

Understanding Videos, Constructing Plots

Learning a Visually Grounded Storyline Model from Annotated Videos

Abhinav Gupta¹, Praveen Srinivasan², Jianbo Shi² and Larry S. Davis¹

¹University of Maryland, College Park, MD

²University of Pennsylvania, Philadelphia, PA

agupta@cs.umd.edu, psrin@seas.upenn.edu, jshi@cis.upenn.edu, lsd@cs.umd.edu

Abstract

Analyzing videos of human activities involves not only recognizing actions (typically based on their appearances), but also determining the story/plot of the video. The storyline of a video describes causal relationships between actions. Beyond recognition of individual actions, discovering causal relationships helps to better understand the semantic meaning of the activities. We present an approach to learn a visually grounded storyline model of videos directly from weakly labeled data. The storyline model is represented as an AND-OR graph, a structure that can compactly encode storyline variation across videos. The edges in the AND-OR graph correspond to causal relationships which are represented in terms of spatio-temporal constraints. We formulate an Integer Programming framework for action recognition and storyline extraction using the storyline model and visual groundings learned from training data.

1. Introduction

Human actions are (typically) defined by their appearances/motion characteristics and the complex and structured causal dependencies that relate them. These causal dependencies define the goals and intentions of the agents. The *storyline* of a video includes the actions that occur in that video and causal relationships [21] between them. A model that represents the set of storylines that can occur in a video corpus and the general causal relationships amongst actions in the video corpus is referred to as a “storyline model”. Storyline models also indicate the agents likely to perform various actions and the visual appearance of actions. A storyline model can be regarded as a (stochastic) grammar, whose language (individual storylines) represents potential plausible “explanations” of new videos in a domain. For example, in analysing a collection of surveillance videos of a traffic intersection scene, a plausible (incomplete) storyline-model is: *When a traffic light turns green traffic starts moving. If while traffic is moving, a pedestrian walks into an intersection, then the traffic suddenly stops. Otherwise it stops when signal turns red.* Not only are the actions “turns green”, “moving” and “walks” observable, there are causal relationships among the actions: traffic starts moving because a light turns green, but it stops because a pedestrian entered an intersection or signal turned

red. Beyond recognition of individual actions, understanding the causal relationships among them provides information about the semantic meaning of the activity in video - the entire set of actions is greater than the sum of the individual actions. The causal relationships are often represented in terms of spatio-temporal relationships between actions. These relationships provide semantic/spatio-temporal context useful for inference of the storyline and recognition of individual actions in subsequent, unannotated videos.

The representational mechanism of the storyline model is very important; traditional action recognition has heavily utilized graphical models, most commonly Dynamic Bayesian networks (DBNs). However, the fixed structure of such models (often encoded by a domain expert) severely limits the storylines that can be represented by the model. At each time step, only a fixed set of actions and agents are available to model the video, which is not sufficient for situations in which the numbers of agents and actions varies. For example, in sports, sequences of actions are governed by the rules of the game and the goals of players/teams. These rules and goals represent a structure that extends beyond a simple fixed structure of recurring events. The set of possible or probable actions and/or agents at any given time may vary substantially. An important contribution of this work is the introduction of AND-OR graphs [22, 26] as a representation mechanism for storyline models. In addition, unlike approaches where human experts design graphical models, we learn the structure and parameters of the graph from weakly labeled videos using linguistic annotations and visual data. Simultaneous learning of storyline models and appearance models of actions constrains the learning process and leads to improved visual appearance models. Finally, we show that the storyline model can be used as a contextual model for inference of the storyline and recognition of actions in new videos.

Our approach to modeling and learning storyline models of actions from weakly labeled data is summarized in Figure 1. The storyline models are represented by AND-OR graphs, where selections are made at OR-nodes to generate storyline variations. For example, in the AND-OR graph shown in the figure, the ‘pitching’ OR-node has two children ‘hit’ and ‘miss’ which represent two possibilities in the storyline, i.e after pitching either a ‘hit’ or a ‘miss’ can

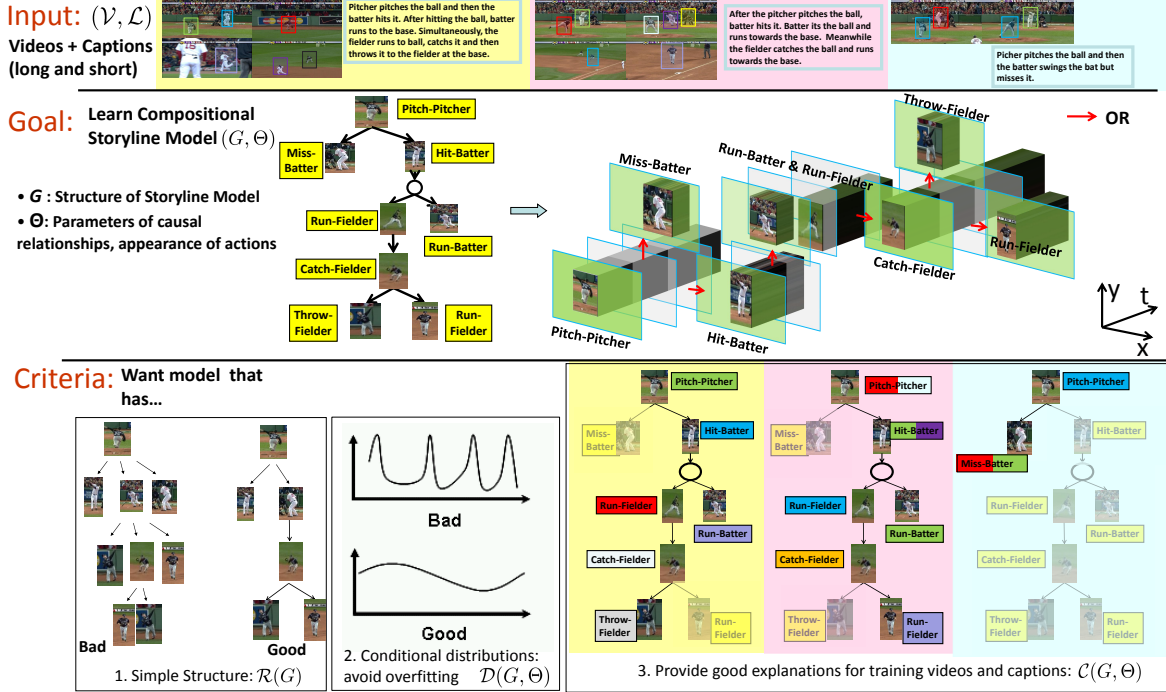


Figure 1. Visually Grounded Storyline-Model: Given annotated videos, we learn the storyline model and the visual grounding of each action. The optimization function for searching the storyline model has three terms: (1) Simple structure. (2) Connections based on simple conditional distributions. (3) Provides explanations for visual and text data in the training set. The figure also shows how our AND-OR graph can encode the variations in storylines (three videos at the top with different storylines (bottom-right)), not possible with graphical models like DBNs.

occur. The edges in the AND-OR graph represent causal relationships and are defined in terms of spatio-temporal constraints. For example, an edge from ‘catch’ to ‘throw’ indicates that ‘throw’ is causally dependent on ‘catch’ (a ball can be thrown only after it has been caught). This causal relationship can be defined in terms of time as $t_{catch} < t_{throw}$. The causal relationship has a spatial constraint also - someone typically throws to another agent at a different location.

Our goal is to learn the storyline model and the visual groundings of each action from the weakly labeled data - videos with captions. We exploit the fact that actions have temporal orderings and spatial relationships, and that many actions either “causally” influence or are “causally” dependent on other actions. Humans learn these “causal” relationships between different actions by utilizing sources of information including language, vision and direct experience (interaction with the world). In our approach, we utilize human generated linguistic annotations of videos to support learning of storyline models.

2. Related Work

Existing datasets for learning action appearance models provide samples for a few classes and in controlled and simplified settings. Such datasets fail to generalize to actions with large intra-class variations and are unsuitable for learning contextual models due to unnatural settings. There has been recent interest in utilizing large amounts of weakly labeled datasets, such as movies/TV shows in conjunction with scripts/subtitles. Approaches such as [6, 15] provide assignment of frames/faces to actions/names. Such

approaches regard assignment and appearance learning as separate process. Nitta et al. [20] present an approach to annotate sports videos by associating text to images based on previously specified knowledge of the game. In contrast we simultaneously learn a storyline model of the video corpus and match tracked humans in the videos to action verbs (i.e, solving the segmentation and correspondence problems).

Our approach is motivated by work in image annotation which typically model the joint distribution of images and keywords to learn keyword appearance models [1]. Similar models have been applied to video retrieval, where annotation words are actions instead of object names [7]. While such models exploit the co-occurrence of image features and keywords, they fail to exploit the overall structure in the video. Gupta et. al [12] presented an approach to simultaneously learn models of both nouns and prepositions from weakly labeled data. Visually grounded models of prepositions are used to learn a contextual model for improving labeling performance. However, spatial reasoning is performed independently for each image. Some spatial reasoning annotations in the images are not incidental and can be shared across most images in the dataset (For example, for all the images in the dataset sun is above water). In addition, the contextual model based on priors over possible relationship words restricts the clique size of the Bayesian network used for inference. Also, it requires a fully connected network, which can lead to intractable inference. In contrast, our approach learns a computationally tractable storyline model based on causal relationships that

generally hold in the given video domain.

There has been significant research in using contextual models for action recognition [11, 23, 10, 2, 18]. Much of this work has focused on the use of graphical models such as Hidden Markov Models (HMMs) [24], Dynamic Bayesian Networks (DBNs) [10] to model contextual relationships among actions. A drawback of these approaches is their fixed structure, defined by human experts. The set of probable actions and/or agents at any given time may vary greatly, so a fixed set of successor actions is insufficient. Our AND-OR graph storyline model can model both contextual relationships (like graphical models) while simultaneously modeling variation in structure (like grammars [2]).

In computer vision, AND-OR graphs have been used to represent compositional patterns [26, 4]. Zhu and Mumford [26] used AND-OR graph to represent a stochastic grammar of images. Zhu et. al [25] present an approach to learn AND-OR graphs for representing an object shape directly from weakly supervised data. Lin et. al [16] also used an AND-OR graph representation for modeling activity in an outdoor surveillance setting. While their approach assumes hand-labeled annotations of spatio-temporal relationships and AND-OR structure is provided, our approach learns the AND-OR graph structure and its parameters using text based captions. Furthermore, [16] assumes one-one correspondence between nodes and tracks as compared to one-many correspondence used in our approach.

Our work is similar in spirit to structure learning of Bayesian networks in [8], which proposed a structural-EM algorithm for combining the standard EM-algorithm for optimizing parameters with search over the Bayesian network structure. We also employ an iterative approach to search for parameters and structure. The structure search in [8] was over the space of possible edges given a fixed set of nodes. However, in our case, both the nodes and edges are unknown. This is because a node can occur more than once in a network, depending upon the context in which it occurs (See figure 2). Therefore, the search space is much larger than the one considered in [8].

3. Storyline Model

We model the storyline of a collection of videos as an AND-OR graph $G = (V_{and}, V_{or}, E)$. The graph has two types of nodes - OR-nodes, V_{or} and AND-nodes V_{and} . Each OR-node $v \in V_{or}$ represents an action which is described by its type and agent. Each action-type has a visual appearance model which provides visual grounding for OR-nodes. Each OR-node is connected to other OR-nodes either directly or via an AND-node. For example, in Fig 1, middle, the OR-node ‘Pitch’ has two OR-children which represents two possibilities after a pitch (i.e either the batter hits the ball (‘Hit-Batter’) or misses it (‘Miss-Batter’)). A path from an OR-node v_i to an OR-node v_j (directly or via an AND-node) represents the causal dependence of action v_j upon action v_i . Here, AND-nodes are dummy nodes and only used when an activity can causally influence two or more simultaneous activities. The causal relationships between two OR-nodes are defined by spatio-temporal constraints. For example, the causal relationship that ‘hitting’

depends on ‘pitching’ the ball can be defined temporally as $t_{pitch} < t_{hit}$ (hitting occurs after pitching) and spatially as the pitcher must be some distance d' from the batter $d(pitch, hit) \approx d'$. Figure 1 shows several examples (top) of videos whose actions are represented by AND-OR graphs. Note that the AND-OR graph can simultaneously capture both long and short duration storylines in a single structure (bottom-right).

4. Learning the Storyline Model

Our goal is to learn a visually grounded storyline model from weakly labeled data. Video annotations include names of actions in the videos and some subset of the temporal and spatial relationships between those actions. These relationships are provided by both spatial and temporal prepositions such as ‘before’, ‘left’ and ‘above’. Each temporal preposition is further modeled in terms of the relationships described in Allen’s Interval Logic. As part of bottom-up processing, we assume that each video has a set of human-tracks, some of which correspond to actions of interest. The feature vector that describes each track is based on appearance histograms of Spatio-Temporal Interest Points (STIPs) [14, 19] extracted from the videos.

Establishing causal relationships between actions and learning groundings of actions involves solving a matching problem. We need to know which human-tracks in the training videos match to different action-verbs of the storyline to learn their appearance models and the storyline-model of videos. However, matching of tracks to action-verbs and storyline extraction of a particular video depends on the structure of the storyline-model, the appearances of actions and causal relationships between them. This leads to yet another chicken-and-egg problem, and we employ a structural EM-like iterative approach to simultaneously learn the storyline-model and appearance models of actions from collections of annotated videos. Formally, we want to learn the structure G and parameters of the storyline model $\Theta = (\theta, A)$ (θ -Conditional Distributions, A -Appearance models), given the set of videos $(\mathcal{V}_1.. \mathcal{V}_n)$ and their associated annotations $(\mathcal{L}_1.. \mathcal{L}_n)$:

$$\begin{aligned} (G, \Theta) &= \arg \max_{G', \Theta'} P(G', \Theta' | \mathcal{V}_1.. \mathcal{V}_n, \mathcal{L}_1.. \mathcal{L}_n) \\ &\propto \arg \max_{G', \Theta'} \prod_i \sum_{M^i, S^i} P(\mathcal{V}_i, \mathcal{L}_i | G', \Theta', M^i, S^i) P(G', \Theta') \end{aligned}$$

- S^i : Storyline for video i .
- M^i : Matchings of tracks to actions for video i .

We treat both S and M as missing data and formulate an EM-approach. The prior, $P(G, \Theta)$, is based on simple structure $(\mathcal{R}(G))$ and simple conditional distributions terms $(\mathcal{D}(G, \Theta))$ and the likelihood terms are based on how well the storyline model generates storylines which can explain both the videos and their linguistic annotations $(\mathcal{C}(G, \Theta))$.

Figure 2 summarizes our approach for learning visually grounded storyline-models of training videos. Given an AND-OR graph structure at the beginning of an iteration, we fix the structure and iterate between learning parameters/visual grounding of the AND-OR graph and the

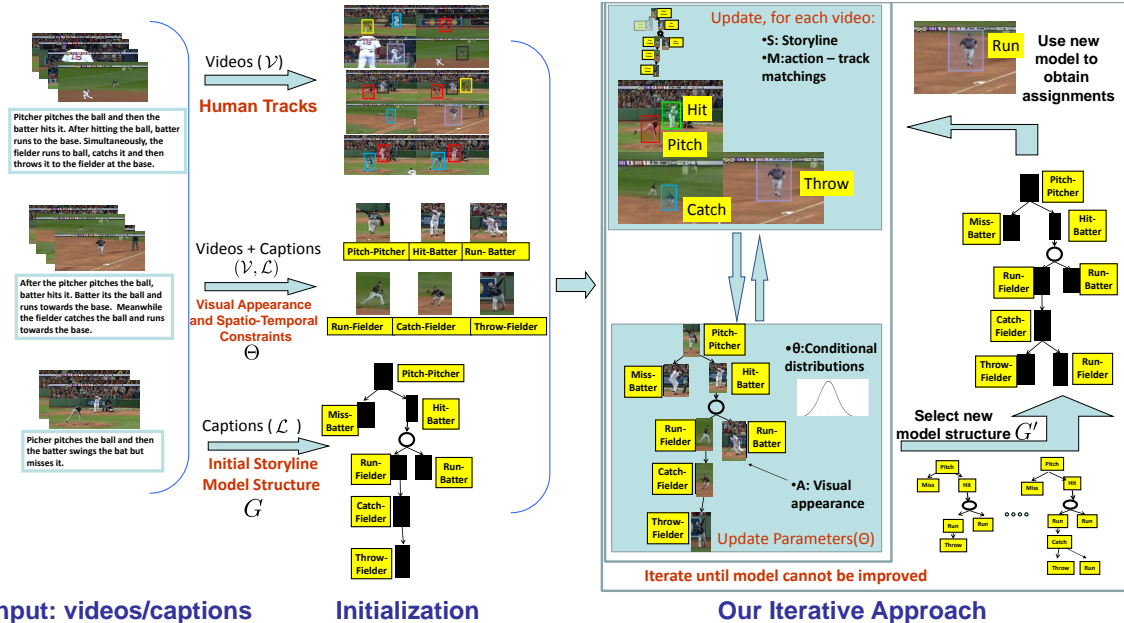


Figure 2. An Overview of our approach; our storyline model (G, Θ) is initialized using videos and captions, and we propose an iterative procedure to improve its parameters Θ and the structure G .

matching of tracks to action nodes(Sec. 4.1). In the hard-E step, we estimate storyline and matchings for all training videos using the current G, Θ . In the M step, we update Θ using the estimated storylines and matchings for all videos. After convergence or a few iterations, we generate new graph proposals by local modifications to the original graph (Sec. 4.2) and select the modification that best represents the set of storylines for the videos(Sec. 4.3). This new storyline model is then used for re-initializing the iterative approach, which iterates between appearances and matchings. The new storyline model, which is a better model of the variations in storylines across the training videos, allows better interpretation. For example, in the figure, the “run-fielder” action after the “catch” was labeled as “throw” since the initial storyline-model did not allow “run” after “catch”. Subsequently, an updated storyline model allows for “run” after “catching”, and the assignments improve because of the new expanded storyline.

4.1. Parsing Videos

We now describe how, an AND-OR storyline model is used to analyze, or parse, videos and obtain their storylines and matchings of human tracks to storyline actions. We provide a one-many matching formulation, where several human tracks can be matched to a single action. Matching of tracks to actions also requires making a selection at each OR-nodes to select one storyline out of the set of possible storylines. While there have been several heuristic inference algorithms for AND-OR graphs, we formulate an integer programming approach to obtain the storyline and matchings, and solve a relaxed version of the problem in the form of a linear program.

Given an AND-OR graph G , a valid instantiation, S (representing a storyline), of the AND-OR graph is a func-

tion $S : i \in V_{and} \cup V_{or} \rightarrow \{0, 1\}$ that obeys the following constraints: (1) At each OR-node v_i there is an associated variable S_i which represents whether the or-node has been selected for a particular storyline or not. For example, in fig 3, ‘hit’ is a part of the storyline, therefore $S_3 = 1$, and miss is not a part of the storyline so $S_2 = 0$. (2) Since OR-children represent alternate possible storyline extensions, exactly one child can be selected at each OR-node. (3) An OR-node, i , can be instantiated (i.e $S_i = 1$) only when all the OR-nodes in the unique path from the root to node i have been instantiated. For example, since the path from ‘pitching’ to ‘catching’ includes ‘hitting’, ‘catching’ can be part of a storyline if and only if ‘hitting’ is part of the storyline.

Given T human tracks in a video, a matching of tracks and nodes is a mapping $M : i \in V_{or}, j \in \{1, \dots, T + 1\} \rightarrow \{0, 1\}$. $M_{ij} = 1$ indicates that the action at the OR-node i is matched with track j . Since some of the actions might not be visible due to occlusion and camera-view, we add a dummy track which can be associated with any action with some penalty. Depending on the constraints imposed on M , different matchings between actions and tracks can be allowed: many-to-many, many-to-one, one-to-many, or one-to-one. We consider those mappings that associate one action to many tracks, which is represented by the constraint $1^T M = 1$. Furthermore, no tracks should be matched to an OR node that is not instantiated: $\forall i \in V_{or}, M_{ij} \leq S_i$.

Finally, to incorporate pairwise constraints (such as temporal ordering and spatial relationships) between matches of two nodes i and k , M_{ij} and M_{kl} , we introduce variables $X : x_{ijkl} \in \{0, 1\}$; $x_{ijkl} = 1$ indicates that the action at node i and track j are matched, and the action at node k and track l are matched. Instead of enforcing a computationally difficult hard constraint $x_{ijkl} = M_{ij} * M_{kl}$, we marginalize both sides over l and represent the constraint

$$\text{Parsing cost: } \mathcal{C}_V(S, M, X) = \mathcal{E}(M) + \mathcal{A}(S, M) + \mathcal{T}(X)$$

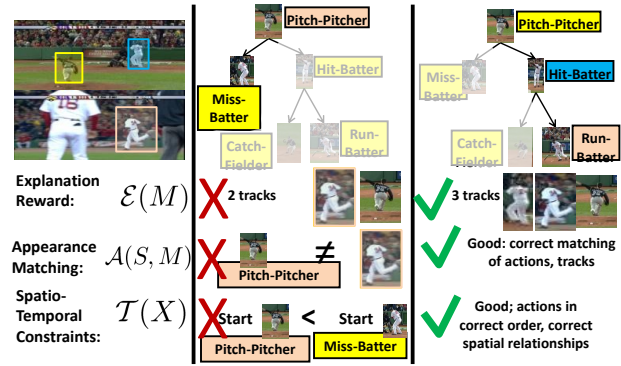


Figure 3. Given the upper left video, we show two possible parses and costs for each parsing cost function component. The left parse is worse since it explains fewer tracks (Explanation $\mathcal{E}(M)$), matches actions to tracks with different appearance (Appearance $\mathcal{A}(S, M)$), and violates spatial and temporal constraints (the pitcher-pitcher action should be matched to a track which occurs before the miss-batter track, and the two tracks should appear in the usual pitcher-batter configuration) (Spatio -Temporal $\mathcal{T}(X)$). The correct parse on the right scores well according to all three cost function components.

$$\text{as: } \forall k, \sum_l x_{ijkl} = M_{ij}.$$

In parsing, we search for a “best” valid instantiation S (representing a storyline from the storyline model) and a matching M of tracks and actions (representing visual grounding of nodes). The optimization function for selection is based on three terms: (1) Select a storyline consisting of nodes for which good matching tracks can be found. (2) Select a storyline which can explain as many human tracks as possible. (3) The matching of nodes to tracks should not violate spatio-temporal constraints defined by the storyline model (See Figure 3). The three terms that form the basis of the objective to be minimized, subject to the above constraints on S , M and X , are:

Appearance Matching: The cost of a matching is based on the similarity of the appearances of instantiated nodes and corresponding tracks. This cost can be written as:

$$\mathcal{A}(S, M) = \sum_i \left| \left(\sum_j M_{ij} \right) t_j - S_i A_i \right| \quad (1)$$

where t_j represents the appearance histogram of track j and A_i represents the appearance histogram model of the action at node i . In many to one matching, multiple tracks combine to match to a single action node. Therefore, the first term, $(\sum_j M_{ij} t_j)$, sums the appearance histograms of human tracks that match to node i . This is then compared to the appearance model at node i by measuring the L1-norm. Figure 3 shows an example of parsing with high(left parse) and low matching costs(right parse). The left parse has high matching cost since the track of a batter running is assigned to the pitching node which are not similar in appearance.

Explanation Reward: Using only appearance matching would cause the optimization algorithm to prefer small storylines, since they require less matching. To remove this bias, we introduce a reward, \mathcal{E} , for explaining as many of the STIPs as possible. We compute \mathcal{E} as:

$$\mathcal{E}(M) = - \sum_j \min \left(\sum_i M_{ij}, 1 \right) \|t_j\| \quad (2)$$

This term computes the number of tracks/STIPs that have been assigned to a node in the AND-OR graph and therefore explained by the storyline model.

Spatio-Temporal Constraints: We also penalize matchings which violate spatio-temporal constraints imposed by causal relationships. If p_{ijkl} encodes the violation cost of having an incompatible pair of matches (node i to track j and node k to track l), the term for spatio-temporal violation cost is represented as: $\mathcal{T}(X) = \sum_{ijkl} p_{ijkl} x_{ijkl}$. This term prefers matchings that do not violate the spatio-temporal constraints imposed by the learned AND-OR graph. For example, the left parse in Figure 3 matches the ‘pitching’ and ‘miss’ actions to incorrect tracks, resulting in ‘pitching’ starting after ‘battering’ in the video, which is physically impossible. The tracks are also not in the typical pitcher-batter spatial configuration. Therefore, this matching has a high cost as compared to the matching shown in the right parse.

The above objective and the constraints result in an Integer Program which is a NP-Hard problem. We approximate the solution by relaxing the variables S , M and X to lie in $[0, 1]$. The result is a linear program, which can be solved very quickly. For the learning procedure, we have the annotated list of actions that occur in the video. We utilize these annotations to obtain a valid instantiation/storyline S and then optimize the function over M, X only. For inference, given a new video with no annotations, we simultaneously optimize the objective over S, M, X .

4.2. Generating new Storyline Model Proposals

After every few inner iterations of the algorithm, we search for a better storyline model to explain the matchings and causal-relationships between actions. To do this, we generate new graph proposals based on local modifications to the AND-OR graph structure from the previous iteration.

The local modifications are: (1) Deletion of an edge and adding a new edge (2) Adding a new edge (3) Adding a new node. The selection of edges to delete and add is random and based on the importance sampling procedure, where deletion of important edges are avoided and addition of an important edge is preferred. The importance is defined on the basis of the likelihood that the head and tail of the edge are related by a causal relationship.

4.3. Selecting the New Storyline Model

Each iteration selects the AND-OR graph from the set of modifications which best represents the storylines of the training videos. The criteria for selection is based on four different terms:

Track Matching Likelihood: The first criteria measures how well a proposal explains the matchings obtained in the previous parsing step. The matching of tracks to actions from the previous step is used to obtain a likelihood of an AND-OR graph generating such a matching. The likelihood of the p^{th} graph proposal, G_p^p generating the pairwise matchings X^{r-1} (at iteration $r - 1$) is given by

$\frac{1}{Z} \exp(-\mathcal{T}_{G^p}(X^{r-1}))$. This likelihood is based on the third term from the parsing cost, but here the penalty terms are computed with respect to the individual graph proposals.

Annotation Likelihood: The AND-OR graph representing the storyline model should not only explain the matching of tracks to actions, but also the linguistic annotations associated with each video. The underlying idea is that the same storyline model is used to generate the visual data and linguistic annotations. The cost function measures how likely an instantiation of the AND-OR graph storyline model accounts for the video’s actions annotations and how well the constraints specified by linguistic prepositions in annotations are satisfied by the AND-OR graph constraints. For example, if the annotation for a training video includes ‘pitching before hitting’, a good AND-OR graph would not only generate a storyline including ‘pitching’ and ‘hitting’ but also have the conditional distribution for the edge $\text{pitching} \rightarrow \text{hitting}$, such that $P(t_{hit} - t_{pitch} > 0 | \theta)$ is high.

Structure Complexity If we only consider likelihoods based on linguistic and visual data, more complex graphs which represent large numbers of possibilities will always be preferred over simple graphs. Therefore, an important criteria for selection of an AND-OR graph is that it should be simple. This provides a prior over the space of possible structures. We use a simplicity prior similar to [9], which prefers linear chains over non-linear structures.

Distribution Complexity The complexity of an AND-OR graph depends not only on its graph structure, but also the conditional distributions of children actions given parent actions. For an action i (OR-node) in an AND-OR graph, we form a distribution over all possible successors, or sets of actions that could appear immediately after action i in a storyline. The individual spatio-temporal conditional distributions between i and its successors are combined into a single distribution over successors, and we compute the entropy of this combined distribution. The entropies of the successor distributions for all OR-nodes in the graph are averaged, providing a measure of the complexity of the conditional distributions contained in the AND-OR graph. Our cost prefers higher entropy distributions; empirically, we have found that this results in better ranking of structures. We can also draw intuition from work on maximum entropy Markov models [17], where higher entropy distributions are preferred in learning conditional distributions to prevent overfitting.

4.4. Initializing the Search

For initialization, we need some plausible AND-OR causal graph to represent the storyline model and appearance models of actions. Establishing a causal sequence of actions from passive visual data is a difficult problem. While one can establish a statistical association between two variables X and Y , inferring causality - whether $X \rightarrow Y$ or $Y \rightarrow X$ - is difficult. For initialization, we use the linguistic annotations of the videos. Based on psychological studies of causal learning, we use ‘time’ as a cue to generate the initial storyline model [13]. If an action A immediately precedes action B , then A is more likely to be the cause and B is more likely to be the effect.

We initialize the AND-OR graph with the minimum number of nodes required to represent all the actions in the annotations of the training videos. Some actions might have more than one node due to their multiple occurrences in the same video and due to different contexts under which the action occur. For example, ‘catch-fielder’ can occur in a video under two different contexts. The action ‘catching’ in the outfield and ‘catching’ at a base are different and require different nodes in the AND-OR graph. Using Allen’s interval temporal logic, we obtain the weight of all possible edges in the graph, which are then selected in a greedy manner such that there is no cycle in the graph. Dummy AND-nodes are then inserted by predicting the likelihood of two activities occurring simultaneously in a video.

For initialization of appearance models, we use the approach proposed in [12]. Using the spatio-temporal reasoning based on the prepositions and the co-occurrence of visual features, we obtain a one-one matching of tracks to actions which is used to learn the initial appearance models.

5. Experimental Evaluation

For our dataset, we manually chose video clips of a wide variety of individual plays from a set of baseball DVDs for the 2007 World Series and processed them as follows: We first detect humans using the human detector[5]. Applied to each frame with a low detection threshold, the output of the detector is a set of detection windows which potentially contain humans. To create tracks, we perform agglomerative clustering of these detection windows over time, comparing windows in nearby frames according to the distance between their centroids, and similarity of color histograms as measured by the Chi-square distance. The resulting tracks can be improved by extending each track forwards and backwards in time using color histogram matching. STIPs that fall within the detection window of a track in a frame contribute to the track’s appearance histogram.

Training: We trained the storyline model on 39 videos (individual baseball plays), consisting of approximately 8000 frames. The training videos contained both very short and very long plays. We evaluate the performance of our training algorithm in terms of number of actions correctly matched to tracks. Figure 4 shows how this accuracy changed over the training process. The figure is divided into three colored blocks. Within each colored block, the structure of the storyline model remains the same and the approach iterates between parsing and parameter update. At the end of each colored block, we update our storyline model and select a new storyline model which is then used to parse videos and estimate parameters. We can see that the accuracy rises significantly over the course of training, well above the initial baselines, validating our iterative approach to training. The percentage improvement over Gupta et. al [12] is as much as 10%.

Figure 5 a) shows an example of how a parse for a video improves with iterations. Figure 5 b) shows an additional example of a video with its inferred storyline and matchings of actions to tracks; we can see that all but the run-fielder action are correctly matched.

Storyline Extraction for New Videos: Our test set in-

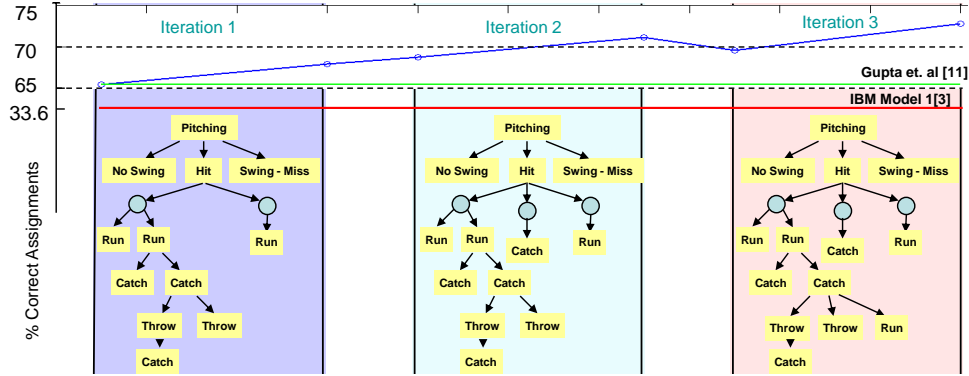


Figure 4. Quantitative evaluation of training performance and how the storyline model changes with iterations. Within each colored block, the storyline model remains fixed and the algorithm iterates between parsing and parameter estimation. At the end of each colored block, the structure of the AND-OR graph is modified and a new structure is learned. The structural changes for the three iterations are shown.

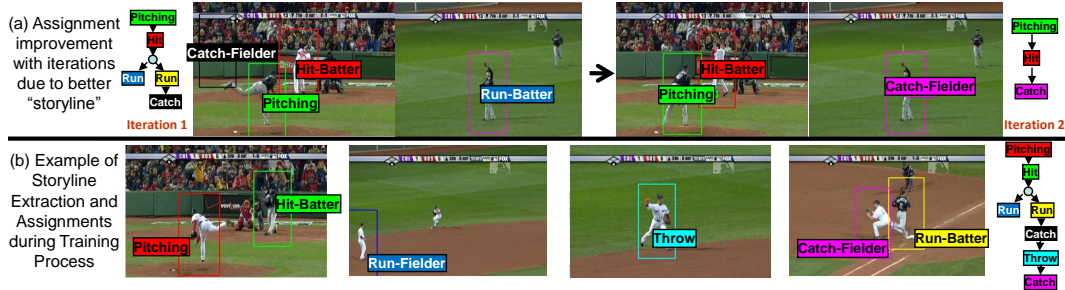


Figure 5. (a) Improvement in assignments with iterations: In the first iteration, the true storyline for the video pitching \rightarrow hit \rightarrow catching, is not a valid instantiation of the AND-OR graph. The closest plausible storyline involves activities like run which have to be hallucinated in order to obtain assignments. However, as the storyline model improves in iteration 2, the true storyline now becomes a valid storyline and is used for obtaining the assignments. (b) Another example of the assignments obtained in training. The assignments are shown by color-coding, each track is associated to the node which has similar color in the instantiated AND-OR graph.

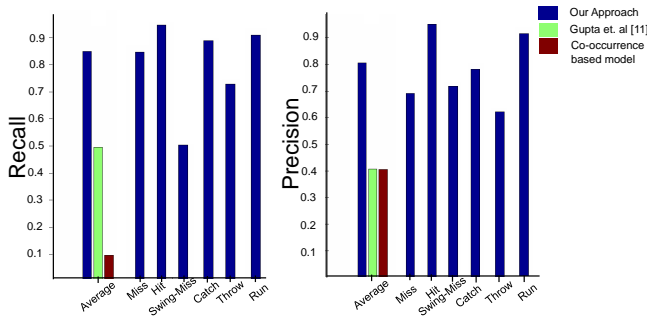


Figure 6. Quantitative Evaluation of Labeling Accuracy

cludes 42 videos from the same baseball corpus. Again, they ranged from very short and simple to longer and more complicated. We first evaluated the performance in terms of storyline extraction. Fig.7 shows some qualitative examples of the storyline extraction in terms of the instantiation of AND-OR graphs, assignment of tracks to actions and the text that is generated from the storyline model. We use recall and precision values of action labeling to measure the performance of storyline extraction. We compare the performance of our approach to the baseline methods of Gupta et.al [12] and IBM Model 1[3]. Figure 6 shows two bar plots, one for recall (left) and the other for precision (right). For the baseline methods, we show the average precision and recall values and compare against our method’s performance (block of blue, red and green bars). Our method

nearly doubles the precision of the baseline methods (.8 vs. .4), and has a much higher recall (.85 vs. 0.5 for [12] and 0.1 for [3]). It performs well over most of the actions, with the exception of the action Swing-Miss (low recall). We also evaluated the number of correct matchings obtained for the actions in the predicted storylines. Quantitatively, we obtained 70% correct assignments of tracks to actions.

We attribute the success of our approach to three reasons: (1) An important reason for improvement in training compared to Gupta et. al [12] is that they did not feedback the contextual models learned at the end of their single iterative loop of training to relearning models of object appearances. (2) During inference, the coupling of actions via the AND-OR graph model provides a more structured model than simple context from co-occurrence statistics and binary relationship words can provide. (3) The one-many (action to track matching) framework used here is more powerful than the one-one framework in [12] and handles the problem of fragmented segmentation.

6. Conclusion

We proposed the use of storyline model, which represents the set of actions and causal relationships between those actions. Our contributions include: (1) Representation of storyline model as an AND-OR graph whose compositional nature allows for compact encoding of substantial storyline variation across training videos. (2) We also pre-

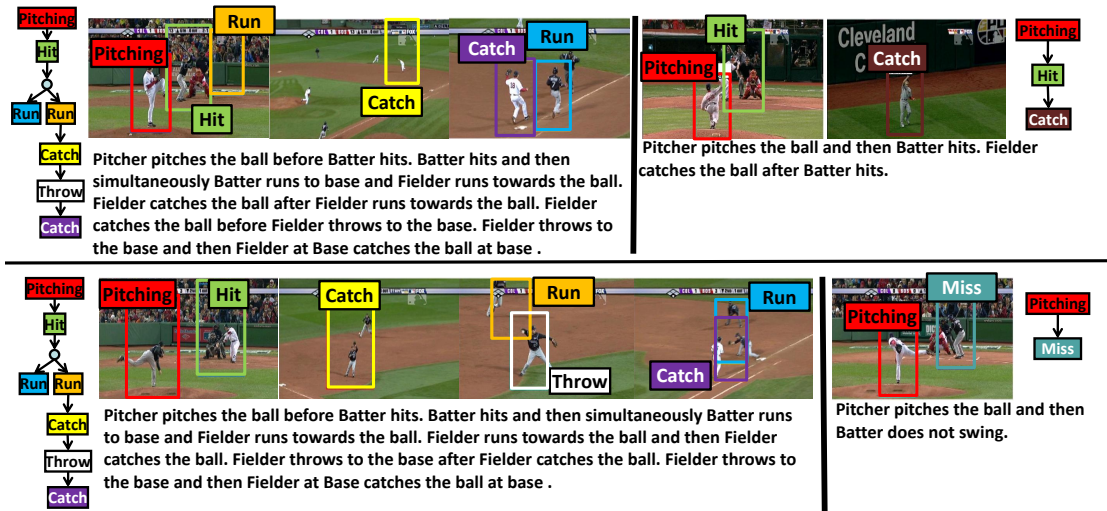


Figure 7. Storyline Extraction for New Videos: We show the instantiation of AND-OR graph obtained for each training video and the story generated in text by our algorithm. The assignments of tracks to action nodes are shown by color coding.

sented a method for learning storyline models from weakly annotated videos. (3) We formulate the parsing problem as a linear integer program. Our formulation permits one-to-many matching of actions to video tracks, addressing the problem of fragmented bottom-up segmentation. Experimental results show that harnessing the structure of videos helps in better assignment of tracks to action during training. Furthermore, coupling of actions into a structured model provides a richer contextual model that significantly outperformed two baselines that utilize priors based on co-occurrence and relationships words.

Acknowledgement: This research was funded by US Governments VACE program, NSF-IIS-0447953 and NSF-IIS-0803538. Praveen Srinivasan was partially supported by an NSF Graduate Fellowship.

References

- [1] K. Barnard, P. Duygulu, N. Freitas, D. Forsyth, D. Blei, and M. I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, pages 1107–1135, 2003.
- [2] A. Bobick and Y. Ivanov. Action recognition using probabilistic parsing. *CVPR98*.
- [3] P. Brown, S. Pietra, V. Pietra, and R. Mercer. The mathematics of statistical machine translation: parameter estimation. *Comp. Linguistics*, 1993.
- [4] H. Chen, Z. Jian, Z. Liu, and S. Zhu. Composite templates for cloth modeling and sketching. *CVPR*, 2006.
- [5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *CVPR*, 2005.
- [6] M. Everingham, J. Sivic, and A. Zisserman. Hello! my name is...buffy - automatic naming of characters in tv video. *BMVC*, 2006.
- [7] M. Fleischman and D. Roy. Situated models of meaning for sports video retrieval. *Human Language Tech.*, 2007.
- [8] N. Friedman. The bayesian structural em algorithm. *UAI*, 1998.
- [9] N. Friedman and D. Koller. Being bayesian about network structure: A bayesian approach to structure discovery in bayesian networks. *Machine Learning*, 2003.
- [10] S. Gong and T. Xiang. Recognition of group activities using dynamic probabilistic networks. *ICCV*, 2003.
- [11] A. Gupta and L. Davis. Objects in action: An approach for combining action understanding and object perception. *CVPR'07*.
- [12] A. Gupta and L. Davis. Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. *ECCV*, 2008.
- [13] D. Laganado and A. Solman. Time as a guide to cause. *J. of Exper. Psychology: Learning Memory & Cognition*, 2006.
- [14] I. Laptev. On space-time interest points. *IJCV*, 2005.
- [15] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. *CVPR'08*.
- [16] L. Lin, H. Gong, L. Li, and L. Wang. Semantic event representation and recognition using syntactic attribute graph grammar. *PRL*, 2009.
- [17] A. McCallum, D. Freitag, and F. Pereira. Maximum entropy markov models for information extraction and segmentation. *ICML*, 2000.
- [18] C. Needham, P. Santos, R. Magee, V. Devin, D. Hogg, and A. Cohn. Protocols from perceptual observations. *AI'05*.
- [19] J. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatio-temporal words. *BMVC*, 2006.
- [20] N. Nitta, N. Babaguchi, and T. Kitahashi. Extracting actors, actions an events from sports video - a fundamental approach to story tracking. *ICPR*, 2000.
- [21] J. Pearl. Causality: Models, reasoning, and inference. *Cambridge University Press*, 2000.
- [22] J. Pearl. Heuristics: Intelligent search strategies for computer problem solving. *Addison-Wesley*, 1984.
- [23] S. Tran and L. Davis. Visual event modeling and recognition using markov logic networks. *ECCV'08*.
- [24] A. Wilson and A. Bobick. Parametric hidden markov models for gesture recognition. *PAMI*, 1999.
- [25] L. Zhu, C. Lin, H. Huang, Y. Chen, and A. Yuille. Unsupervised structure learning: Hierarchical recursive composition, suspicious coincidence, competitive exclusion. *ECCV'08*.
- [26] S. Zhu and D. Mumford. A stochastic grammar of images. *Foundations and Trends in Comp. Graphics and Vision*, 2006.