# Ontology and Taxonomy Collaborated Framework for Meeting Classification

Asaad Hakeem and Mubarak Shah
School of Computer Science
University of Central Florida, Orlando, FL 32816
{ahakeem,shah}@cs.ucf.edu

## Abstract

*A framework for classification of meeting videos is proposed in this paper. Our goal is to utilize this framework to analyze human motion data to perform automatic meeting classification. We use a rule-based system and state machine to analyze the videos, utilize three levels of context hierarchy, namely movements (and their attributes), events(actions), and behavior to identify the activities and classify the meeting type based on the meeting ontology. We also define a meeting ontology that is determined by the knowledge base of various meeting sequences. This ontology validates and refines the taxonomy based on the hierarchy of events and behaviors, and regroups similar meetings in one category, refining the classes. Ontology is the process of determining the class of a meeting video based on relationships, and taxonomy is the categorization of meetings based on a certain criteria. The rule-based system is the primary framework manager, which recognizes behaviors based on the events detected by the state machine. It also periodically rolls back the state machine from erroneous state-space to a stable state. The state machine detects the events using a sliding temporal window of human movements. Our approach is appropriate for classifying meetings in complex sequences involving various actions and partial occlusion of tracked objects. Our framework is unique and scalable, with the capability to add new meeting types to the framework with little or no modification to the current framework. Using our framework, we are able to correctly classify various meeting sequences such as voting, argument, presentation, and object passing in our experiments. This framework is applicable to automated video surveillance, video segmentation and retrieval (multimedia), human computer interaction, augmented reality, etc. Our framework can also be used as a model for different areas of high-level vision understanding.*

# 1. Introduction

Human activity recognition is an important area in the field of computer vision that has applications like video surveillance, human computer interaction and augmented reality. A lot of research has been done ranging from low-level tracking, to medium-level segmentation and action recognition, and finally high-level semantics, activity recognition and language generation to understand the videos. Various researchers have used different methodologies for action and activity recognition. These include Hidden Markov Models, Finite State Machines, Rule-based systems and using their own taxonomy of actions and activities. Unfortunately, most of these works take a small section of a problem and solve it using a specific technique. These techniques are chosen based on a narrow scope of application for solving the problem. There is no framework, as such, that defines a methodology for solving the problem and keeping the broader view in perspective. Our goal in this paper is to propose such a framework for classification. We chose meeting videos like voting, discussion, arguments and presentation for classification, as they have many applications in surveillance and are relatively complex involving multiple agents and their interactions.

In order to detect the events, we analyze the temporal data in a sliding temporal window. Using the relative positions of hands and head along with the movement attributes, for a certain time-frame, we are able to correctly identify the underlying event using the state machine. In order to properly determine the events, we built an ontology for different events based on the movements. This event ontology is further extended to recognize behaviors and finally genres from behaviors. These ontologies are determined by the knowledge base of various meeting sequences. The main emphasis in this paper is the ontology and taxonomy framework for activity recognition, and we implement this framework using state machines and rule-based expert system, although other implementations such as SCFG and Bayesian Networks can also be used with our framework.

We use three levels of context hierarchy, namely movements (and its attributes), events, and behavior to identify the activities and classify the meeting type based on the meeting ontology. The rule-based system recognizes behaviors based on the events detected by the state machine. It also periodically rolls back the state machine from erroneous state-space to a stable state. We need a state machine for detection of an event based on the movements, since an event is a defined sequence of movements. In order to detect the behavior, we need the rule-based system, as we cannot model behaviors using state machines, due to the interaction of different persons, each having their own state machine. This forms a disjoint set of state machines and we need a rule-based system to coordinate and analyze the behavior, the roles of different persons and classify the meeting video.

The rest of the paper is organized in six sections. Section 2 deals with the related work in activity recognition and various taxonomies and methodologies for single and multi-agent tracking. Several approaches are discussed in this section along with their limitations. Section 3 describes our proposed framework in detail. In this section we provide the framework for video classification continuing with the meeting example, suggest a meeting ontology, describe how the state machine works with the rule-based system to correctly identify the movements, events and behaviors, and classify the meeting videos. Section 4 provides the experimental results for meeting classification, and summarizes the results using a table of ground truth and actual results. Section 5 details the current limitations of the framework and techniques used; and also the suggested future work.

## 2. Related Work

Aggarwal and Cai [1] proposed an overview of Human Motion Analysis, where they described the different methods of tracking and recognition of motion. The various tracking methods involve the use of single or multiple cameras, with point, blob or volume tracking; and state-space or template matching, with point, line, blob or mesh recognition. It also involved an overview of the 2D and 3D approaches for motion analysis with or without a priori shape models, and tracking without body parts.

Kojima and Tamura [5] use a concept hierarchy of action rules called case frames, to determine an action grammar for the sequence of events and generate a sentence from the actions performed. Their method works well for single person action recognition in a concept hierarchy using a case framework. The case frame is a kind of frame expression used in natural language processing and consists of 8 categories of cases like agent, object, locus, source, etc. Our method is their counterpart of single person activity recognition, where we use a different approach to classify meetings that involve multiple people.

Intille and Bobick [6], [7] talk about closed-world tracking with their object taxonomy of precise, approximate and amorphous objects. The objects are detected based on context-specific features (domain knowledge) forming a template. The tracking is continued based on the correlation between the object and its template. These tracked results are input into the multi-agent belief network, which is a framework for recognizing multi-agent action from visual evidence. It is based on the probabilistic maximization of a belief, based on the evidence provided by the tracked data. Final recognition is based on the combination of rules attaining a goal, which is the maximized result or the actual recognition.

Hongeng and Nevatia [8] use a state machine to model the tracking of events, also making use of graphical notation for multi-agent event classification. They use a ground plane assumption and Kolmogorov-Smirnov statistics for region merging. Their approach is based on statistical training data and requires large datasets.

Ayers and Shah [2] use a state machine for action recognition and use that data for key frame extraction by taking a snapshot of the frame on the occurrence of an event. Their model tracks motion in specified areas, which restricts the system and requires a priori knowledge of the environment. Badler [17] proposed a four level hierarchy to define *motion verbs* which we term as *behavior*. He used state graphs and primitive rules on artificial environments with static images to describe these motion verbs. His methodology was interesting for that time, but it had limitation to sampling rate with no error correction techniques. Also, since his method was implemented on artificial environment where complete knowledge of system was available, he used that knowledge to resolve complex events rather than using movement data for event detection.

Jebara and Pentland [4] utilize time-series of perceptual measures and predict the action using Conditional Expectation Maximization (CEM), and synthesize a reaction based on the predicted action. Tracking is done by skin detection and Expectation Maximization (EM). Yacoob and Black [10] track cyclic human motion using their parameterized model. The recognition is done through eigenspace warping of the observed data to the model data using Principal Component Analysis (PCA) methodology. Their methodology is limited to recognition of activities consisting of repeated patterns. Davis and Bobick [11] track the human movement using temporal templates. They use a combination of Motion Energy Image (MEI) and a scalar valued Motion History Image (MHI) to construct temporal templates. This method is view specific and sensitive to partial occlusions as the motion of the trained temporal template will not have the occluded MEI and its associated MHI. Hidden Markov Models (HMMs) [12, 13, 9] or its variations such as Coupled Hidden Markov Models (CHMMs) [14] and Layered Hidden Markov Models (LHMMs) [15] have been widely used in the area of action and activity recognition. Ivanov *et al.* [9] uses Stochastic Context Free Grammar (SCFG) that is a non-deterministic probabilistic expansion of context free grammar (CFG) to recursively look for a completed tree in the grammar, hence recognizing the activity. It then generates a language of the detected events, giving the sentential description of activities. Wilson *et al.* [12] rightly points out the inherent weakness of the Hidden Markov Models, in the fact that they need a huge training data of varying actions and events in the spatio-temporal domain that could occur in the sequences. A slight variation in the spatio-temporal data may confuse the model and recognition will fail. For example, if the training data was not translation or scale variant, then change of spatial positioning of user will not be recognized by the trained model.

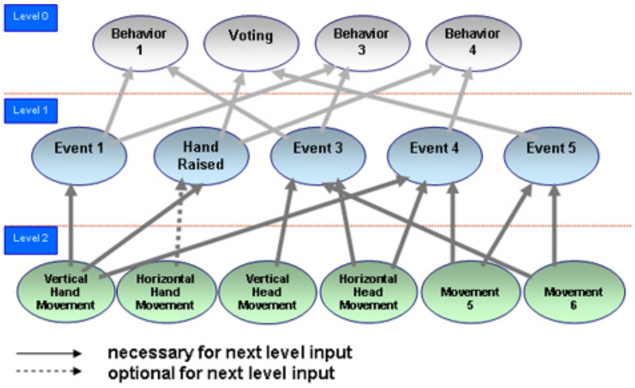

Figure 1: Voting behavior example for elaborating multi-level meeting ontology.

Some of the above methods require either a specific number of people for training the system, or tracking is done with non-occluding entities, or require large training data for better prediction and recognition. Retraining the system when a new meeting type is added to it might not be feasible in case of systems which are trained on a large number of videos - whereas using our modular approach will be feasible, with just adding relevant data to the knowledge-base. Our approach is very systematic, where we use the ontology of a particular behavior from coarse movement data, to refined detection of events and recognition of behavior and finally genre (if present). This ontology refines the taxonomy, which categorizes the different types of behaviors based on the similarities in them, provided by the ontology. Once the framework of ontology and taxonomy is ready, it is mapped to an implementation, which in our case is state machine and rule-based system.

# 3. Video Classification

The methodology involved in our framework includes defining a taxonomy of meetings based on person-to-person interactions for the initial taxonomy. Then using observations on different types of meetings (knowledge base) we construct an ontology of meetings. This ontology is used to refine the meeting taxonomy, categorizing similar behaviors under the same class of meetings. Once the refinement process is complete, the state machines and rules for the rule-based system are mapped from the event and behavior ontology respectively. New sequences involving varying activities are detected and tracked using CONDENSATION [3] and input to the system, which recognizes these activities and outputs the analysis results and meeting classification. Our framework is generic and can be applied to various areas like action grammar generation and recognition of other activities like sports, violence detection, crowd behavior recognition, etc. Using this methodology, create the ontology and taxonomy of the above mentioned areas and then map them to the state machine and rule-based system (see implementation section for details). The meeting ontology is based on a multi-level context hierarchy that is described in detail in the next section.

## 3.1. Multi-Level Classification using Context Hierarchy

In order to determine the meeting type we need to define a multi-level context hierarchy for meeting ontology; from a low level *movement* of hand or head to a medium level *event* using a sequence of movements and its attributes, to the high level *behavior* consisting of different events combined to depict a behavior. Explanations of the different levels of the context hierarchy are as follows:

*Movement* ($\delta$): The *movements* are determined using low level tracking data. We consider movement of human body parts such as hands and head for analysis of meeting videos. These movements have attributes such as position, displacement, direction, and speed. Examples of movements include vertical movement with large displacement, and position of hand above the head.

*Event* ($\alpha$): An *event* is a single action based on the movement of the head and hands, and their attributes. These are also called *actions*, and constitute a unique sequence of movements over time within the sliding temporal window. The event is detected by the state machine for that particular action in our framework. Examples of events include hand raised, hand lowered, pick up object, and putdown object.

*Behavior* ($\beta$): The correct semantics of an event can be determined by checking the other events in the neighborhood of the sliding temporal window, to determine the higher level *behavior*. This is also called *activity*, and is defined by a sequence of events over time that form a unique set. The behavior depicts the main theme of the meeting in the absence of genre. The behavior is recognized by the rule-based system in our framework and classification starts at the behavior level. For example, if a person raises his hand for a few seconds and then lowers it, that sequence of two events form a voting behavior. If a hand is raised longer than the voting threshold, then it is not a voting behavior. In that case we look for other events present to determine the behavior. Other examples of behaviors include hand shaking, discussion, and argument.

*Genre* ($\gamma$): A *genre* is a superset of meeting activities consisting of more than one behavior and forming a unique set in a meeting sequence. For example, a presentation is a genre of meeting consisting of behaviors like questions, discussion, and object passing. The genre is classified by the rule-based system in our framework. The presence of a behavior is necessary for meeting classification, whereas that of genre is optional.

To explain the proposed framework we present a pictorial representation with an example of voting behavior in Figure 1. Voting consists of horizontal (optional) and vertical hand movements, raising the hand above the bottom of the head, leaving it there for a short time, then lowering it, in sequence. The combination of these movements form the events hand raised (above head) and hand lowered. The voting behavior is recognized if the hand was above the lower head bounds for greater than a minimum amount of time and less than a maximum amount. Otherwise it can be confused by other behaviors like question or waving. These movements and events are tracked for each individual and are analyzed by the rule-based system for behavior recognition and genre classification. In the absence of genre, the recognized behavior is classified as the meeting type. The next sections describe the ontology and taxonomy of meetings depicting the methodology for building a framework to classify the meetings. The section ends with an implementation of the framework showing how ontology and taxonomy are mapped to an algorithm.

## 3.2. Meeting Ontology and Classification

The ontology for classification of meeting sequences uses a hierarchical framework of defining relationships between *movements* $\delta$ to form *events* $\alpha$, that have relationships with each other to form *behaviors* $\beta$. The different behaviors combine to form *genres* $\gamma$ and the voting ontology is shown in Figure 1. The movements in the horizontal and vertical directions are given meaning at a higher level based on conditions to generate a hand raised event. This is further refined by checking different events in the context hierarchy to determine the behavior. This whole process of refining and finding relationships between different levels of the context hierarchy is called voting ontology. Similarly ontologies of different events and behaviors accumulate to form the meeting ontology. A table detailing the list of movements, events and behaviors observed and used for activity recognition and meeting classification is given in Table 2.

## 3.3. Taxonomy of Meetings

The taxonomy of meetings is important for understanding the categorization of meeting classes. It specifies how categories in a particular level are linked with the lower level. We used the person-to-person interaction criteria for the taxonomy, and there are four types of meetings:

*One-to-One*: This type of meeting refers to a *single* person interacting with another person. Examples include informal meeting, and interviews.

*One-to-Many*: This type of meeting refers to a *single* person interacting with *multiple* persons. This single person acts as a moderator like announcer, conductor, or lecturer. This taxonomy does not portray that the 'one' will initiate the many, rather it depicts the 'one' acting as the moderator. Examples of this meeting type include presentation, announcement, and board meeting.

*Many-to-One*: This type of meeting refers to a *multiple* people interacting with a *single* person. Examples of this meeting type include interview by multiple interviewers, and hiring of a person by a panel of people.

*Many-to-Many*: This type of meeting involves *multiple* people interacting with *another group* of people. Examples of this type include gang fights, mob and crowd behaviors, and their interactions. The taxonomy does not expand beyond the behavior, since the ontology recognizes the behavior from the lower levels of events and movements. The taxonomy gives the higher level classification of genre depending upon the behaviors observed in the meeting.

| **Movements** | **Events** | **Behavior** |
|---|---|---|
| **Entities:**<br>head, left hand,<br>right hand<br>**Attributes:**<br>-displacement (none, small,medium,large)<br>-position (x,y)<br>-direction (left, right,up,down)<br>-speed (none,slow, medium,fast) | hand raised<br>hand lowered<br>pick up object<br>put down object<br>hand extended<br>hand retracted<br>hand waving<br>moderator present<br>hand pointing<br>head shakes/nods | hand shaking<br>object passing<br>greeting<br>congratulate<br>question answer<br>discussion<br>argument<br>decision<br>voting |

Table 2: List of movements, events and behavior.

## 3.4. Implementation of the Framework

The event ontology is mapped into different state machines, whereas the behavior ontologies map to the rules of the rule-based system for behavior recognition. The genre is determined by the sequence of behaviors present in the taxonomy for the matching meeting category. For example, if the meeting is between two people, who greet each other by hand waving, and discuss something during that meeting; the system matches the informal meeting type by taxonomy look-up in one-to-one person interaction category. The *rule-based system* is the primary framework manager, with a set of rules in a hierarchy that determines behavior based on the events. The events are detected by a *state machine* on the observance of movements. If the state machine goes into an invalid state, the rule-based system rolls it back to the last valid state and ignores the current observation. This method involves insertion, deletion and substitution error recovery. For example, if the observation string was 'abaaba' of observation movements 'a' and 'b', and if the correct event was 'ababa' where the extra 'a' was a false observation (insertion), then the system would ignore that observation and update the event history list. If no event exists after ignoring the observed data, then it would ignore all the observed data for the sliding temporal window and roll back the state machine to the start state. Even though there might be a loss of an event, the multi-level nature of the system is flexible enough to handle this case and is able to recognize the behavior even in the absence of a few events. We need a rule-based system for the final meeting classification based on the recognized behaviors and genres, as there are multiple persons involved and modelling it using a state-machine and classifying the meeting type is not possible.

# 4. Experiments and Results

We conducted experiments for recognizing three types of meeting behaviors and one genre using 15 different meeting sequences, each having frames ranging from 150 to 2500. All the sequences were unconstrained, where the people in the meetings were free to do different combinations of various actions at varying pace and positioning in meetings. The activities in the sequences were correctly recognized with proper classification of meeting types by the framework, even in the presence of crowding and occlusion, showing its robustness. We also tested our system on PETS dataset, with successful results.

| **No. of Frames** | **Events Detected** | **Ground Truth** | **Events Missed** | **False Positive** | **Classification** |
|---|---|---|---|---|---|
| 272 | 10 | 10 | 0 | 0 | Argument |
| 311 | 12 | 11 | 0 | 1 | Argument |
| 330 | 11 | 12 | 1 | 0 | Argument |
| 161 | 8 | 8 | 0 | 0 | Object Passing |
| 187 | 5 | 5 | 0 | 0 | Object Passing |
| 184 | 3 | 3 | 0 | 0 | Voting |
| 165 | 4 | 4 | 0 | 0 | Voting |
| 247 | 11 | 9 | 0 | 2 | Voting |
| 342 | 10 | 11 | 1 | 0 | Voting |
| 2153 | 25 | 22 | 0 | 3 | Voting |
| 2441 | 22 | 27 | 5 | 0 | Argument+Voting |

Table 3: Summary of Results.

We now show the events detected and behaviors classified in the different meeting videos by the system during the experiments. Figure 2 shows the frames of voting sequences, that were analyzed and classified correctly. The summary of different experiments involving various meeting sequences are shown in Table 3. Since we used a hierarchical approach we were able to classify all the meetings correctly, even with missing or incorrectly detected events as those events recur and missing a few of them would not make our system fail. If the event does not recur then it is an unimportant event, and is not necessary for meeting classification.

# 5. Limitations and Future Work

The system has the limitation of initializing the bounding box of the human head and hands (for the CONDENSATION tracking), which can be corrected with enhancement in techniques in computer vision to better detect human body parts. Also the current techniques like template matching, contour matching, heuristic based, etc. are not robust enough for all situations; for example not view invariant, and color dependent, hence we did not use those techniques. Also, when a body part (head or hand) is completely occluded or merged and emerges or splits the tracking system loses track of it or incorrectly labels the tracks and we need to manually intervene and correct the situation.

The future work could use speech information for getting better results and making the system less prone to error. It would detect the number of people, the moderator in the meeting, and recognize the activities involving sounds like clapping and thumping. Addition of facial gesture recognition could help in better understanding the state of the people (happy, sad, angry, etc.) and can further improve the system's accuracy. Finally, we could use a view invariant representation and action recognition [16], using spatio-temporal curvatures, making our framework more robust.

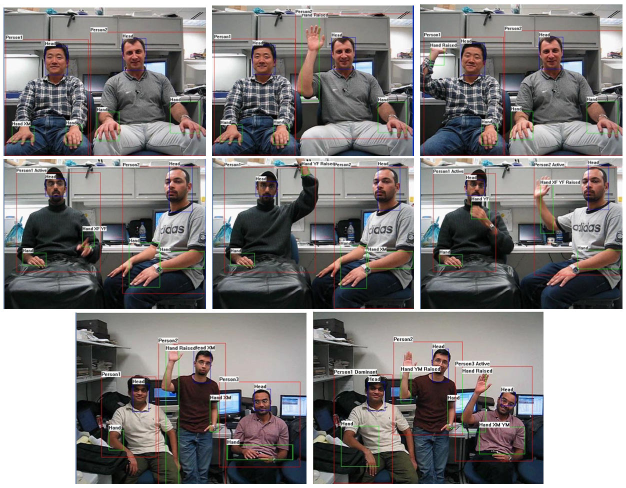

Figure 2: Sequence of frames for various Voting videos.

# References

[1] J. K. Aggarwal, Q. Cai, "Human Motion Analysis: A Review," *Computer Vision and Image Understanding*, Vol. 73, pp. 428-440, 1999.

[2] D. Ayers, M. Shah, "Monitoring Human Behavior from Video Taken in an Office Environment," *Image and Vision Computing*, Vol. 19, pp. 833-846, 2001.

[3] M. Isard, A. Blake, "A Mixed-State CONDENSATION Tracker with Automatic Model-Switching," *International Conference on Computer Vision*, pp. 107-112, 1998.

[4] T. Jebara, A. Pentland, "Action Reaction Learning: Analysis and Synthesis of Human Behavior," *IEEE Workshop on the Interpretation of Visual Motion*, 1998.

[5] A. Kojima, T. Tamura, "Natural Language Description of Human Activities from Video Images Based on Concept Hierarchy Actions," *International Journal of Computer Vision*, Vol. 50, pp. 171-184, 2001.

[6] S. S. Intille, A. F. Bobick, "Closed-World Tracking," *International Conference on Computer Vision*, pp. 672-678, 1995.

[7] S. S. Intille, A. F. Bobick, "A Framework for Recognizing Multi-Agent Action from Visual Evidence," *M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 489*, pp. 518-525, 1999.

[8] S. Hongeng, R. Nevatia, "Multi-Agent Event Recognition," *International Conference on Computer Vision*, pp. 84-91, 2001.

[9] Y. A. Ivanov, A. F. Bobick, "Recognition of Visual Activities and Interactions by Stochastic Parsing," *IEEE Transactions of Pattern Analysis and Machine Intelligence* , Vol. 22, pp. 852-872, 2000.

[10] Y. Yacoob, M. J. Black, "Parameterized Modeling and Recognition of Activities," *International Conference of Computer Vision*, Vol. 73, pp. 232-247, 1998.

[11] J. W. Davis, A. F. Bobick, "The Representation and Recognition of Human Movement Using Temporal Templates," *Computer Vision and Pattern Recognition*, Vol. pp. 928-934, 1997.

[12] A. D. Wilson, A. F. Bobick, "Realtime Online Adaptive Gesture Recognition," *International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*, pp. 111-117, 1999.

[13] D. J. Moore, I. A. Essa, M. H. Hayes, "Exploiting Human Actions and Object Context for Recognition Tasks," *International Conference of Computer Vision*, Vol. 1, pp. 80-86, 1999.

[14] N. M. Oliver, B. Rosario, A. P. Pentland, "A Bayesian Computer Vision System for Modeling Human Interactions," *IEEE Transactions of Pattern Analysis and Machine Intelligence* , Vol. 22, pp. 831-843, 2000.

[15] N. Oliver, A. Garg, E. Horvitz, "Layered Representation for Learning and Inferring Office Activity from Multiple Sensory Channels," *Fourth IEEE Conference on Multimodal Interfaces*, pp. 3-8, 2002.

[16] C. Rao, A. Yilmaz, M. Shah, "View-Invariant Representation and Recognition of Actions," *International Journal of Computer Vision*, Vol. 50, pp. 203-226, 2002.

[17] N. Badler, "Temporal Scene Analysis: Conceptual Description of Object Movements," *University of Toronto Technical Report No. 80*, 1975.