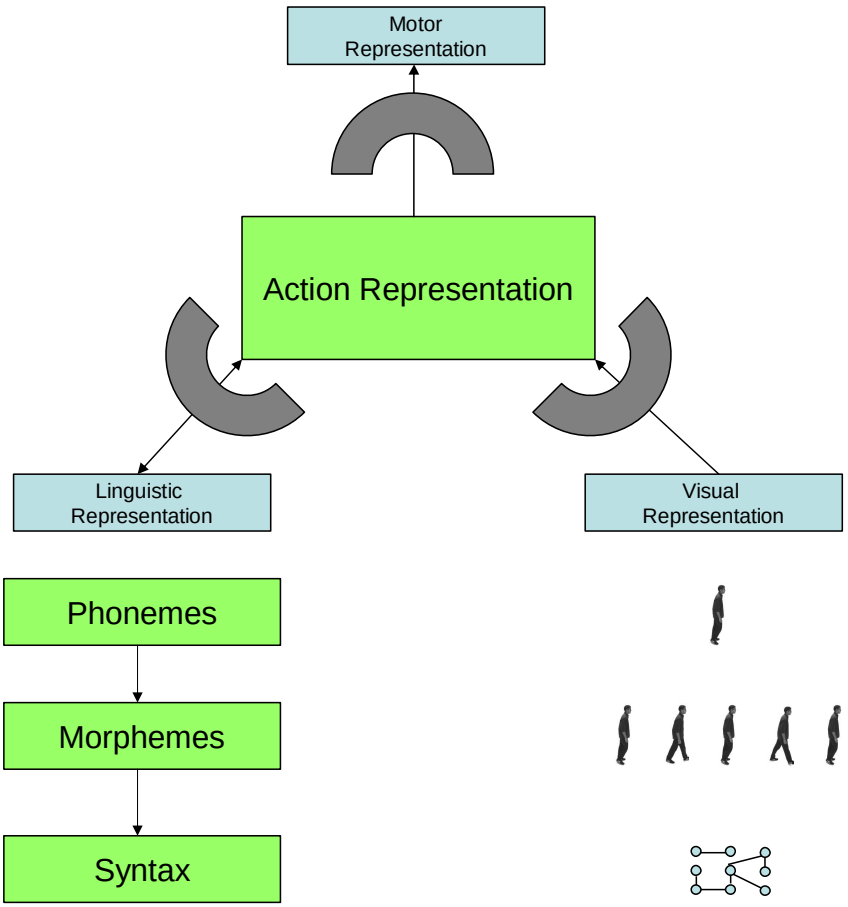
# Modeling Human Activities

**Alap Karapurkar**
Department of Computer Science
University of Maryland, College Park 20742
karapurk@cs.umd.edu




**Abstract**

The ability to recognize the actions of our conspecifics constitutes one of our fundamental cognitive abilities. The ability to understand, and to react to, human actions is thus crucial for cognitive systems of the future. Actions are complex entities, possessing several representations - linguistic, visual, cognitive, and motor. The key to understanding actions is to understand the relationship between these different spaces. In this paper, we study the visual representation of actions and find parallels to natural language processing by representing actions as Probabilistic Context Free Grammars. We also provide a review of previous work done in the field of human action recognition.

# 1 Introduction

**Vision and Action.** The visual space is baffling in its complexity, yet biological systems seem to almost effortlessly sieve through extensive optical sensory input to make sense of the geometry and affordance of their surroundings and of themselves. The perceived effortlessness of human vision conceals the fact that over half of the human brain is engaged in visual processing, and yet the computational principles behind visual perception remain amongst nature's greatest enigmas.

If we wish to build artificial systems that can mimic our visual behavior, these systems need to mimic our representations of the visual world. Evidence from the neuroscience literature[9] increasingly suggests that these representations are dependent in a very fundamental way upon the way we represent our actions. Intuitively, an agent's actions have a direct effect on its perceptual input, and hence, it is plausible that the resulting perceptual input is organized according to the structure present in the agent's actions - structure already known to the agent. Within this view, perceptual processing will be "grounded" in the space of the agent's actions. The possibility of grounding even higher level knowledge by exploiting the relationship between perception and action[5] motivates our study into the nature of human actions and its possible representations.

**The different representations of actions.** Actions have several representations. We can understand a piece of text describing an action, such as "John entered the room, placed his bag on the table, and sat down on the chair". We can even imagine a simulation of the events described in the statement. This is the cognitive/linguistic representation of actions. We also have the cognitive/visual representation of actions - when we see a video of the above action taking place, we can form a mental model of the action and recognize it. Finally, we have the ability to physically imitate the actions of others by transferring their motions/events with respect to other objects to our motor space. Thus we have a cognitive/motor representation. The holy grail of research into actions is to find the relationship between these spaces, and it is our belief that the different spaces are interfaces for a shared cognitive representation. In a completely general setting, we might term actions as a sequence of events undergone by objects in the world. Although a broad definition, it allows us to sketch some of the properties that an action representation system should have. Consider a video of a ball bouncing on a floor. At some level, the representation derived from viewing this video and understanding the sentence "the ball bounced on the floor" must be equal. Linguistics informs us of the semantic structure present in this statement: the subject is the ball, the object is floor and the event or predicate is bounce. Vision provides an opportunity to ground these notions into sensory input. The verb bounce stands for the causal visual event structure in the action, and its representation needs to be independent of the precise form of the subject and object, i.e., the representation needs the ability to abstract away details. The sequence of events themselves specifies a verb, but the precise way in which we move through the sequence, or the dynamics, specifies the action more precisely and determines the adverb. The representation also needs to be compositional - we need the ability to put together several actions to form a complex action. Finally, the existence of mirror neurons[8], hints to a relationship between the ability to visually recognize actions and the ability to imitate the same action.

## 2 Previous Work

The previous section sketched a general notion of actions and the similarity between their different representations. In this paper, however, we are concerned specifically with the visual representation of human actions and to present one such representation by drawing parallels with natural language processing. The problem of human action recognition is complicated by the complexity and variability of shape and movement of the human body, which can be modeled as an articulated rigid body. Moreover, two actions can occur simultaneously, e.g., walk and and wave. Most work on action recognition involving the full human body is concerned with actions completely described by motion of the human body, i.e., without considering interactions with objects. We review a few representative methods from related work in action representation and recognition.

Feng and Perona[4] represent an action using a series of codewords (called movelets). Each codeword is a vector consisting of the parameters of the 10 main body parts (each represented by a rectangle with 5 degrees of freedom) in two successive frames. During the training phase, the set of all observed codewords is clustered using K-means. Each action is thus represented by a series of codewords. Using this data, an HMM is trained for each action. The input video, after background subtraction consists of the silhouette of a single person performing a single action. A probability model for observing two successive silhouettes given the codeword is specified and used for the output probabilities of the HMMs. Given an input video, each HMM calculates its likelihood of generating the video. The input is classified as that action whose HMM yields the maximum of these likelihoods.

Sminchiescu et al.[11] use a conditional random field (CRF) to model human motion. A CRF does not provide a generative model of the motion, but rather learns the conditional probability of motion labels given the data. Their CRF consists of a state node and an output node for each frame in the input sequence. Edges are placed between neighboring state nodes, and between a state node and neighboring output nodes. Each output node is represented by a feature vector (histograms of shape context and edge features) and each state node represents the label for a particular action. The parameters to be learnt consist of parameters of the clique potentials, and is achieved by maximizing the likelihood of the training data. The maximum likelihood estimate of the motion labels is obtained from the input feature vectors using a viterbi style algorithm.

Bregler[2] proposed a hierarchical framework for recognizing actions, invoking the similarity to speech processing. Input pixels in each video frame are grouped into regions (called blobs) using cues of coherent motion, color, spatial proximity, and groupings in previous frames. The states of the blobs in successive frames are grouped together using second order linear dynamical systems (SLDS). The final level consists of HMMs, one for each action, whose nodes output the SLDS currently in action. Given the number of dynamical models and the topology of the HMMs, all of the remaining parameters are learnt automatically. Furthermore, no background subtraction is required. Given an input video, hypothesis at each level are propagated upwards, enabling the calculation of the likelihoods of the HMM models, which are used to infer the actions being performed.

Bobick and Ivanov[1] divide the problem of action recognition into two levels. At the first level, a bank of HMMs is trained, one for each action. These HMMs are then run over the input to be recognized, each HMM generating a series of discrete events indicated by the time intervals at which the HMM (the particular low-level motion) is substantially active. This low level input stream is parsed using a user specified stochastic context free grammars, with the output of HMMs forming the terminal symbols. Usage of the grammar provides the ability to enforce long range causality between lower level motions. This extra structure helps disambiguate low-level motions.The system is tested on hand movements tracked by stereo system.

Schuldt et al.[10] compute local features, i.e., spatio-temporal interest points at multiple scales to recognize human action. The interest points are based on a criterion dependent on the local spatio-temporal image gradient. Each interest point is represented by a vector of higher order spatio-temporal image derivatives around the interest point. During the training phase, all such descriptors are collected and clustered using K-means. Each cluster represents a primitive event. An action is represented using a histogram of primitive events. Since each action is represented by a vector (global features, or causality of primitive events is ignored), the actions can be classified using support vector machines.

Kojima et al.[7] use a hierarchy of case frames to represent the semantics associated with an action and to generate its natural language descriptions. From the input video, the location of orientation of head and hands are extracted by modeling skin color and segmenting out skin colored regions. Objects are identified by comparing their edge images against those present in a database. A case frame is simply the specification of a predicate (e.g. walk), an agent (e.g. human), a goal (e.g. walk by the door), and location (e.g. in front of the table). Predicates such as move, move slow, move fast, high, low, loadable, etc., are defined in terms of geometry extracted from the vision preprocess. A hierarchy of frames derives from observations such as move slow is a specialization of move, stand or sit are specializations of a pose, etc. Additionally, edges corresponding to verb case frames are added between pose case frames (e.g., 'standup' between sit and stand). For each frame in the video, the case frame hierarchy is evaluated starting from the root and moving down the specializations. If the case frame changes between two frames, any case frame corresponding to the transition is activated. These case frames are defined for each body part. Additional rules are built in for merging case frames from separate body parts. The case frames thus obtained are subsequently translated into natural language expressions.

The problem of identifying people and their body parts, while not explicitly addressing the action recognition issue, can serve as preprocesses to some of the above approaches especially when they do not require manual initialization. Felzenszwalb and Huttenlocher[3] use a tree graphical model to represent the body, using one node for each body part. Appearance parameters and constraints between body parts are learnt from training data, and a dynamic programming algorithm is described to infer the state of the graphical model (and hence location of the body parts) from a single image. Since a tree based model does not explicitly take into account phenomenon such as occlusions, Ioffe and Forsyth[6] use a mixture of trees - one for each subset of parts to model different views of the body.

# 3 Action Recognition using Grammars

In this section we describe a framework for recognizing actions in a controlled environment from a database of actions. Our motivation is to formalize the notion of compositionality of actions - i.e., we can form a new action by stringing several known sub-actions together - just like we can form a new sentence by stringing together several known words. Invoking such a parallel allows us to analyze actions in terms of concepts developed in natural language processing (NLP). Higher level actions (which are similar to phrases) are composed of several simpler actions (which are similar to words or morphemes). Each of the simpler actions is in turn composed of several low-level image based representations called keyframes (which are similar to speech phonemes). Borrowing terminology from NLP, the problem of recognizing actions is thus three fold (Fig. 1):

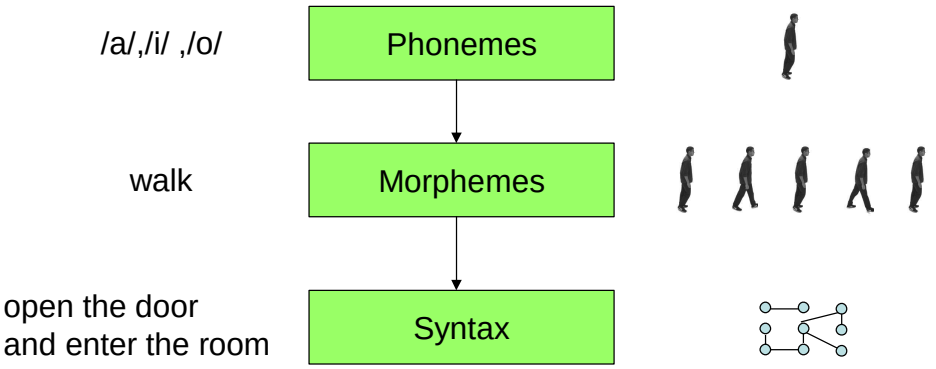



Figure 1: *The Language of Actions: Drawing a parallel between NLP and Actions*

- **Phonology.** Just as phonemes are the low-level structure present in the audio signal, action phonemes are low-level visual descriptors present in the visual signal. We choose a particularly simple form for our action phonemes which is described in the next section. The study of such phonemes is termed phonology.
- **Morphology.** Next in the hierarchy come words or morphemes which are formed by stringing together several phonemes. Similarly, action morphemes (or visual verbs) - the lowest level of actions - are formed by a composition of action phonemes, and the study of such morphemes is called morphology.
- **Syntax.** Meaningful sentences arise only when words or morphemes are combined in certain specific ways. Similarly, meaningful actions occur only when action morphemes are combined in specific ways. The problem of syntax is to find precisely what these specific ways are. The model for action syntax is a direction of future work and not described in this paper.

It is plausible that these stages are not separate but have intricate feedback connections - e.g, the morphemes being recognized might influence the choice of phonemes to search for. However, for our present purposes, we consider them as separate modules. This allows us to analyze and formulate techniques for each of the modules independently and then consider interactions between them at a later stage.

## 3.1 Action Phonemes

The first issue we face is the choice of low-level visual descriptors for actions. In other words, what low level information should we extract from a given video that sufficiently encodes all actions occurring within it? Several models can be used as phonemes - e.g. the optical flow at each frame, or contiguous frames that fit a simple dynamical model, or location and direction of motion of individual body parts from a segmentation process. Using the optical flow at each frame leads to too much information - we want to compress information coming from lower levels in order to efficiently test their combinations as candidate morphemes. The third suggestion, finding location and motion of body parts, provides possibly maximal compression, but at the cost of more processing at the lower level. Segmenting the frames based on a dynamic model lies midway between the two. For our current implementation, we are not interested in the intra-class variability of actions, i.e., we do not want to differentiate between two different types of walks. Staying with the example of a walk, we observe that what is common to different types of walks is the sequence of the extremal positions of the joint angles (legs separated, legs together, legs separated). A simple and useful phoneme representation is thus the human silhouette in those frames where the joint angles hit an extrema (We assume the existence of a segmentation procedure which finds the silhouette). However, the problem of finding the joint angles in a given video is known to be a hard problem. We can avoid this problem by observing that the joint extrema are strongly correlated with the extrema of the average optical flow (Fig. 2). Intuitively, at their extremal positions, the joint angles are either stationary or moving very fast, hence the correlation with optical flow. Given an input video, our phoneme representation is thus the silhouette

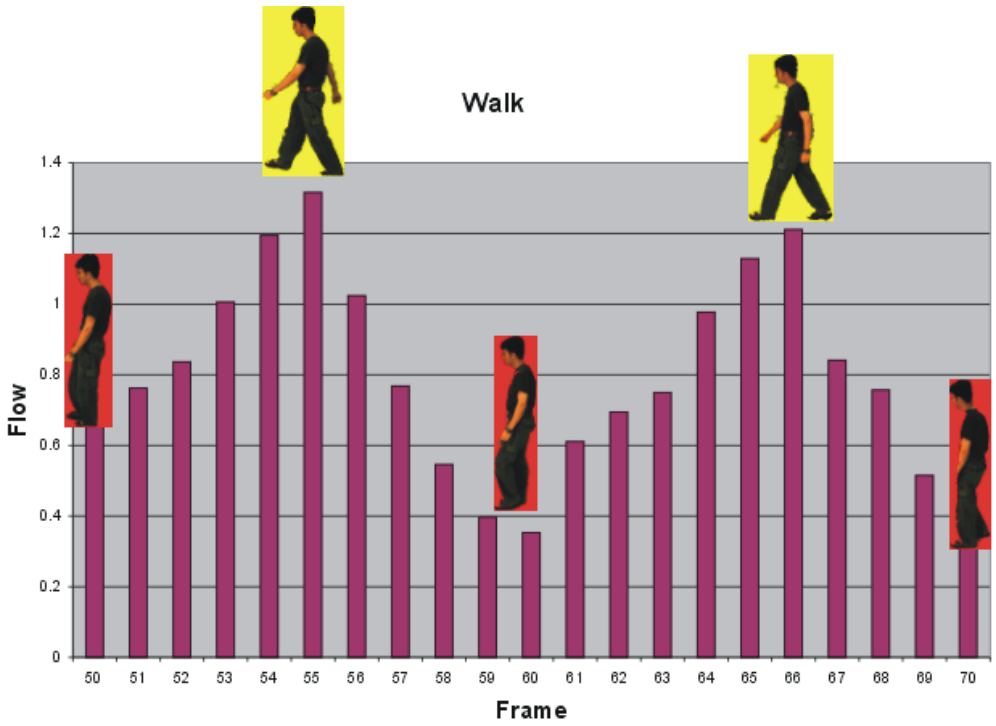



Figure 2: *The extrema of optical flow strongly correlate with the extrema of joint angles.*

of the human body in those frames where the average optical flow hits an extrema (maximum, minimum, or inflection). Undoubtedly, this is a very shallow phoneme representation as we neglect a lot of spatio-temporal information, and furthermore, as mentioned before, it is possible that the choice of phonemes is affected by through feedback by the higher level modules performing action recognition. We discuss such issues in the final section. We have found that this simple representation suffices for the kind of actions under considerations and allows us to build higher level modules on top of it. In addition to the phoneme representation, we also need a distance function to evaluate the similarity of two phonemes. In our current implementation, we employ a similarity metric based on phase correlation of the silhouette images (to maximally register the two images) and similarity of optical flow within the two registered images.

## 3.2 Action Morphemes

An action morpheme is represented as a sequence of action phonemes. Action morphemes thus form the first level of actions - those which cannot be divided into smaller actions. For example, the action morpheme for walk is defined by the five phonemes shown in Fig. 2. We currently do not use any information of the dynamics between two phonemes. Such information could be stored as attributes in the grammar symbols and used to specify intra-class variability, e.g. slow or fast walk. Now, given an input video, if we run our phoneme detector, we obtain a stream of phonemes. These phonemes need to be grouped into their respective morphemes. This simultaneous task of segmentation and recognition is performed using a Probabilistic Context Free Grammar (PCFG). We formulate a PCFG which extracts the most likely sequence of morphemes which generated the observed sequence of phonemes. The detailed PCFG is described subsequently, but at a higher level, the grammar is structured as follows:

- Video $\rightarrow$ Action | Action Action | ...
- Action $\rightarrow$ $\text{Action}_1$ | $\text{Action}_2$ | ...
- $\text{Action}_i \rightarrow \text{ModelPhoneme}_1$ $\text{ModelPhoneme}_2$ ...
- $\text{ModelPhoneme}_i \rightarrow \text{ObservedPhoneme}_j$

In words, each video generates a sequence of actions. Each action generates a sequence of model phonemes. And finally, each model phoneme generates an observed phoneme with a certain probability which depends on the distance function between two phonemes. The grammar thus specifies a generative model for the set of actions it models. The probabilities associated with each production encodes a probability distribution over the set of all possible strings generated by this grammar. Since the grammar is unambiguous, there is a one-to-one correspondence between strings in the generated language and their respective parse trees. Thus, the grammar encodes a probability distribution of parse trees given a particular input (stream of observed phonemes). The most likely parse tree will segment the set of observed phonemes into the set of morphemes that most likely generated them.

### 3.3 Implementation and Results

**Data collection.** We collected training data consisting of 11 actions (jump, walk, kick, kneel, squat, pickup, sitstand, punch, wave, handshake, turn) performed by 10 actors, each action being observed through 8 synchronized cameras (Fig. 3). All videos were captured against a white background to enable trivial background subtraction.

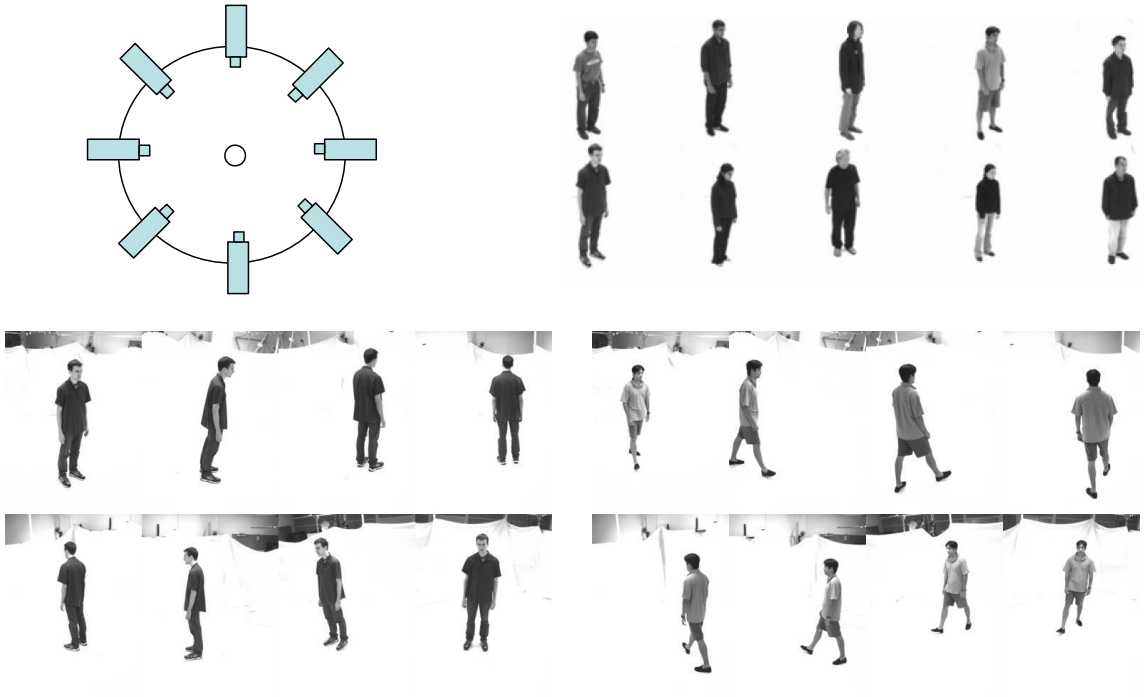


Figure 3: *Experimental setup: 11 actions were performed by 10 actors (top right) and captured by eight synchronized cameras a in surround configuration (top left). The two bottom images show eight views of a single frame for two actions.*

**Phoneme Extraction.** Our first task is to represent each action in the database by a short sequence of keyframes (action phonemes). For each action in the database, we run our phoneme detector on the video stream from each view. This yields a set of keyframes for each view. The union of the set of keyframes from all views represents the set of keyframes for the action. Thus, each action $a_i$ consists of a sequence of phonemes $(p_1, p_2, \ldots, p_n)$, where each phoneme consists of eight views of the corresponding body pose $p_i = (p_i^1, p_i^2, \ldots, p_i^8)$. At the end of this process, each action is represented by short sequence of phonemes. However, several actions may share the same phoneme, and the same action can have multiple instances of the same phoneme. Therefore, we have to reduce the set of phonemes obtained from all actions independently into a joint phoneme set. Two phonemes are considered equal if the distance function (described in the previous section) between each of the corresponding views lies below a threshold. Using this criterion, each set of equivalent phonemes is replaced by a single representative phoneme and the set of phonemes is compressed into a smaller, unique set. The phoneme set found from the training data is shown in Fig. 4

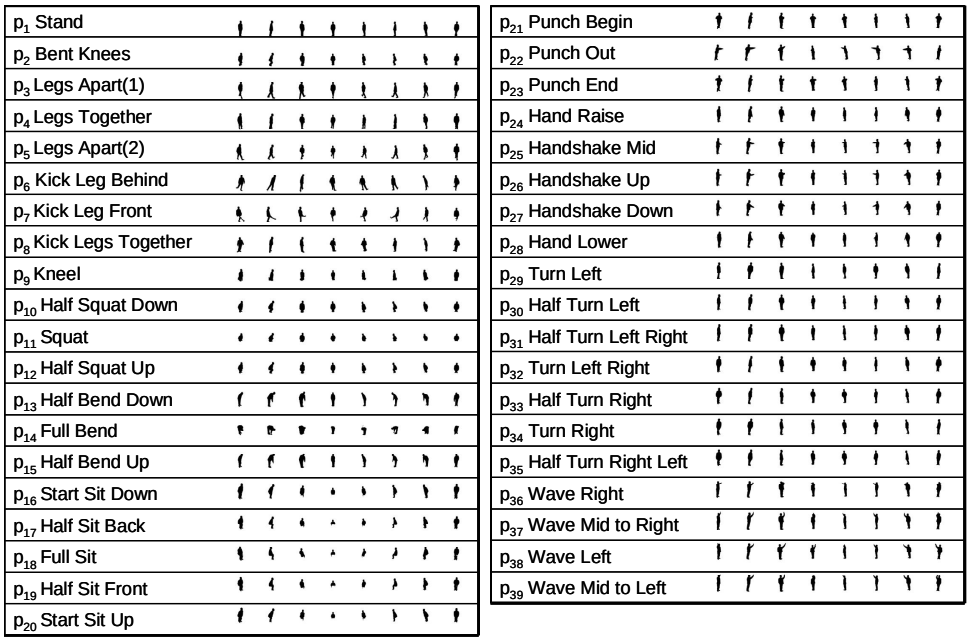



Figure 4: *Set of unique phonemes obtained from the training data*

**Grammar Construction.** We use these phonemes to specify an action grammar as show below. The first rule specifies that each video $V$ is composed of a sequence of upto $f$ actions. All of the actions $A_1, A_2, \ldots, A_g$ are equiprobable. The third rule states that each action is composed of a sequence of an ordered pair of phonemes. Each phoneme in the phoneme stream for an action is duplicated, except the first and the last, and successive phonemes in this new stream are grouped together to form the ordered pair of phonemes. This is necessary to allow sharing of border phonemes between actions. The fourth rule specifies that each ordered pair generates a particular view from each of the constituent phonemes. Only those productions are considered where the view of the two phonemes either remains the same or changes to an adjacent camera, with higher probability assigned to staying in the same view. The final production states that each view in a model phoneme generates an observed phoneme with a probability derived from the distance function between the two. Since the observed phonemes change for each input video, these productions are formed at runtime.

| | |
|---|---|
| $V \rightarrow A \mid AA \mid \ldots \mid A^f$ | $\forall i, p(A^i \mid V) = 1/f$ |
| $A \rightarrow A_1 \mid A_2 \mid \ldots \mid A_g$ | $\forall i, p(A_i \mid A) = 1/g$ |
| $A_i \rightarrow q_{ab} q_{bc} q_{cd} \ldots$ | $p(q_{ab} q_{bc} q_{cd} \ldots \mid A_i) = 1$ |
| $q_{cd} \rightarrow p_c^u p_d^v$ | $\sum_{\substack{allowed \\ u,v}} p(p_c^u p_d^v \mid q_{cd}) = 1$ |
| $p_i^v \rightarrow s_k$ | $p(s_k \mid p_i^v)$ *obtained at runtime* |

**Sample Parses.** We used the Viterbi PCFG parser provided in the Natural Language Toolkit (NLTK, *http://nltk.sourceforge.net*) for implementing the grammar. The test data consisted of single camera video of a sequence of actions. Two sample parses are shown in Fig. 5. The first example consists of a video of a person performing a succession of four actions - walk, turn, kick, and kneel, seen from a single camera. The grammar correctly separates the extracted phonemes into the corresponding morphemes. The second video shows a person walking along a circle followed by a pickup action. This shows the ability of the system to exploit the viewer/camera duality (a moving viewer with stationary camera is equivalent to a stationary camera with moving viewer) to find the correct orientation of the body. The viewpoint changes in accordance with direction of movement and the sequence is correctly parsed into its constituent actions.

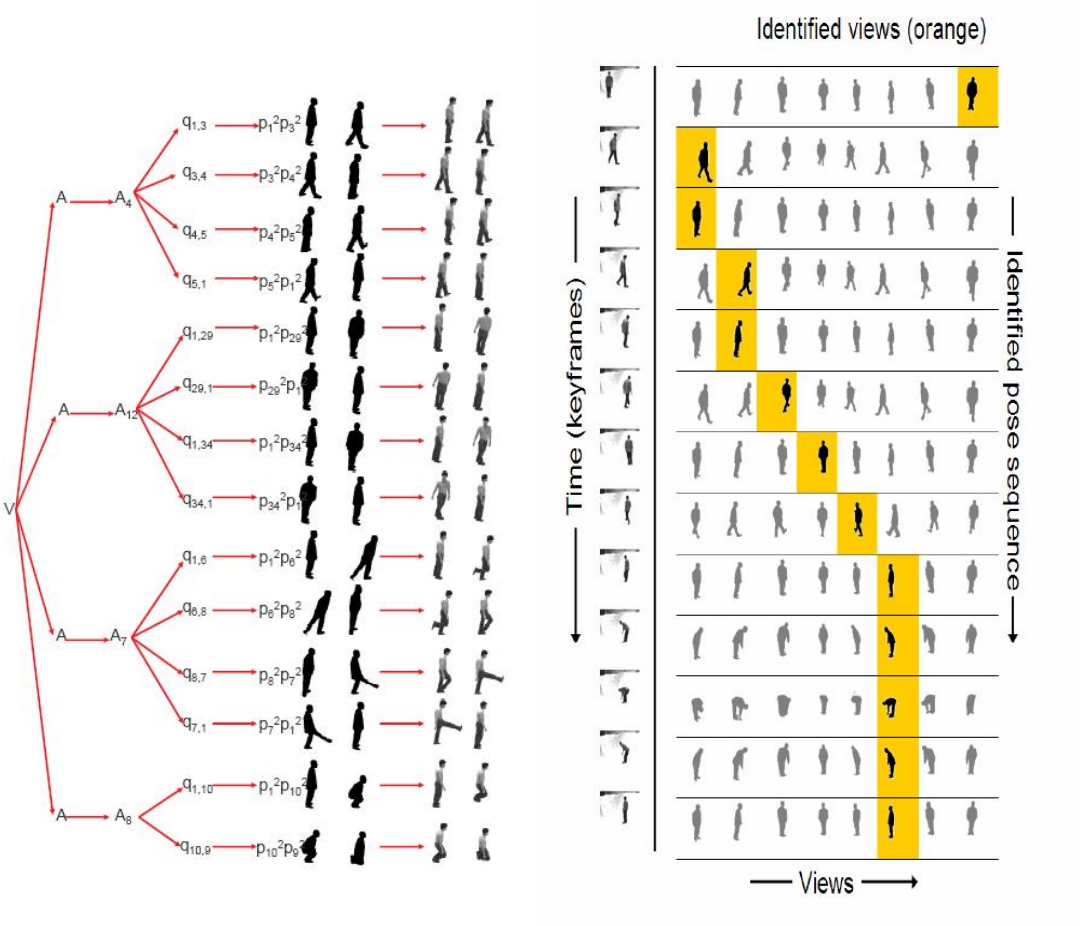



Figure 5: *Sample parses obtained by the system. Left: A sequence comprising of walk, turn, kick, and kneel. The terminal symbols are the observed phonemes, the black silhouettes are the identified model phonemes. As can be seen, the video is correctly segmented into its constituent actions. Right: Identified pose sequence for a walk and turn sequence. The left column displays the observed phonemes. Corresponding to each observed phoneme is the inferred model phoneme, with the particular inferred view highlighted.*

# 4 Conclusions and Future Work

In this paper, we have presented a framework for representing human actions by dividing the problem into three stages. We have proposed simple techniques for realizing the first two stages - representing action phonemes by silhouettes with extremal average optical flow, and morphemes as a sequence of phonemes with constraints placed using a probabilistic context free grammar. Furthermore, the phonemes and morphemes are generalized to handle multiple views. Given the goals outlined in Section 1, the approach described is very basic and several improvements can be made within the framework. The action phonemes and morphemes do not take into account the dynamics between keyframes. The phonemes are composed of full-body silhouettes which is too restrictive. Moreover, we only consider actions involving a single person with no object interaction, the level of syntax is unspecified. Future work will lie along incorporating these issues into the framework, which will possibly involve relaxing the rigid grammar structure into a more loosely organized hierarchical/compositional structure with the capability of representing both schema-like structures and low-level geometric entities.

# References


[1] A. F. Bobick and Y. A. Ivanov. Action recognition using probabilistic parsing. In *CVPR '98: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Washington, DC, USA, 1998.

[2] C. Bregler. Learning and recognizing human dynamics in video sequences. *Proceedings of the IEEE Conf. Computer Vision and Pattern Recognition*, pages 568–574, 1997.

[3] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79, January 2005.

[4] X. Feng and P. Perona. Human action recognition by sequence of movelet codewords. *Proceedings of First International Symposium on 3D Data Processing Visualization and Transmission*, pages 717–721, 2002.

[5] Vittorio Gallese and George Lakoff. The brains's concepts: the role of the sensory-motor system in conceptual knowledge. *Cognitive Neuropsychology*, 21(0), 2005.

[6] Sergey Ioffe and David A. Forsyth. Human tracking with mixtures of trees. In *ICCV '01: Proceedings of the International Conference on Computer Vision*, pages 690–695, 2001.

[7] A. Kojima, T. Tamura, and K. Fukunaga. Natural language description of human activities from video images based on concept hierarchy of actions. *International Journal of Computer VisionV*, 50(2):171–184, November 2002.

[8] G. Rizzolatti and M. A. Arbib. Language within our grasp. *Trends in Neurosciences*, 21(5):188–194, 1998.

[9] Giacomo Rizzolatti and Vittorio Gallese. *Problems in Systems Neuroscience*, chapter Do perception and action result from different brain circuits? The three visual systems hypothesis. 2005.

[10] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: A local svm approach. In *ICPR '04: Proceedings of the 17th International Conference on Pattern Recognition*, volume 3, pages 32– 36, August 2004.

[11] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Conditional models for human motion recognition. In *ICCV '05: Proceedings of the International Conference on Computer Vision*, Beijing, China, 2005.