Improving Concept Detection by Utilizing Temporal Relationships

Christian Weigandt
Department of Mathematics
and Computer Science
High Point University
High Point, NC 27262
Email: weigac10@highpoint.edu

Levi Smith
School of Electrical Engineering
and Computer Science
University of Central Florida
Orlando, FL 32816
Email: l.smith@knights.ucf.edu

Khurram Soomro
School of Electrical Engineering
and Computer Science
University of Central Florida
Orlando, FL 32816
Email: ksoomro@eecs.ucf.edu

***Abstract*—Detection of concepts, or actions, within a video is a challenging and important task. Detecting concepts as simple as a person walking or as subtle as a wheel turning is challenging for many reasons. We introduce a method to improve concept detection through the use of known temporal relationships between concepts. We compare our approach with baseline concept detection accuracies to show that the notion of temporal relationships is beneficial to the task of concept detection. We apply our method to two datasets - TRECVID and a self-annotated soccer dataset. Each dataset shows varying levels of improvement and helps to indicate the generality of our approach.**

I. Introduction

Detection of concepts in a video is a complex task in computer vision. Concepts may be thought of as simple actions such as *walking*, *flipping a skateboard*, *hammering*, etc. Temporally isolating these concepts is essential in determining important keyframes of a video. The task of concept detection is applicable to many fields, such as security or sports analysis, and aids in reducing the amount of video to only those frames that contain important events. The typical process for detecting concepts in a video is to split the video into a sequence of clips and use a support-vector machine, or SVM, to classify the concept that occurs in each clip. When classifying individual clips within a video, the context for that clip is dismissed, and any beneficial information gained therein is lost. Thus, we aim to improve concept detection by taking advantage of known temporal relationships between concepts. For example, for an event such as fishing, we expect the concept of $casting$ to always precede the concept of $reeling$. We can use this information to boost the accuracy of our concept detector and correct mistakes that it may make in detection.

II. Related Work

Concept detection is essential to a variety of tasks such as event classification, compactly summarizing videos, rare event detection, etc. The original aim of our research was to improve action detection across soccer broadcast videos in order to isolate highlights. Many existing works [1], [2] take advantage of broadcasting techniques such as slow motion replays, logo transitions, and visible grass ratio in an attempt to isolate highlights. Our approach was to isolate highlights by recognizing the actions themselves in order to make action detection more robust across different broadcasting styles. Many works exist on the task of action detection. For example, Yang and Shah [3] use low-level features across multiple modalities, along with DBNs, to recognize higher-level semantics of a video. There are also many works on the broader task of reducing videos to their important actions. Rodriguez [4] categorizes spatio-temporal regions and temporally shifts them so that a summarization video may contain multiple actions at the same time. Kwon and Lee [5] form a graph of spatial and temporal regions in a video and implement an energy-minimization algorithm to summarize a video by its rare events. Masumitsu and Echigo [6] use reinforcement learning to summarize a video. Li and Sezan [7] perform sports video summarization by temporally localizing the occurrence of plays. However, this only applies to non-continuous sports such as baseball and American football. Tu, Sun, and Luo [8] use low-rank subspace segmentation to summarize a video based on its important keyframes. Shmuel Peleg [9], [10], [11], [12] also has many papers on summarizing videos and clustering of similar actions. These works combine actions into compact segments in order to best reduce the number of clips that are used to show actions.

III. Causality

Our main work is to improve concept detection using temporal relationships. We do this through the use of *causality*. We define causality as $p(B|A)_t$ where

$$p(B|A)_t = \frac{|B \cap A|_t}{|A|}. \quad (1)$$

We define $|B \cap A|_t$ to be the number of times that B is the first concept that occurs within t seconds of concept A, and $|A|$ as the total number of occurrences of concept A. Thus $p(B|A)_t$ is simply the probability that concept A *leads to* or *causes* concept B. This gives us a temporal relationship between concepts which we can take advantage of to improve detection. We find the causality between all concepts in a given event and create a causality matrix to be used in our model. One such causality matrix is given in Figure 1. As can be seen in the figure, relationships such as goal leading to celebration or corner leading to an attempt at goal are highlighted. Concepts or actions such as a goal can be seen to have a strong temporal relationship with themselves. This is given by the fact that soccer broadcast videos tend to replay highlights, and thus actions such as a goal lead to replays of itself, and our model

will show this self-relation. We then use this causality matrix in our model to increase the accuracy of concept detection by giving more weight to the most probable concept to come next in a sequence.

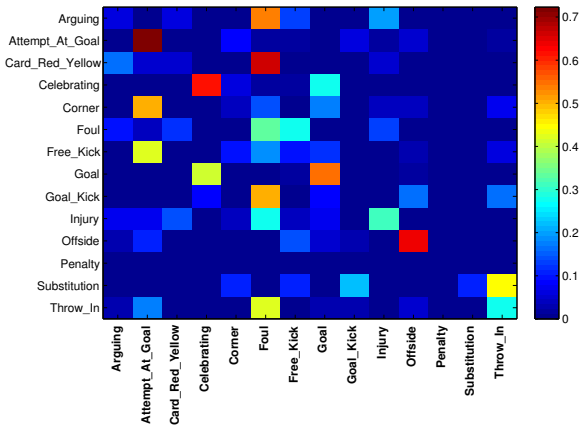


Figure 1. Causality matrix for soccer dataset with $t = 12s$, normalized across rows to highlight relationships

IV. Graph-based Approach

We propose a modified Markov chain (MMC) method to improve the performance of concept detection throughout a video. This method takes advantage of the temporal relationships that are present within an event, thus causing the transitions between detected concepts to be made smoother and more consistent. Using the notion of causality, we are able to better detect concepts in a video based upon the previous concept that occurred. This helps to correct for inaccuracies in the SVM classification of each clip.

A. Graphical Representation

Our model can best be visualized as a graph with concepts as the nodes and transitions as the edges, as seen in Figure 2. The nodes are representative of the probability that a certain clip is of a certain concept. For example, the node for goal in the first layer would hold the probability that the first clip contains a goal. The convoluted edges between layers are representative of the transition probabilities of one concept leading to another, i.e. the causality, and the straight edges are representative of SVM probabilities at a clip. The graph is non-cyclic and each clip, other than the initial clip, is represented by two layers of the graph. The first layer of our model is initialized to the SVM probabilities of each concept at the first clip. These are multiplied by the probabilities of transitioning between concepts and added to get the probabilities of each concept at the following clip. We then multiply by the concept probabilities for this clip to move to the next layer, and the concept with the maximum probability at each odd layer is assigned to that clip.

B. Markov Chain Process

Our method is based off a Markov chain model where the transition probabilities are taken from the causality matrix. We iterate through each clip of the video, incorporating the SVM probabilities of each concept as we go. Our model is given by the equation

$$\begin{bmatrix} P_{K,1} \\ \vdots \\ P_{K,N} \end{bmatrix} = \left(\begin{bmatrix} C_{1,1} & \dots & C_{1,N} \\ \vdots & \ddots & \vdots \\ C_{N,1} & \dots & C_{N,N} \end{bmatrix}^T \begin{bmatrix} P_{K-1,1} \\ \vdots \\ P_{K-1,N} \end{bmatrix} \right) \cdot \begin{bmatrix} S_{K,1} \\ \vdots \\ S_{K,N} \end{bmatrix}, \tag{2}$$

where $P_{K,i}$ is the updated probability that clip K contains concept i as given by our model, $C_{i,j}$ is the causality probability that concept i leads to concept j (previously defined as $p(j|i)_t$), $S_{K,i}$ is the original probability that clip K contains concept i (generated by an SVM in our case), and N is the number of concepts. We initialize the probabilities at clip number 1 to simply be those generated by an SVM, thus $P_{1,i} = S_{1,i}$. One can see that the initial matrix multiplication within the parentheses is simply a single iteration of a Markov process over a state space. The causality matrix is the transition matrix for our model. However, we incorporate SVM probabilities as priors in the Markov Chain approach. Together, the fused probability score better represents the causal dependence of concepts. We involve this prior that is independent of the previous state, and thus we lose any advantageous, and more importantly disadvantageous, properties of a Markov chain. Without the final variable, our model would converge to a stationary distribution. This property is neither needed nor wanted, as we would not expect videos with enough clips to converge to clips all containing the same concept. We want to allow for the unique properties of each clip to affect how our model interprets its concept, and thus we include the final multiplication. In order to more easily describe the model, the multiplication for a single concept, concept 1, is given by the equation

$$\begin{aligned} P_{(K,1)} = (P_{(K-1,1)}C_{1,1} + P_{(K-1,2)}C_{2,1} + \dots + \\ P_{(K-1,N)}C_{N,1}) * S_{(K,1)}. \end{aligned} \tag{3}$$

This is simply the dot product of the probabilities from the previous clip with the first column of the causality matrix, times the probability that the current clip contains concept 1. We find that this equation is a good representation for the flow of one concept to the next in a video. Once we have the updated probabilities for each clip in the video, we take the maximum of each clip and assign that clip to contain that concept.

V. Other Methods

Apart from the MMC approach discussed above, we tested both an iterative and a dynamic programming-based approach to solve the task of finding the optimal sequence of concepts. We aim to show the benefit of causality to the task of concept detection rather than simply the benefit of our unique approach.

A. Iterative Approach

In the first approach, we use a graph which only contains one layer per clip and is shown in Figure 3. Each node in the graph contains the probability for that concept occurring in that clip. The edge weights between the nodes are the values from the corresponding entries in the causality matrix. (see Figure 1) Our task is to find the path through this graph with the highest weight.

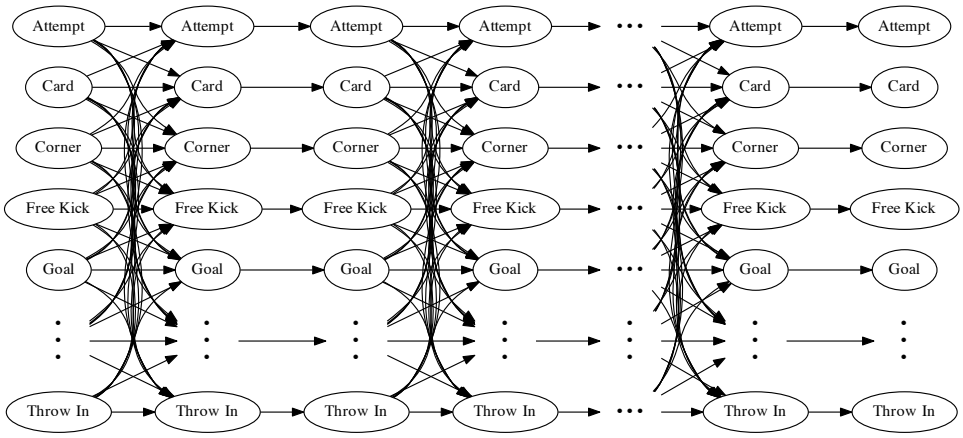


Figure 2. Graphical representation of our model as applied to the soccer dataset.

1) Initialization: The first step is to initialize the graph by choosing a random path through the graph. This step is to merely provide a starting point for the algorithm.

2) Traversal: Starting from the first layer, the algorithm selects the node with the maximum sum of probability and outgoing edge weight to the second node, which is fixed as either the randomly selected node, or the selected node from a previous iteration. When choosing the node at layer i, the nodes at layer $i-1$ and $i+1$ are fixed from the previous step in the algorithm or from a previous iteration, respectively. As such, at the second node, and all subsequent nodes until the last node, the algorithm chooses the node with the maximum sum of incoming edge, probability, and outgoing edge. This continues until the last node, where the algorithm must only select the node with the maximum sum of incoming edge and probability.

3) Iteration: The algorithm takes as input the desired number of iterations. The aim of the iterations is to converge towards the optimal solution over numerous iterations, using the output from the previous iteration as input to the next iteration.

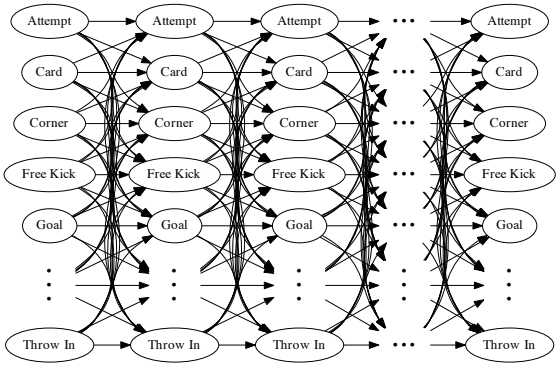


Figure 3. Architecture for both the iterative and dynamic programming approaches as applied to the soccer dataset.

B. Dynamic Programming

Another approach uses dynamic programming to achieve an optimal path through the graph. We used the same graph that is used in the iterative method (see Figure 3). The dynamic programming approach to our problem is akin to finding the path with the least cost through a graph [13]. This approach is effectively a shortest-path approach to our problem and performs quite well as can be seen in Section VI-C.

VI. Experimental Results

A. Datasets

We tested our methods on two separate datasets: the 25-event TRECVID dataset and also a self-annotated soccer dataset. These datasets provided a wide spectrum of events and concepts on which to test our proposed method. The TRECVID dataset is divided into two subsets, EC11 and EC12, with each having a unique set of events and concepts. The first, EC11, is composed of fifteen events containing 92 unique concepts dispersed among them. Events range from *attempting a board trick* to *making a sandwich*. The second, EC12, is composed of ten events and 93 unique concepts, with events such as *cleaning an appliance*, *rock climbing*, and *winning a race without a vehicle*. The final dataset is a collection of self-annotated soccer clips taken from nine different broadcasted soccer matches. There are over 1600 annotated clips of the fourteen concepts that comprise this dataset. Concepts include common soccer actions such as *throw-ins*, *corners*, and *goals*.

For EC11 and EC12, we applied SVM concept detectors that were trained on MBH features to sliding windows of size 180 frames with a displacement size of 30 frames for each video. Thus each video is described by an N-by-K matrix of SVM probabilities where N is the number of concepts, 92 for EC11 and 93 for EC12, and K is the number of clips. The same technique was used for the soccer dataset but for STIP features and with sliding windows of size 5 seconds with a displacement size of 5 seconds. So videos in the soccer dataset are described by an N-by-K matrix of SVM probabilities where N is 14.

B. Intermediary Results

1) Time Intervals: Due to the temporal nature of causality, different events have unique optimal t values, or time steps, for their causality matrices. As can be seen Figure 4, for event *flash mob gathering*, the ideal value for t in creating the causality matrix is 8 seconds. As we increase the time step, it can be seen that the accuracy of our method decreases. This is because our method is creating relationships between concepts that should not exist, i.e. it is recognizing a concept to *cause* another concept when it occurs over 20 seconds later. This is not true within an event such as *flash mob gathering* where concepts or actions, such as *dancing*, *running*, or *waving*, occur fairly rapidly. Thus we take the uniqueness of each event into account and attempt to find the best time interval for each event.

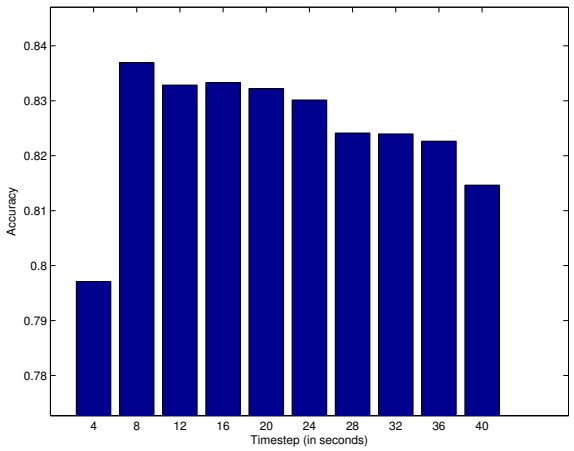


Figure 4. Accuracies for concept detection on event *flash mob gathering* using different time intervals for causality

2) Updating Probabilities: With our approach to concept detection, we calculate a sequence of concept probabilities for each clip and are then able to create an updated sequence of probabilities using causality. In order to better visualize how our method affects concept detection for a video, Figure 5 is shown. The sequences are for the event of *fishing*. We see that the SVM classifier recognizes concept 2, *animal grabbing offered food and eating*, as existing within the video. Even though this concept does not exist in this event, one can see how this would be easily confused with baiting a hook or something of that matter. Our method corrects for this and removes that concept from the sequence (other than at clip 1 as this is the initialization step). We see that after applying our method, only relevant concepts are highlighted and there is a smoothing effect which takes place across clips.

C. Results

We tested the MMC method on all three datasets. As stated in Section VI-B1, each event is unique in its optimal time step, and, therefore, we show results for each event when using the time step that provided the highest accuracies for that event. We determined the optimal time step for each event empirically, testing on t values from 120-1200 frames at intervals of 120 frames. To compute the accuracies, we look only at

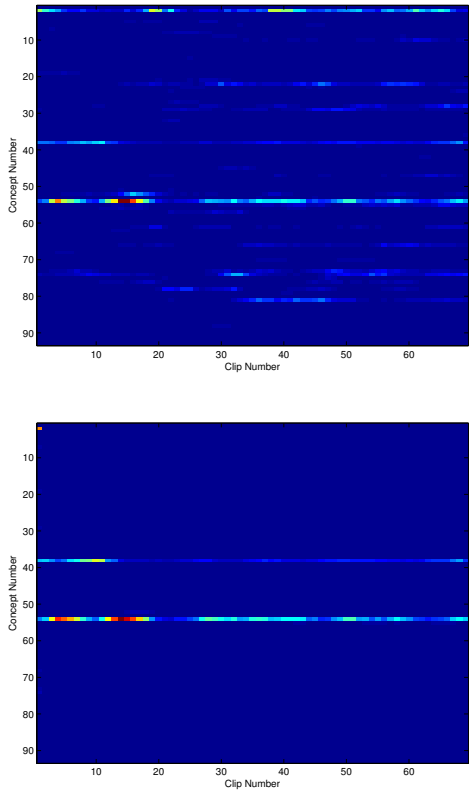


Figure 5. Probability sequences from SVM (left) and our method (right)

clips in each video which contain concepts and then calculate the accuracy on those clips. For both EC11 and EC12, we use 5-fold cross validation across all videos for each event and compute the causality matrices using 4 folds while testing on the remaining fold. Figure 6 shows results when applied to the EC11 dataset. We see that all but one event see improvement when causality is applied with at least one of the proposed methods. Figure 7 shows results when applied to the EC12 dataset. We see that 8 of the 10 events show improvement after incorporating causality. The SVM probabilities for each dataset were retrieved on classifiers that were trained on the same dataset, and thus our results are shown to be more accurate than expected. However this is true with our baseline as well, and thus an improvement on this dataset is still indicative of the usefulness of our approach. We also add small amounts of noise ($\pm 10\%$) in an effort to more accurately simulate probabilities that we would expect to see from a dataset that was not trained on itself. We find that 3 of the 25 events across EC11 and EC12 did not show improvement from the use of causality. They are *wedding*, *marriage proposal*, and *winning a race without a vehicle*. The events which showed the greatest improvement after applying causality were *grooming an animal* and *town hall meeting*. It is difficult to quantify how *causal* an event's concepts are. However, events with stronger relationships between concepts are expected to see greater improvements from our method. We also notice that the concepts for the two events with greatest improvement are more prone to being confused amongst themselves, and thus using causality aids in correcting the probabilities.

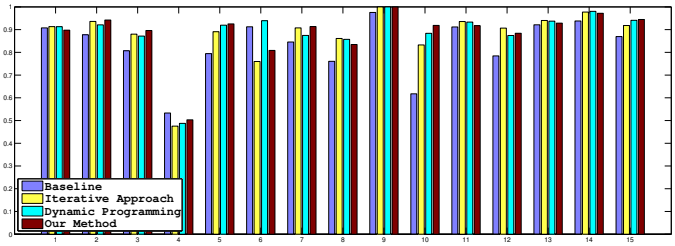


Figure 6. Results over the 15 events of the EC11 dataset

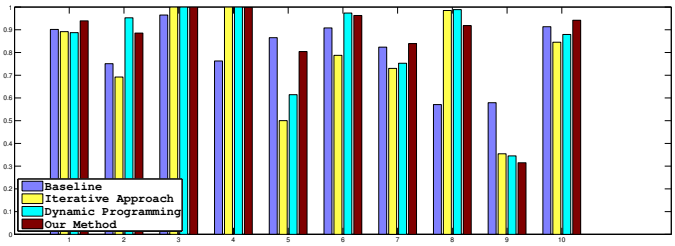


Figure 7. Results over the 10 events of the EC12 dataset

The final results are shown in Table I. We see that in each dataset, causality is able to improve the accuracy of concept detection through at least one of the proposed methods. We see that for EC11, the dynamic programming approach performs best while our method performs nearly as well, both resulting in over a 5% increase over the baseline. For EC12, our method performs much better than both other approaches, once again improving the accuracy by over 5% above the baseline. For our soccer dataset, our method once again performs best, giving an increase of almost 3% over the baseline using a 13-second interval for causality. We use 9-fold cross validation for the soccer dataset, creating the causality matrix using 8 of the 9 soccer videos and testing on the remaining video. We also use 9-fold cross validation in the same way to compute the SVM probabilities for this dataset. We would expect our accuracy to improve as we incorporate more information and data into our soccer dataset. However, these preliminary results show that causality is useful for improving concept detection for the event of soccer.

Table I. AVERAGE ACCURACY ACROSS ALL DATASETS FOR EACH METHOD

	Baseline	Iterative	Dynamic	Our Method
EC11	83.04%	87.56%	88.89%	88.57%
EC12	80.40%	77.86%	83.94%	86.04%
Soccer	21.77%	22.32%	21.48%	24.65%

VII. CONCLUSION

We proposed a method for improving concept detection which takes advantage of known temporal relationships between concepts. We introduced causality as a good measure for quantifying temporal relations and incorporated this into our graph-based approach to achieve increased accuracy for concept detection. We also applied iterative and dynamic programming approaches with varying levels of improvement over the baseline. We find that taking temporal relationships into account is a beneficial method of improving overall concept detection.

VIII. FUTURE WORK

Our method may be improved to detect non-concepts. Non-concepts are clips of a video that do not contain any concepts and are thus of little importance to many applications. As of current, the method is able to highlight areas where there are concepts and correctly identify which concept exists. However it is not always accurate in recognizing where concepts do not exist. Our method could also be applied to more datasets and tested further to more generally prove its accuracy. We also would have liked to approach the problem of concept detection using models such as second-order Markov chains and Conditional Random Fields. These models show promise for being able to utilize causality and are standard methods for incorporating temporal information into a model. There is also room for improvement with generating the causality matrix. Currently we are using one time step for each event. However we would like to explore using one time step for each relation between concepts to more accurately portray the complex relations within an event. We also would like to incorporate more features into the soccer dataset, such as MBH, MFCC, and SIFT. Currently we are using STIP features and are achieving an accuracy of 38% for action classification. Using MFCC features we achieve about a 20% accuracy. We would like to look into using post-fusion with these and other features to improve the baseline action classification before implementing our method.

A. Event Classification

Because our method is capable of generating an updated sequence of probabilities, we applied our method to the task of event classification. One approach to event classification is to create a vector of the maximum probabilities of each concept over each clip in a video and give this to an SVM that is trained to recognize videos as belonging to events in this way. Another approach is to use histograms of concepts/actions as the feature vector. We approach the task of event classification in both ways except with using our updated probability sequence. We train using SVM sequences that are generated using MBH features on the EC11 dataset and test on the DEVT1 dataset. We find that there is much potential for improvement in this area as it relates to our method, and that is one task which can examined further.

B. Hidden Markov Models

Hidden Markov Models, or HMMs, are advantageous to use when there are two sets of variables - observations and hidden variables [14], [15]. In order to apply them to our problem, we formulated our model such that the concepts were both the observations and the hidden states. The theory was that we could correct for confusion among concepts by using a Hidden Markov Model and incorporating causality. An HMM is defined by four sets of data - the hidden states, the observed states, the transition probabilities, and the observation probabilities. We set both sets of states to be the set of concepts. We set the transition probabilities to be those from the causality matrix, and we set the observation probabilities to

be those from a confusion matrix after applying our method. We then are able to give the HMM a sequence of concepts and use the Viterbi algorithm to return a new sequence of concepts that would be corrected for confusion and that incorporates the concept of causality. We found that this greatly reduces accuracy due to an over-reliance on causality. In order to correct this, we would have needed to alter the Viterbi algorithm to incorporate the SVM probabilities at each clip. However we instead focused on improving our own method. Using an HMM in this way would be a novel approach to correcting for confusion and is something that we could focus more on in the future.

REFERENCES

[1] A. E. A., M. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization," *IEEE TRANSACTIONS ON IMAGE PROCESSING*, vol. 12, no. 7, July 2003.

[2] C.-L. Huang, H.-C. Shih, and C.-Y. Chao, "Semantic analysis of soccer video using dynamic bayesian network," *IEEE TRANSACTIONS ON MULTIMEDIA*, vol. 8, no. 4, August 2006.

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

[4] M. Rodriguez, "Cram: Compact representation of actions in movies," in *CVPR*, 2010.

[5] J. Kwon and K. M. Lee, "A unified framework for event summarization and rare event detection," in *Computer Vision and Pattern Recognition (CVPR)*, 2012.

[6] K. Masumitsu and T. Echigo, "Video summarization using reinforcement learning in eigenspace," in *Image Processing*, vol. 2, 2000, pp. 267–270.

[7] B. Li and M. Sezan, "Event detection and summarization in sports video," in *Content-Based Access of Image and Video Libraries*, 2001, pp. 132–138.

[8] Z. Tu, D. Sun, and B. Luo, "Video summarization by robust low-rank subspace segmentation," in *Bio-Inspired Computing: Theories and Applications*, 2013.

[9] Y. Pritch, A. Rav-Acha, and S. Peleg, "Nonchronological video synopsis and indexing," in *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, vol. 30, no. 11, November 2008.

[10] Y. Pritch, S. Ratovitch, A. Hendel, and S. Peleg, "Clustered synopsis of surveillance video," *Advanced Video and Signal Based Surveillance*, 2009.

[11] Y. Pritch, A. Rav-Acha, A. Gutman, and S. Peleg, "Webcam synopsis: Peeking around the world," in *ICCV*, 2007.

[12] A. Rav-Acha, Y. Pritch, and S. Peleg, "Making a long video short: Dynamic video synopsis," in *CVPR*, 2006.

[13] S. J. D. Prince, *Computer Vision: Models, Learning, and Inference*. New York, NY: Cambridge University Press, 2013.

[14] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, February 1989.

[15] B. H. Juang and L. R. Rabiner, "Hidden markov models for speech recognition," *Technometrics*, vol. 33, no. 3, pp. 251–272, August 1991.