Using Probabilistic Ontologies for Video Exploration

Miguel Bustamante
Computer Science and Engineering, State University of New York at Buffalo, Buffalo, NY, United States., mbustama7@hotmail.com

Jason Corso
Computer Science and Engineering, State University of New York at Buffalo, Buffalo, NY, United States., jcorso@buffalo.edu

Follow this and additional works at: http://aisel.aisnet.org/amcis2012

Recommended Citation

This material is brought to you by the Americas Conference on Information Systems (AMCIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in AMCIS 2012 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.
Using Probabilistic Ontologies for Video Exploration

Miguel A. Bustamante (author contact)  
Department of Computer Science and Engineering  
State University of New York at Buffalo  
miguelbu@buffalo.edu

Jason J. Corso  
Department of Computer Science and Engineering  
State University of New York at Buffalo  
jcorso@buffalo.edu

ABSTRACT

Video data is being collected at alarming rates and yet there exists no comprehensive forensic toolset that enables the analyst to quickly examine video in the context of the massive collections. This research builds a System that studies video at a semantic level by means of a joint solution to semantic entity extraction, entity-entity relationship extraction, and dynamic event recognition. The working of the System is grounded in formal ontology. This ontology is jointly induced from the data and established by the human domain experts (i.e., interactive machine learning). Specifically, we implement a Multi Entity Bayesian Network (a form of a probabilistic ontology); we test our System on two-on-two basketball game videos, and our results demonstrate state of the art detection rates on activities like passing the ball, and shooting, consequently promising that the presented methodology is an encouraging direction for semantically rich video analysis.

Keywords
Ontology, Probabilistic ontology, video semantic analyses, Multi Entity Bayesian Network.

INTRODUCTION

How does a person access the content of a 700-page book? Without reading the whole book, he or she usually goes to the table of contents to find which parts of the book suit his or her needs. If he or she has a specific issue in mind, such as finding a term, the person can go to the index at the end of the book and find the corresponding pages addressing that issue. (Xiong, Zhou, Tian, Rui and Huang, 2006). Similarly, we have the same questions with video data. But, discerning the informational content of video is difficult; so, procedures are needed for automatically (or semi-automatically) constructing video table of contents, video highlights and video indices to facilitate summarization, browsing and retrieval. (Xiong et al., 2006).

To these ends this paper uses Multi Entity Bayesian Networks (MEBN), a form of probabilistic ontology, and video analysis methods from computer vision (identification and tracking) to build a computational model that can represent video data and provide accurate responses to queries. Specifically, there has never been a study (that we are aware of) incorporating ontology based reasoning into video analysis. Most studies that do use ontologies in processing have largely been in language processing e.g., the Semantic Web (Maedche and Staab, 2001).

The study of video data is complex. It requires the transduction of the original high-dimensional video signal into a high-level description of its content, including entities, their inter-relationships, and how they form dynamic events in the video. To deal with this complexity our paper, using MEBN, presents a joint solution to semantic entity extraction, entity-entity relationship extraction, and dynamic event recognition.

To assess our approach our paper shows an implementation of an MEBN for the two-on-two basketball game video domain. Our research tests our system on real video data. We instantiate our probabilistic ontology for primitive activities like holding the ball and compound activities such as shooting and found state of the art activity detection rates; our findings proves that the proposed framework can be used as an initial step to build a toolset that will allow users to study video data in real time.

The remainder of the paper is organized as follows: section 2 provides an overview of ontologies, and probabilistic ontologies. Section 3, describes MEBNs. Section 4, presents a particular MEBN developed for video analysis using our operational research model (to study two-on-two basketball game videos). Section Five shows the research results. And finally, the paper concludes with a summary and discussion section.
ONTOLOGIES, AND PROBABILISTIC ONTOLOGIES

Ontologies
An ontology specifies a particular vocabulary for representing entities and relationships characterizing a domain. Ontologies facilitate interoperability by standardizing terminology, allowing automated tools to use the stored data in a context-aware fashion, enabling intelligent software agents to perform better knowledge management, and providing other benefits of formalized semantics. (Laskey, Wright and Costa, 2010). In this paper our ontology will describe the types of elements (entities), and the multi-entity relationships (events) from the video data.

Probabilistic Ontologies
Probabilistic ontologies were born because of the necessity to include uncertainty in ontologies. Probabilistic ontologies are used for the purpose of widely describing knowledge about a domain and the uncertainty associated with that knowledge. This description should be structured and sharable; and furthermore it should be made in a format that can be read and processed by a computer. (Laskey, 2008; Costa and Laskey, 2005).

Laskey et al., 2010 Present a clear and concise definition: A probabilistic ontology is an explicit, formal knowledge representation that expresses knowledge about a domain of application. This includes:

a. Types of entities that exist in the domain;
b. Properties of those entities;
c. Relationships among entities;
d. Processes and events that happen with those entities;
e. Statistical regularities that characterize the domain;
f. Inconclusive, ambiguous, incomplete, unreliable, and dissonant knowledge related to entities of the domain; and
g. Uncertainty about all the above forms of knowledge;

where the term entity refers to any concept (real or fictitious, concrete or abstract) that can be described and reasoned about within the domain of application.

MULTI ENTITY BAYESIAN NETWORKS (MEBNS)
Multi Entity Bayesian Networks (MEBN) extend Bayesian Networks (BN); its logic combines the flexibility of Bayesian Networks with the representational power of First-Order Logic. MEBN represents a particular domain as composed of entities that have attributes and are related to other entities. The knowledge about the attributes of entities and their relationships to each other is represented as a collection of MEBN Fragments (MFrags), which are organized into MEBN Theories (MTheories). (Laskey, Costa and Janssen, 2008). An analogy between a finite automaton definition an a MFragment definition is presented in Figure 1.
There are three types of random variables (RVs): Resident, Input, and Context. A MFragment represents a conditional probability distribution of the instances of its Resident (RVs) given the values of instances of their parents in the fragment graphs and given the context constraints. RVs are graphically represented in a MFragment either as resident nodes, which have distributions defined in their home fragment, or as input nodes, which have distributions defined in another place. (Laskey et al., 2008) Context nodes are the third type of MFragment nodes, and represent conditions assumed for definition of the local distributions. Usually, MFrags are small, since their main purpose is to model little parts of domain knowledge that can be reused when needed. This feature is helpful when modeling complex, intricate situations and is one that can be seen as the knowledge representation version of the divide and conquer paradigm for decision-making; (Laskey et al., 2008) the main idea is that MFrags split a difficult decision problem in a set of smaller ones. This decomposition is accomplished by modeling a complex situation as a collection of small MFrags, each representing some specific element of a simpler situation. In reality, MFrags provide a flexible means to represent knowledge about specific subjects within the domain of discourse, but the true gain in expressive power is revealed when aggregating these "knowledge patterns" to form a coherent model of the domain in study that can be instantiated to reason about specific situations and polished through learning; this group of MFrags is called an MTheory, and it represents a joint probability distribution for a number of instances of its random variables. This joint distribution is specified implicitly through the local and default distributions inside each MFragment, together with the conditional independence relationships implied by the fragment graphs. A MTheory summarizes statistical regularities that describe a domain. These regularities are captured and programmed in a knowledge base using some combination of expert judgment and learning from observation. A MTheory let the analyst reason about particular scenarios, he or she needs to provide the system with specific information about the individual entity instances involved in the scenario. After having this information, Bayesian inference can be used both to answer specific questions of interest (e.g., how likely is it that a particular black car that enter the garage is planning to perform a specific attack?) And to improve the MTheory (e.g., each new scenario includes additional statistical data about the possibility of a given attack for that set of circumstances). In a elegant way Bayesian inference can be used to perform both problem specific inference and learning. (Laskey et al., 2008)

**RESEARCH METHODOLOGY**

**Research Model**

Recall our goal is to build and to implement a system that can optimally represent, construct and evaluate video data while providing (probabilistic) accurate responses to queries. We will explore the use of probabilistic ontologies of semantic entities models based on Multi Entity Bayesian Networks to build such a system. The operational research model is presented in Figure 2.
Comments on the operational research model

Fig 2, our operational research model, summarizes the steps needed to analyze a video.

In a first stage, (arrow which goes form video input to box on the center) the analyst and user decide which questions the model is expected to answer; which information is needed to answer the questions (evidence); and the entities, relations and rules that apply to them in our particular domain.

In a second stage, the transformation from video to text takes place. This stage includes: the detection of the geometrical position (coordinates) of our entities; the transformation of our coordinates data according to our domain's rules; and the consolidation of data by unit of analysis. The output of this stage is our GROUND TRUTH; the data that feeds our situation specific Bayesian networks.

In a third stage, the building of the MEBN takes place. The output of this stage is our INFERENCE; a collection of situation specific Bayesian networks that are stored in our Complex and Dynamic Database (CXDB).

Finally, to answer queries the system uses a situation specific Bayesian network and the GROUND TRUTH.

NOTE: in the next paragraphs, we explain our research model using an specific MEBN developed for video analysis of two-on-two (2-on-2) basketball game videos.
Specific MEBN developed for video analysis using our operational research model to study two-on-two basketball game videos

We use two-on-two (2-on-2) basketball game videos to test our system. The game is a variation of standard basketball in which four players (two each on two teams) use half of the court and compete under the same hoop. 2-on-2 presents a convincing and yet tractable subject matter for our inquiry: the collection of entities is simple enough to model and report on. Yet, the ways in which they interact remains rich enough to test the representational and inferential capabilities of the system. For example, in 2-on-2 there are only four people on the court, but their roles are constantly switching between offense and defense depending on the recent play and ball possession.

Given a video of a 2-on-2 basketball game our goal is to specify when a primitive activity -e.g., holding the ball, ball is bouncing, etc.- is taking place with a confidence (or probability) and between which frames. These activities are directly observable from the data with static detectors, using various computer vision techniques.

Compound Activities To specify when a compound activity -a sequence of primitive activities, e.g., passing, shooting, etc.- is taking place with a confidence (or probability) and between which frames. These activities are directly observable from the data, but require some temporal sequence of events to happen that can be modeled with techniques in pattern recognition; the probabilistic ontology provides a language to capture such temporal sequences of events.

Although these represent a small part of possible inferences for a full system, these would clearly permit high-level semantic video indexing and afford rich analyses of the video content, such as team possession and behavior metrics.

Data set

For our study, we use the OSUPEL Basketball data set created in (Brendel, Fern, and Todorovic, 2011). The videos have been acquired with a camcorder on a tripod at 960 x 540 pixels at 29fps.

The two-on-two videos in the data set (videos 4 to 7) contain a total of 21 minutes 56 seconds or 38146 frames. Each frame in the data set is annotated with one of the following 10 action labels: passing, catching, holding ball, shooting, jumping, dribbling, trajectory (the ball is in moving in the air), contact, bouncing, and near rim. In our study, we use videos 5 to 7 for training and reserve video 4 for testing. We have selected action labels shooting, passing, catching, holding ball, trajectory, near-rim, open [Dunkin] (jointly dribbling, contact, and bouncing) for our study, joining three of them into one open group to keep the study tractable.

Capturing frames from the video

Since our domain demands the decoding (understanding) of video frames, we base on high-level models on the following geometrical observations made on the video frames. First, semi-interactive methods from computer vision have been used to provide detections and tracks of the core entities in the scene: namely the players, the hoop-rim, and the ball. In each case, an entity \( e \) is represented as a sequence of tuples \( (x,y,w,h) \) with \( x,y \) being the upper-left hand corner of the entity's bounding box, \( w,h \) are the width and height of the box. See Figure 3 captioning frames from the video.
Second, we further reduce the observation space into temporal units of analysis based on the relative configuration of the ball and the other objects in the scene. We use the ball here because it signifies the dominant activity factor in 2-on-2 basketball. At each frame, we intersect the bounding box of the ball with the players and the hoop-rim. Essentially, this means that the relationship between the ball and the environment is fixed for a given unit of analysis. Our modeling is based entirely on sequential units of analysis. See Figure 4 example of unit of analysis.

Implementing Model

We have followed and present our model in sequence following the Probabilistic Ontology Modeling Cycle (Carvalho, Laskey, Costa, Ladeira, Santos, and Matsumoto, 2010) based on the goals set in the previous section, (i) define entities, attributes, and relations, (ii) isolate elements of uncertainty and define appropriate MFrags to represent them, (iii) define/learn local probability distribution functions for resident random variables and (iv) finalize the knowledge base by populating it with instances. We separately discuss the structure of the probabilistic ontology and the instantiation of it to support queries in our domain. As an aside, we note the convenience of the graphical representations of the MEBN to circumvent a need to belabor the discussion with complex probability and logic equations, which are all directly available from the figures.
Entities
We define four entities for our case study: Ball, Team, Player, and Rim-Hoop. As stated earlier, we present a conveniently simple set of entities that interact in a comparatively complex manner.

Uncertainty and MFrags
The relationship between the ball and the players on the court is the primary source of uncertainty in our test. We model our MFrags based on this relationship and subsequently capture the primitive and composite actions instances of the MFrags. Some of our MFrags are presented in Figure 5 and explained below; we do not include singleton MFrags, which are often referred to as “support” or “service” in the literature and are the input variables that do not appear as resident in other MFrags, such as Team(p)

PlayerOnCourt
Parameterized by a player, the input variables represent the teams at play and their rosters (i.e., team A, ) has players A1 and A2 and not B1 and B2. The resident random variable establishes if this player is in play for a specific team.

BallIntersectsPlayer
Parameterized by the ball and a player, this MFrag (the resident random variable within) geometrically relates the ball’s position in the video frame to the player’s position. The local distribution is directly computed by the intersection vector

BallOnAir
To support reasoning about passing, shooting and catching, this MFrag models the probability that the ball is not being held by a player; i.e., the ball is in a trajectory.
HoldingBall
Analogously, the resident variable in this MFrag captures if a player $p$ is holding the ball.

Passing
This MFrag captures the compound activity of passing, taking BallOnAir and HoldingBall as input variables and computing the probability that the player $p$ is now passing.

Catching
Following (and conditioned on) passing, this MFrag models if player $p$ has passed to player $q$ and $q$ is catching it.

Instantiation
Recall the MEBN definition of a probabilistic ontology defines a parameterized set of Bayesian network fragments that can be instantiated to model events in the domain. In the MEBN/PR-OWL literature this instantiation is recorded in the knowledge base (Costa and Laskey, 2006; Carvalho et al., 2010). In our case, we instantiate two teams, A and B, four players, A1, A2, B1, B2, a single hoop, and a single ball. In addition, we query (events) situation specific Bayesian networks (SSBN) (Laskey, 2008) to model specific primitive and compound activities. Each SSBN instantiates a set of MFrags by assigning specific instances to its arguments. For example, we can instantiate the event Shooting Mfrag for each of the four players see Figure 6.

This SSBN case is the most complex in our system since it involves a long sequence of interdependent primitive actions (captured as instantiations of their respective MFrags). For example, the player is identified and in play via instantiated PlayerOnCourt MFrag, and the ball NearRim MFrag explains the relation between the ball and the rim. For the shooting activity to have a high probability, the HoldingBall resident variable needs to be True for this player and the NearRim resident variables also needs to be True. For each SSBN, the local probability distributions are learned as a combination of domain constraints followed by maximum likelihood parameter estimation on labeled training examples (in our data set, videos 5-7). During query resolution and video analysis processing, the observed variables from each unit of analysis are iteratively run through the SSBNs, resulting in updates to the local distributions of all resident variables in all SSBNs and hence updates to the current state of the world. See figure 6.
Our full model has been implemented in the UnBBayes software environment (Carvalho et al., 2010) and the ontology stored in a PR-OWL language (Costa et al., 2006). The system is able to report on specific queries about the activities in the domain and generate various summary and detailed reports, which we describe in the next section.

**RESEARCH RESULTS**

Using our implemented MEBN, we evaluate its capabilities to respond to queries on the two activity types (primitive and compound) in the 2-on-2 basketball problem. For evaluation, we use the video 4 from the OSUPEL 2-on-2 basketball dataset, which has 7427 frames. (Brendel et al., 2011). Our first processing step extracts a total of 385 relevant units of analysis on which our processing is based.
As a first experiment, we have implemented the system to index (classify) six primitive and compound activities shooting, passing, catching, holding the ball, trajectory, and near rim; the remaining three activities are grouped into an open class dunkin and no classifier is trained on them. We show the results of the system on the test video via a confusion matrix in Figure 7.

For example, in this confusion matrix, of the twenty-three actual shooting events, the system predicted that four were Ball-Trajectory, and of the twenty-seven Near-Rim events, it predicted that one was a shooting, six were Ball-Trajectory, and one was Dunkin. We can see from the matrix that all correct guesses are located in the diagonal of the table, so it’s easy to visually inspect the table for errors, as they will be represented by any non-zero values outside.

The confusion table shows that the system is capable of achieving or exceeding state of the art classification scores for five of the six activities. (See Brendel et al., 2011 for state of the art classification scores). Limited success is observed on only the ball trajectory activity, where it is most often mistaken for holding ball, which suggests the simple intersecting box feature is insufficient to let the MEBN/SSBN distinguish between whether the ball is being held or is in the air.

These classification results demonstrate the system can suitably index domain-specific activities at a unit-level. We are also able to generate detailed reports see Figure 8 which show further information to the user about the content of a particular video such as when and how confident a specific activity occurred.

<table>
<thead>
<tr>
<th>Actual Class</th>
<th>Predicted class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shooting</td>
<td>Passing</td>
</tr>
<tr>
<td>Shooting</td>
<td>19</td>
</tr>
<tr>
<td>Passing</td>
<td>35</td>
</tr>
<tr>
<td>Catching</td>
<td>3</td>
</tr>
<tr>
<td>Holding_the_Ball</td>
<td>3</td>
</tr>
<tr>
<td>Ball-Trajectory</td>
<td>4</td>
</tr>
<tr>
<td>Near-Rim</td>
<td>1</td>
</tr>
<tr>
<td>Dunkin</td>
<td>10</td>
</tr>
<tr>
<td>Total = 23</td>
<td>Total = 63</td>
</tr>
</tbody>
</table>

Figure 7. Confusion Matrix

<table>
<thead>
<tr>
<th>Detailed Report: By Event: SHOOTING</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instance</td>
</tr>
<tr>
<td>----------</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>11</td>
</tr>
<tr>
<td>12</td>
</tr>
<tr>
<td>13</td>
</tr>
<tr>
<td>14</td>
</tr>
<tr>
<td>15</td>
</tr>
<tr>
<td>16</td>
</tr>
<tr>
<td>17</td>
</tr>
<tr>
<td>18</td>
</tr>
<tr>
<td>19</td>
</tr>
<tr>
<td>20</td>
</tr>
<tr>
<td>21</td>
</tr>
<tr>
<td>22</td>
</tr>
<tr>
<td>23</td>
</tr>
</tbody>
</table>

Figure 8. Detailed Report for Shooting Event
Our system is also able to generate sequences of keyframes from the videos to visually describe the activity. For example, we present two sets of such extracted keyframes sequences from a correct shooting activity (Figure 9) and an incorrect one (Figure 10). In this case the incorrect detection is fired because the ball flies near enough to the rim to trigger the shooting SSBN.

Figure 9. Detailed Report for Shooting Event (correct shooting activity)
Figure 10. Detailed Report for Shooting Event (incorrect shooting activity)
SUMMARY AND DISCUSSION

This paper describes a coherent, comprehensive probabilistic ontology framework based on Multi Entity Bayesian Networks, that provides a means of optimally represent, construct and evaluate video data while providing (probabilistic) accurate responses to queries. The proposed framework is an initial step towards a more comprehensive effort focused on creating a friendly toolset that will allow analysts to investigate video data in real time.
REFERENCES


