simple dynamical systems [9, 12] and finally represent each event as a histogram of these attributes. By contrast, we: (a) use bag of visual words model at the attribute detection phase and (b) model events using ideas from Linear Dynamical Systems (LDS) [12, 21].

In our approach, a video is decomposed into a sequence of overlapping fixed-length temporal clips, on which low-level feature detectors are applied. Each clip is then represented as a histogram (bag-of-visual-words) which is used as a clip level feature and tested against a set of pre-trained action concept detectors. Real-valued confidence scores, pertaining to the presence of each concept are recorded for each clip, converting the video into a vector time series. Fig. 2 illustrates sample vector time-series from different event classes through time. We model each such vector time series using a single linear dynamical system, whose characteristic properties are estimated using two different ways. The first technique (termed SSID-S) is indirect and involves computing principal projections on the Eigen decomposition of block Hankel Matrix [21] constructed from the vector time series. The second one (termed H-S) involves directly estimating harmonic signature parameters of the LDS using a method inspired by PLiF [14]. The representations generated by SSID-S and H-S are individually compact, discriminative and complementary, enabling us to perform late fusion in order to achieve better accuracies in complex event recognition.

Linear dynamical systems have been employed for a variety of tasks in video, including video texture analysis [6], tracking moving objects [17, 22, 25], motion segmentation [3, 18] and human action recognition [4, 13]. However, to our knowledge, ours is the first work to compute discriminative video event signatures from vector time series modeled as LDS.

2. Temporal Dynamics

In our formulation, we represent a video V as a sequence of n fixed-length clips, with a certain number of overlapping frames (n differ for videos of different lengths). On each clip, a fixed set of C concept detectors are applied and their respective responses denoting the probability of presence of the corresponding concepts, are recorded. Thus $V \equiv \{\mathbf{c}_0, \mathbf{c}_1, \ldots \mathbf{c}_{n-1}\}$, where each $\mathbf{c}_t \in \mathbb{R}^C$, is a vector containing concept detector responses. Thus each corresponding element of vectors $\mathbf{c}_0, \ldots, \mathbf{c}_{n-1}$ form a sequence of probabilities of detection of a given concept across time in a video.

Now, each concept sequence could be treated independently as an individual time series or modeled using techniques such as Auto Regressive Moving Average [1] processes, from which features or model parameters could be computed. However, doing so ignores the interactions across such series within a video (see Fig. 2). Linear dynamical systems provide a more natural way to model such interactions, with an additional advantage of dimensionality reduction, and can be defined as:

$$\begin{aligned} \mathbf{c}_t &= K\mathbf{x}_t + \mathcal{N}(0, \alpha) && (1) \\ \mathbf{x}_t &= \phi\mathbf{x}_{t-1} + \mathcal{N}(0, \beta); && (2) \end{aligned}$$

where K is the observation matrix $\in \mathbb{R}^{C \times d}$ that maps each observed vector $\mathbf{c}_t$ to a relatively lower dimensional hidden state vector $\mathbf{x}_t \in \mathbb{R}^d$, and ϕ is the dynamics or transition matrix $\in \mathbb{R}^{d \times d}$ that maps between previous and current hidden states. α, β are variances for the Gaussian noise models.

The parameters K, ϕ, α, β and the initial state $\mathbf{x_0}$, need to be identified to characterize a system defined in Eqn.(2).

However, this being a non-convex problem [9, 14, 21], the parameters can only be locally approximated. Moreover, representing these parameters in an appropriate feature space for discriminative classification (for event recognition), is another challenging task. This motivates us to explore strategies discussed next, to derive compact discriminative signatures from a vector time series representation of a video.

2.1. SSID Signatures

In linear systems theory [21], a vector time-series that obeys Eqn.(2) can be arranged in a block Hankel matrix pattern with constant entries along its skew-diagonals as follows:

$$H = \begin{pmatrix} \mathbf{c}_0 & \mathbf{c}_1 & \mathbf{c}_2 & \dots & \mathbf{c}_{n-r} \\ \mathbf{c}_1 & \mathbf{c}_2 & \mathbf{c}_3 & \dots & \mathbf{c}_{n-r+1} \\ \mathbf{c}_2 & \mathbf{c}_3 & \mathbf{c}_4 & \dots & \mathbf{c}_{n-r+2} \\ \dots & \dots & \dots & \dots & \dots \\ \mathbf{c}_{r-1} & \mathbf{c}_r & \mathbf{c}_{r+1} & \dots & \mathbf{c}_{n-1} \end{pmatrix}, \tag{3}$$

where r captures the temporal range over which dynamics can be computed. Note that entries in each successive column of H are shifted by a single time step and each column is itself of dimension $C \times r$ (since each $\mathbf{c}_j$ element denotes a C-dimensional vector of concept responses). Thus, with a reasonable r, one can intuitively identify a reduced number of distinct time groups (columns) required to represent the system in Eqn.(2). In application, H is typically normalized in a preprocessing step using the Frobenius norm:

$$H = \frac{H}{\mathrm{Tr}(H \cdot H^T)^{\frac{1}{2}}}. \tag{4}$$

It is well known from the system identification community that the singular value decomposition (SVD) of a Hankel matrix provides the state space realization of its underlying system. Since H is typically not square, we perform singular value decomposition on $H \cdot H^T$ to obtain the orthonormal bases (singular vectors, E) and singular values in diagonal matrix D.

The SSID signature (SSID-S) is constructed from the m largest singular values along with their corresponding vectors by flattening the matrix given by:

$$[S]_{rC\times m} = [E]_{rC\times m} \cdot [D]_{m\times m} \tag{5}$$

into an rCm-dimensional vector.

Fig. 3 shows an intuitive visualization of SSID-S's benefits over the mid-level concept feature space, computed using 100 videos from each of 5 event classes. Fig. 3(left) shows inter- video Euclidean distance between max-pooled concept detection scores, with each concept score max-pooled temporally to generate a C-dimensional vector per video. While there is some block structure, we see significant confusion between classes (e.g., Birthday Party vs. Vehicle Unstuck). Fig. 3(right) shows the Euclidean distance matrix between videos represented using the proposed SSID signature. The latter is much cleaner, showing improved separability of event classes, even using a simple distance metric.

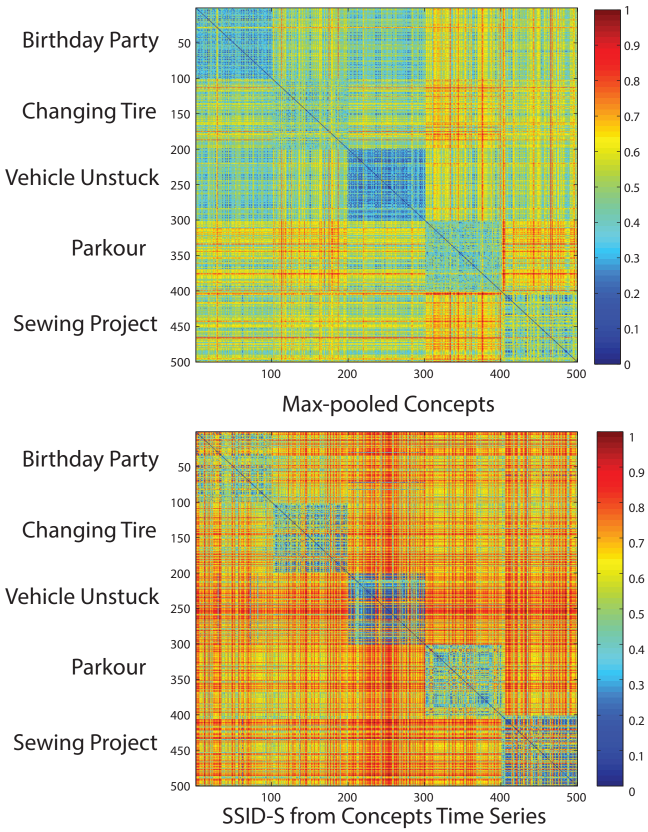

Figure 3. Visualizing distances between videos in conventional max-pooled concept feature space (Top) and in the proposed SSID-S space (Bottom). The strong block diagonal structure of the second distance matrix shows that complex events are better separated in SSID-S.

2.2. Harmonic Signatures

As observed by Li et al. in the context of the PLiF feature [14], the hidden state vector, $\mathbf{x}_t$ can be analyzed as a second-order system, whose behavior (e.g., damped oscillation) is characterized by the eigenvalues ($\{\lambda_i \dots \lambda_L\}$) of the dynamics matrix (ϕ). Following Li et al., we decompose ϕ:

$$[\phi] = [U]_{\theta\times L} \cdot [\Lambda]_{L\times L} \cdot U^T, \tag{6}$$

where Λ is the diagonal matrix of eigenvalues, grouped into their conjugate pairs and ordered according to their phases, and U contains the corresponding eigenvectors. This can be used to transform the hidden state variables, which are otherwise not directly comparable, into a canonical form where their meaning is consistent across videos:

$$\hat{\mathbf{x}}_0 = U^T \cdot \mathbf{x}_0, \quad \text{and,} \tag{7}$$

$$\hat{\mathbf{x}}_{t-1} = U^T \cdot \mathbf{x}_{t-1}. \tag{8}$$

Similarly, the dynamics matrix in Eqn.(2) can be canonicalized as:

$$K_H = K \cdot U. \tag{9}$$

K_H is the Harmonics mixing matrix, where eigenvectors of U are grouped appropriately. Using the above relation in Eqn.(8) enables us to model the observed concept detector responses for the given video ($\mathbf{c}_t$) in the canonical state space, using the eigen dynamics matrix (Λ):

$$\hat{\mathbf{x}}_t = \Lambda^{t-1} \cdot \hat{\mathbf{x}}_0 + \mathcal{N}(0, \beta) \tag{10}$$

$$\mathbf{c}_t = K_H \cdot \Lambda^{t-1} \cdot \hat{\mathbf{x}}_0 + \mathcal{N}(0, \alpha). \tag{11}$$

As in Li et al. [14], we recover an estimate for the harmonic mixing matrix, K_H, in Eqn.(11) using an Expectation-Maximization based algorithm. We observe that estimates for K_H require fewer than 10 iterations to converge in practice. The dimensionality of the hidden state vector (d) is a free parameter but as shown in Section 3, the overall mAP is relatively insensitive to its setting.

Finally, we generate a real-valued Harmonic Signature (H-S) for a given video by flattening the magnitudes of the entries in K_H (similar to PLiF). However, unlike PLiF which is computed in batch over a set of data, the harmonic signature for each video is computed individually, enabling parallelized processing of videos.

Fig. 4 shows an intuitive visualization of H-S's benefits over the mid-level concept feature space, computed using the same 100 videos from each of five event classes as seen in Fig. 3. Dots (corresponding to videos) in the max-pooled concept feature space are C-dimensional, whereas those in H-S space are Cd dimensional. The scatter plot is generated by projecting each point in the feature spaces to two dimensions using PCA. We observe that the videos are much more separable in H-S space as compared to the max-pooled concept feature space: four of the five complex event classes are visually separable, even in the 2D visualization.

3. Experiments

This section details our experimental protocols and describes the video datasets. We also discuss the baseline methods against which we compare the proposed approach. A small section on selection of parameters is provided towards the end, as well as brief look at the computational complexity of the entire system.

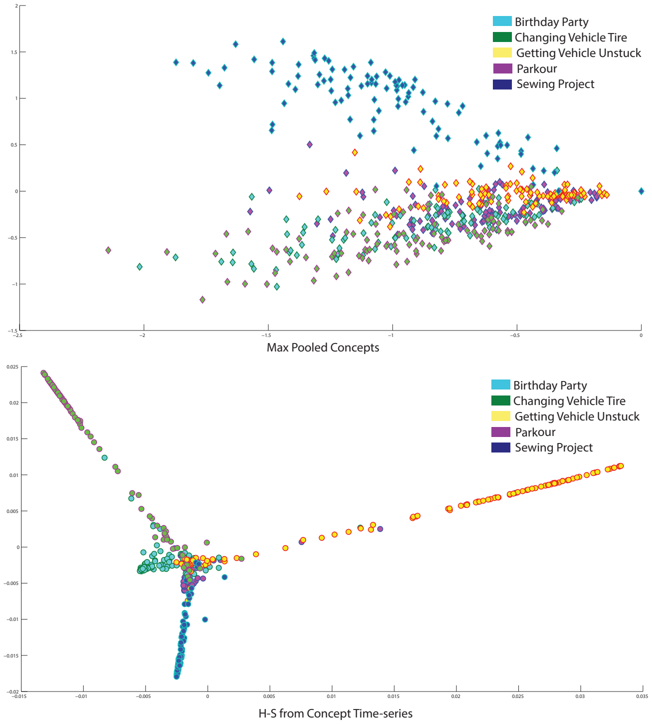

Figure 4. Visualizing separability between events in max-pooled concept space (top) and H-S space (bottom). Each data point is a video from one of five event classes.

3.1. Datasets

A number of datasets have been released by NIST as part of TRECVID MED competition[1] that has been organized since 2010. We have selected two datasets for our evaluation. The first, referred to as MED11 Event Collection (MED11EC) was released in 2011 and consists of 2,062 videos from 15 different complex event categories. The second, referred to as MED12 Event Collection (MED12EC) is similar to the first, containing 2,000 videos pertaining to 10 additional events. The event names and the corresponding videos available in each event category are summarized in Table 1.

We also directly compare the performance of our proposed approach against two recent state-of-the-art algorithms on TRECVID MED DEV-T, a dataset consisting of 10,723 videos from five categories *Attempting a board trick*, *Feeding an animal*, *Landing a fish*, *Wedding ceremony*, and *Working on a woodworking project* as well as a large number of *null class* videos that do not correspond to any of these classes. Given 150 training videos from each class, the goal is to retrieve the desired videos among distractors.

[1] http://www.nist.gov/itl/iad/mig/med.cfm

Table 1. Summary of two datasets used in this paper. The first half (E001–E015) is from MED11EC, while the second half (E021–E030) is from MED12EC. The number of videos ranges from 111 for to 299 per class.

ID	Event Name	[N]	ID	Event Name	[N]
E001	Board-trick	287	E002	Feeding Animal	299
E003	Landing Fish	234	E004	Wedding	251
E005	Woodworking	263	E006	Birthday	173
E007	Changing Tire	111	E008	Flash-mob	173
E009	Vehicle Unstuck	132	E010	Grooming Animal	138
E011	Making Sandwich	126	E012	Parade	138
E013	Parkour	112	E014	Repairing Appl.	123
E015	Sewing Project	120			
E021	Bike-trick	200	E022	Giving Directons	200
E023	Dog-show	200	E024	Wedding	200
E025	Marriage Proposal	200	E026	Renovating Home	200
E027	Rock-climbing	200	E028	Town-hall Meet	200
E029	Winning Race	200	E030	Metal crafts	200

3.2. Spatiotemporal Concept Detectors

Following Jiang et al. [11], we identify a set of 93 unique spatiotemporal concepts by parsing the textual definition of events provided within NISTs TRECVID MED 11–12 database. We obtain a small number of training samples for each of these concepts (<30) from a subset of videos (<50 videos per event class) in the TRECVID MED 11–12 dataset. As an example, the following mid-level concepts: *person clapping, person blowing candles*, etc. are all trained from clips taken from videos of the *Birthday Party* complex event. Human annotators are asked to mark the approximate beginning and ending frame of the concepts in the videos. A list of the 93 concepts is available on the project website. We use a bag of words technique on dense trajectory based spatiotemporal features [24], extracted from each annotated clip to represent it. A vocabulary size of $2,048$ is observed to deliver optimal trade-off between performance and computational constraints, and is hence chosen as the default vocabulary size for all experiments. A publicly available implementation[2] of binary SVM classifiers with histogram intersection kernels are used as our concept detectors. These concept detectors are applied on BoVW representations of each fixed-length clips (300 frames with an overlap of 60 frames) from every video. Confidence scores $\in (0,1)$ corresponding to each of the 93 concept detectors are collected to create a vector time series for every video.

3.3. Baseline Methods

Since, the datasets are relatively new and research on concept detection is still in its infancy, it is very difficult to compare our work with published methods [2, 20] that involve fusion of multiple low-level feature representations. To make fair comparisons, we implemented three independent baseline methods that all share the same mid-level concept representation.

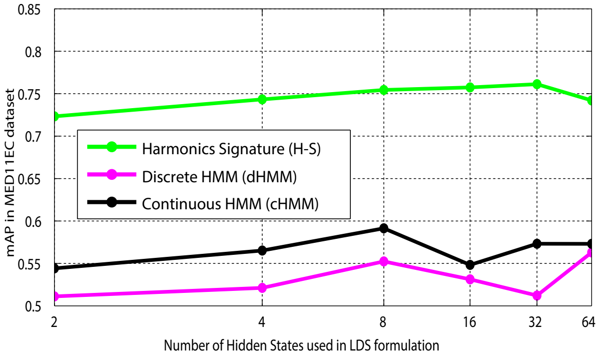

Figure 5. Mean Average Precision (mAP) scores of event detection on a subset of MED11EC dataset is shown against hidden state dimensionality ($d \in \{2 \dots 64\}$) for learning LDS parameters using three similar approaches, with green – Harmonic Signatures (H-S), purple – Discrete HMMs (dHMM) and black – Continuous HMMs (cHMM). H-S has significantly higher mAP scores compared to dHMMs and cHMMs and is relatively insensitive to the dimensionality of the hidden state vector d.

The first baseline extracts Discrete Cosine Transform (DCT) coefficients for each concept detector response sequence and coefficients from all sequences from a video are concatenated to form the final temporal descriptor. A linear SVM is used to predict event labels. The best performance is achieved for 64 coefficients per time-series, requiring a 93×64 dimensional feature per video.

The other two baselines are implemented with two variants of HMMs: discrete HMM and continuous HMM. In both cases we perform experiments with 6 different choices for the dimensionality of the hidden state vector, $d \in \{2 \dots 64\}$ (doubling at each step). Initial parameters for both experimental settings (refer Eqn.(2)), such as the prior ($\mathbf{x_0}$), transmission matrix (ϕ) and observation matrix (K) are determined from a stochastic process input with $\mathcal{N}(0,1)$.

For the discrete HMM baseline, we obtain the maximum confidence at each time step from every observation ($\mathbf{c}_t$) in a given vector time series, and associate the corresponding concept label to generate the input symbol sequence. This step is not required in a continuous HMM framework, as each state variable in this case is modeled using the distribution of confidences at each time step. In both cases, event-specific models are trained. Given a testing sequence, the maximum likelihood of generating the input sequence given an event model is computed using a forward Viterbi algorithm, with the highest-scoring event returned as the prediction from the model.

3.4. Parameter Selection for SSID-S

Fig. 6 empirically illustrates how different combination of parameters affect recognition performance over subsets of videos from MED11EC and MED12EC datasets, using

[2] http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/

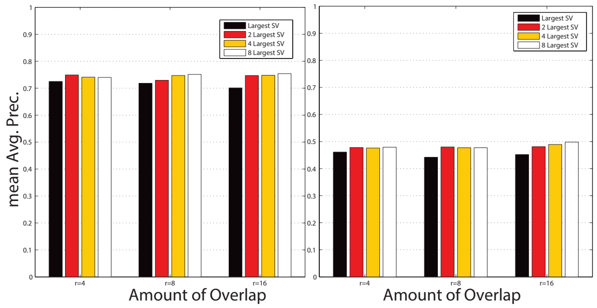

Figure 6. mAP for different overlap settings (r) and largest eigenvectors used for projection (m). Subsets of samples from both MED11EC (left) and MED12EC (right) are shown here.

the proposed SSID-S algorithm. While constructing block Hankel matrix, we experiment with three different overlap settings, i.e. $r \in \{4, 8, 16\}$. For each overlap setting, we evaluate four settings for dimensionality reduction $m \in \{1, 2, 4, 8\}$ in computing the SSID-S feature. Performance is relatively insensitive to parameter choice; we use $r = 4$, and $m = 2$ in our subsequent experiments.

3.5. Computational Complexity

Since our approach employs existing concept detectors, our analysis focuses on the specific computational cost of the proposed descriptors. For the SSID-S algorithm, the Block Hankel Matrix involves only stacking vectors in a specific pattern depending on the overlap (r) owing to $O(r)$. The complexity of Singular Value decomposition on $H \cdot H^T$ is $O(m(Cr)^3)$, where m is the number of largest eigenvalues, C being the number of concepts. The main computational complexity of the Harmonics Signatures algorithm is in the iterative E-M stage, which while theoretically unbounded, never took more than 200 steps among the videos in our datasets. In terms of wall-clock time, on a standard laptop (2.6 GHz CPU with 4GB physical memory), the average run time for extracting SSID-S from a single video (20 clips) is 540ms, while that for H-S is 2200ms.

4. Results

This section presents details about our experimental results. In Sec. 4.1, we compare against baselines that employ the same mid-level concept time series representation as the proposed method. Sec. 4.2 presents insights on how our approach can be directly fused with concept-level classifiers for complex event recognition. Finally, Sec. 4.3 compares our approach against two state-of-the-art methods that also exploit temporal information for complex event recognition.

4.1. Comparison with Baselines

Here we report results on the MED11EC and MED12EC datasets. For each event category, we use around 100 positive sample videos (not used during training of concept detectors) and equal number of negative samples (videos from all other event classes). The same mixture is used for evaluation across all three baselines and our variants of our proposed approach. For our approach, we consider SSID-S alone, H-S alone and a weak fusion of the two. Fig. 7 reports comprehensive summary of mAP across all 25 events using the six methods (3 baselines and 3 variants of the proposed method).

We make the following observations. First, we see that all three baselines under-perform all of the variants of the proposed method, even though all methods employ exactly the same time series mid-level concept representation. The DCT baseline is poorest at representing the vector time series data, and this may be attributed to its inability to capture the complexity of the concept dynamics, particularly with a limited number of bases.

As there is no principled way to determine the optimal number of hidden states for the two HMM baselines, we experiment with a different number of hidden states along with corresponding Gaussian stochastic prior matrices. We note that the best-performing discrete HMM employs 64 hidden states while the best continuous HMM yields highest mAP with just 8 states. The continuous HMM consistently outperforms the discrete HMM, both on MED11EC and MED12EC datasets (Fig. 7). However, estimating the mixture parameters is computationally intensive and for a significant fraction of the videos in our datasets, the training does not converge.

Both of our proposed signatures perform better than the baseline methods by a significant margin $(22--35\%)$. We observe a significant reduction $(\sim\ 24\%)$ in mAP on the MED12EC dataset in comparison to MED11EC. This is primarily because the mid-level concept detectors (trained only on MED11EC), do not have sufficient coverage on the 10 new event classes in MED12EC. While our proposed methods still outperform the baselines on MED12EC, the cross-dataset drop in performance underscores the importance of employing a sufficiently broad set of concept detectors so as to cover the observed actions in the test set. We also observe that the late fusion of SSID-S and H-S (denoted “combined temporal representation”) helps more on the MED12EC dataset.

4.2. Fusion with Direct Video-Level Predictions

Our earlier experiments employed SSID-S or H-S (or their fused combination) in exclusion to the mid-level features. However, a natural question is whether we can further improve performance by performing late fusion using SSID-S and H-S with video-level complex event predictions generated directly from the mid-level concept detectors.

Table 2 presents mAP results from this experiment. *BoC* denotes the video-level prediction generated from average-pooling the concept predictions through time. *BoVW* de-

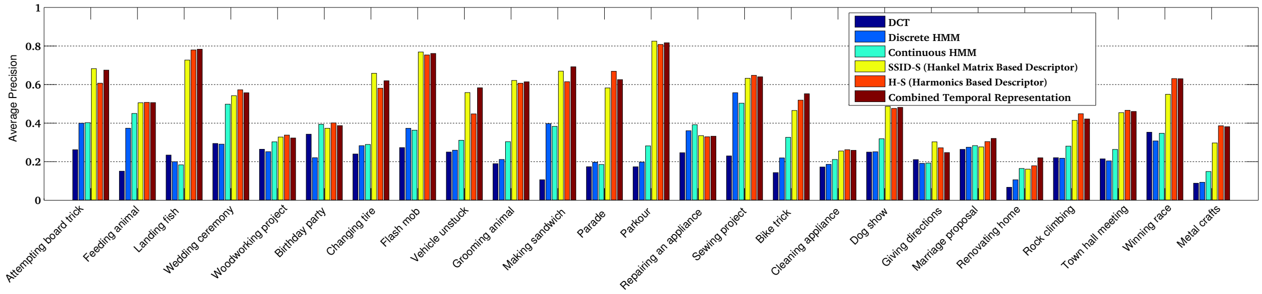

Figure 7. mAP for complex event recognition of all 25 categories in MED 11-12 EC using three baselines (DCT, Discrete HMM, Continuous HMM) and three variants of the proposed method (SSID-S alone, H-S alone, weak fusion of both signals). All variants of the proposed method consistently dominate the baselines, with weak fusion providing a slight benefit among variants of the proposed method.

Table 2. Results of fusion with low-level event classifiers.

Datasets	BoC	BoVW	CTR	F1 [CTR+BoC]	F2 [CTR+BoVW]
MED11EC	0.72	0.75	0.76	0.77	0.79
MED12EC	0.46	0.48	0.53	0.50	0.56

notes direct video-level predictions generated using a large vocabulary of MBH features. *CTR* denotes the late fusion of SSID-S with H-S (Combined Temporal Representation), discussed above. The remaining two entries, *F1* and *F2* show the fusion of CTR with video-level predictions from BoC and BoVW, respectively. The results indicate that fusing our approach with video-level predictions helps both on MED11EC and MED12EC.

4.3. Comparison with State-of-the-Art Methods

Following the same protocol as suggested in [16, 23], we report our performance on five events from the DEV-T dataset in Table 3. Both Tang et al. [23] and Li et al. [16] analyze the temporal structure of videos using different dynamical system formulations. Columns 3 and 4 are quoted from the results published in [16]. Our results are shown in the last three columns, with CTR denoting late fusion of SSID-S and H-S.

It is interesting to observe that despite their relatively simple formulations, both SSID-S and H-S (as well as their fusion – CTR) outperform both of the more complicated earlier methods. Note also that the mAP scores reported here are lower than those in Table 2 and Fig. 7. This is mainly due to the fact that over $9,500$ videos contained in the DEV-T dataset do not match the small number of event categories from which the concept detectors are trained. We conjecture that expanding the pool of concept detectors to a substantially larger set could address this limitation.

5. Conclusion

Modeling the temporal dynamics of spatiotemporal concepts occurring in a video can provide useful cue towards understanding its semantic structure. We introduce two different techniques to model the temporal relationships among spatiotemporal concepts of a video using foundations from Linear dynamical systems. Through several detailed experiments, we demonstrate the efficacy of our proposed method over contemporary methods that are used extensively by computer vision and multimedia researchers to analyze the temporal structure of videos. Our proposed approach is straightforward to implement and computationally inexpensive and could also be used effectively on other tasks, such as multimedia event recounting, which demand a better understanding of the temporal structure in multimedia data. As part of future work, we plan to extend these ideas to large corpora of concepts, which may be learned in a less supervised fashion, given a collection of videos.

Table 3. Performance comparison with the state of the art for 5 training events from the DEV-T dataset (> 10,000 videos).

Events	Random	[23]	[16]	SSID-S	H-S	CTR
Board-Trick	0.011	0.15	0.29	0.31	0.33	0.33
Feed Animal	0.010	0.03	0.07	0.11	0.09	0.12
Land Fish	0.008	0.14	0.28	0.26	0.27	0.27
Wedding	0.007	0.15	0.22	0.29	0.32	0.31
Woodworking	0.009	0.08	0.18	0.17	0.21	0.22
mAP	0.009	0.11	0.21	0.23	0.23	0.24

Acknowledgments

This work was supported in part 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.

References

[1] G. E. P. Box and G. Jenkins. *Time Series Analysis, Forecasting and Control*. Holden-Day, Incorporated, 1990. 2

[2] L. Cao, Y. Mu, A. Natsev, S.-F. Chang, G. Hua, and J. R. Smith. Scene Aligned Pooling for Complex Video Recognition. In *ECCV*, pages 688–701, 2012. 1, 5

[3] A. B. Chan and N. Vasconcelos. Modeling, clustering, and segmenting video with mixtures of dynamic textures. *TPAMI*, 30(5):909–926, 2008. 2

[4] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal. Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In *CVPR*, pages 1932–1939, June 2009. 2

[5] C. I. Connolly. Learning to Recognize Complex Actions using Conditional Random Fields. In *Proceedings of the 3rd international conference on Advances in visual computing - Volume Part II*, ISVC'07, pages 340–348, 2007. 1

[6] A. Doretto, G.and Chiuso, Y. Wu, and S. Soatto. Dynamic textures. *International Journal of Computer Vision*, 51(2):91–109, 2003. 2

[7] C.-L. Huang, H.-C. Shih, and C.-Y. Chao. Semantic Analysis of Soccer Videos using Dynamic Bayesian Network. *IEEE Transactions on Multimedia*, 8(4):749 –760, 2006. 1

[8] S. S. Intille and A. F. Bobick. Recognizing Planned, Multiperson Action. *CVIU*, 81(3):414 – 445, 2001. 1

[9] E. A. Jackson. *Perspectives of Nonlinear Dynamics, Chapter 2*. Cambridge University Press, 1991. 2, 3

[10] Y.-G. Jiang, S. Bhattacharya, S.-F. Chang, and M. Shah. High-level event recognition in unconstrained videos. *International Journal of Multimedia Information Retrieval*, 2(2):73–101, 2013. 1

[11] Y.-G. Jiang, X. Zeng, G. Ye, S. Bhattacharya, D. Ellis, M. Shah, and S.-F. Chang. Columbia-UCF TRECVID2010 Multimedia Event Detection: Combining multiple modalities, contextual concepts, and temporal matching. In *Proc. NIST TRECVID Workshop*, Gaithersburg, MD, 2010. 1, 5

[12] T. Kailath. A view of three decades of linear filtering theory. *IEEE Transactions on Information Theory*, 20(2):146–181, 1974. 2

[13] B. Li, M. Ayazoglu, T. Mao, O. I. Camps, and M. Sznaier. Activity recognition using dynamic subspace angles. In *CVPR*, pages 3193–3200, 2011. 2

[14] L. Li, B. A. Prakash, and C. Faloutsos. Parsimonious Linear Fingerprinting for Time-series. *Proc. VLDB Endow.*, 3(1-2):385–396, Sept. 2010. 2, 3, 4

[15] W. Li and N. Vasconcelos. Recognizing activities by attribute dynamics. In *NIPS*, pages 1115–1123, 2012. 2

[16] W. Li, Q. Yu, H. Sawhney, and N. Vasconcelos. Recognizing activities via bag of words for attribute dynamics. In *CVPR*, pages 2587–2594, 2013. 1, 2, 7

[17] H. Lim, O. Camps, M. Sznaier, and V. Morariu. Dynamic appearance modeling for human tracking. In *CVPR*, volume 1, pages 751–757, June 2006. 2

[18] R. Lublinerman, M. Sznaier, and O. I. Camps. Dynamics based robust motion segmentation. In *CVPR*, pages 1176–1184, 2006. 2

[19] P. Natarajan and R. Nevatia. Online, Real-time Tracking and Recognition of Human Actions. In *Proc. IEEE Workshop MVC*, pages 1–8, 2008. 1

[20] P. Natarajan, S. Wu, S. Vitaladevuni, X. Zhuang, S. Tsakalidis, U. Park, R. Prasad, and P. Natarajan. Multimodal feature fusion for robust event detection in web videos. In *CVPR*, pages 1298–1305, 2012. 1, 5

[21] P. V. Overschee and B. D. Moor. N4SID: Subspace Algorithms for the Identification of Combined Deterministic-Stochastic Systems. *Automatica*, 30(1):75–93, 1994. 2, 3

[22] B. Ristic, S. Arulampalam, and N. Gordon. *Beyond the Kalman filter: particle filters for tracking applications*. Artech House, 2004. 2

[23] K. Tang, L. Fei-Fei, and D. Koller. Learning latent temporal structure for complex event detection. In *CVPR*, pages 1250–1257, 2012. 7

[24] H. Wang, A. Kläser, C. Schmid, and L. Cheng-Lin. Action Recognition by Dense Trajectories. In *CVPR*, 2011. 5

[25] S.-K. Weng, C.-M. Kuo, and S.-K. Tu. Video object tracking using adaptive Kalman filter. *J. Vis. Comun. Image Represent.*, 17(6):1190–1208, Dec. 2006. 2

[26] J. Yamato, J. Ohya, and K. Ishii. Recognizing Human Action in Time-Sequential Images using Hidden Markov Model. In *CVPR*, 1992. 1