**Human Action Recognition and Prediction: A Survey**

**INTRODUCTION**

Video analysis research work has progressed from understanding the present state to predicting the future events. Human actions have an inherent goal driving it. AI research work targets to predict human actions so that it can provide effective services. For e.g., a medical assistant robot should be capable of perceiving patient’s actions, analyze efficacy of exercises prescribed to him and thereby prevent further injuries.

The major applied difference between action recognition and action prediction lies when a decision is made. Human action recognition is to infer the action label after the entire action execution has been observed, whereas prediction requires arriving at future state conclusion before it happens, which mandates faster inference engines.

**Real world applications**

Human action recognition and prediction methods powers a plethora of real world applications such as video surveillance, video content auto tagging for retrieval, entertainment & gaming industry, human robot interaction and last but not the least autonomous driving technology.

State-of-the-art algorithms such as temporal segment network, spatiotemporal multiplier convolution network and activity prediction using LSTM, remarkably reduces human labor in analyzing a large-scale video data and provide understanding on the current state and future state of an ongoing video data.

RGB-D sensors like Kinect provide depth info in addition to RGB info. This extra depth dimension encodes rich structural data and facilitates action recognition as it simplifies intra class motion variations and reduces cluttered background noise.

Action prediction algorithms are used to predict a person’s intention and trajectory of motion in a short period of time by studying action evolution so as to avoid collision in autonomous driving vehicles.

**Research Challenges**

Action recognition and prediction research work involves numerous hurdles such as distinguishing between Intra- and Inter-class variations, Cluttered background and camera motion, insufficient annotated data and uneven predictability.

Deep learning has shown promising results on data collected from uncontrolled settings, but it requires large amount of annotated training dataset.

**HUMAN PERCEPTION OF ACTIONS**

Different perspectives describing an action needs to answer what is the action, why the action was performed and who executed the action. Computation models for former two concerns have been thoroughly studied and implemented with considerable success, but the studies on agent’s identity and social context which provides a better description of the agent, i.e. narrative understanding has fewer work in computer vision community. As humans we draw inference about the goal of an action by observing the end state after an action is performed. It is believed that the inference is arrived at drawing similarities between observed action and observer’s own motor representation for that action with the help of mirror neuron system.

Action prediction cues are also drawn from emotional and attentional information such as facial expression or gaze of the individual.

**Action Recognition**

**Shallow Approaches**

* **Action Representation**

Action recognition problem statement resolution requires that we figure out how to represent an action in a video. Actions appearing in a video differs in speed, camera point of view, appearance and pose variation for same set of objects.

One of the major challenges in action recognition is large appearance and pose variations in one action category, making the recognition task difficult. The goal of action representation is to convert an action video into a feature vector and extract representative and discriminative information of human actions, and minimize the variations, thereby improving the recognition performance.

* **Holistic Representations:**

Holistic representation technique captures action as a 3D volume in space time by encoding spatial position of body over time. It employs 3 popular techniques to create a holistic representation, namely Motion Energy Image (MEI), Motion History Image (MHI) and 3D motion history volume (MHV) technique. The MEI procedure represents the spatial distribution of motion where bright region signifies both the action occurring and the viewing condition.

Whereas, Motion History Image (MHI) procedure illustrates both position and trajectory of the motion. MHI leverages pixel intensity to denote motion history at a particular location, the brighter values correspond to a recent movement. However, MHI is not immune to viewpoint changes, therefore 3D motion history volume (MHV) technique is applied where 3D occupancy of body is captured from various viewpoints using 3D voxels. And, then Fourier transform post processing provides a feature vector invariant to locations and rotations.

The motion in horizontal and vertical axis is computed by the so-called optical flow algorithms which indicates pattern of apparent motion of objects on two consecutive frames.

But holistic representations are inherently marred with sensitivity to noise in motion data. This is resolved by ***Local representation*** techniques described below.

* **Local Representations**

Local representations capture salient motion information in localized areas to overcome the problem in holistic representations. Space-time interest points (STIPs) based approaches successfully captures motion trajectory, that are immune to translator and appearance variation.

A spatio-temporal separable Gaussian kernel is applied on a video to obtain its response function for finding large motion changes in both spatial and temporal dimensions. Bregonzio et al. detected spatial-temporal interest points using Gabor filters and spatiotemporal Hessian matrix. And, gradients over optical flow fields are computed to build the so-called motion boundary histograms (MBH) for describing trajectories.

Feature trajectory is a straightforward way of capturing long-duration information where spatio temporal representation cannot be used.

Then a hierarchical context information is captured through this method to generate more accurate and robust trajectory representation.

**Action Classifiers**

Action classification is the next step in action recognition process once a representation has been computed.

* **Direct Classification:**

This methodology leverages off the shelf classifiers such as SVM, kNN or bag of words model. The bag of words model encodes distribution of local motion patterns using a histogram of visual words. It uses STIP to detect local salient regions and then assigned visual words to perform bag of model computation.

* **Sequential Approaches:**

This approach captures temporal evolution of appearance or pose using sequential state models such as hidden Markov models (HMMs), conditional random fields (CRFs) and structured support vector machine (SSVM). This line of work involves treating videos as temporally aligned frames where HMMs are used to plot the trajectory.

* **Space-time Approaches:**

This approach tries to accommodate spatiotemporal correlations between local features.

It learns a global Gaussian mixture model (GMM) using the relative coordinates features and uses multiple GMMs to describe the distribution of interest points over local regions at multiple scales.

The feature is calculated by using the transform which is defined as an extended 3D discrete Radon transform. Such feature captures the geometrical information of the interest points and keeps invariant to geometry transformation and robust to noise.

Graph is used to capture the spatial and temporal relationships among local features where local features are used as the vertices of the two-graph model and the relationships among local features in the intra-frames and inter-frames are characterized by the edges.

However, these methods are limited to small datasets as they need to model the correlations between interest points which are explosive on large datasets.

* **Part-based Approaches:**

It is straightforward to model human actions using motion information from body parts as it approaches consider motion information from both the entire human body as well as body parts.

A constellation model was proposed which models the position, appearance and velocity of body parts. Inspired by a part-based hierarchical model in which a part is generated by the model hypothesis and local visual words are generated from a body part.

* **Manifold Learning Approaches:**

Human actions can be described by temporally varying silhouettes. However, their representation is high dimensional, thus in order to achieve efficient action recognition manifold learning approaches are applied for dimensionality reduction. A novel manifold embedding method finds the optimal embedding that maximizes the principal angles between temporal subspaces associated with silhouettes of different classes.

* **Mid-Level Feature Approaches:**

Bag of words model may not well represent actions due to the large semantic gap between low-level features and high-level actions. Thus, Hierarchical approaches are proposed to learn an additional layer of representations and expect to better abstract the low-level features for classification by learning mid-level features from low-level features to recognize action tasks.

* **Feature Fusion Approaches:**

A multi-task sparse learning (MTSL) model is used to fuse multiple features for action recognition. In addition to this, a multi-feature max-margin hierarchical Bayesian model (M3HBM) is used to learn a high-level representation by combining a hierarchical generative model (HGM) and discriminative max margin classifiers in a unified Bayesian framework.

A video set is modeled as an optimized probabilistic hypergraph, and a robust context-aware kernel is used to measure high order relationships among videos.

* **Classifiers for Human Interactions:**

Instead of directly modeling action co-occurrence, the approach proposes to learn phrases that describe the motion relationships between body parts. This will describe complex interactions in more details and introduce human knowledge into the model.

* **Classifiers for RGB-D Videos**

Classifiers for RGB-D videos get skeleton data provided by a Kinect sensor. The method projects various types of features including skeleton features and local HOG features into a shared feature space learned by minimizing the reconstruction loss. A different work contains a method that jointly learns RGB-D and skeleton features and action classifiers. The projection matrices are computed by minimizing the noise after projection and classification error using the projected features.

**Deep Architectures**

Recent works provide feature learning using deep learning techniques has received increasing attention due to their ability of designing powerful features that can be generalized very well. Action features learned by deep learning techniques has been popularly investigated with convolution operation being one of the fundamental components in deep networks for action recognition.

3D convolutional networks (3D Conv Nets) directly create hierarchical representations of spatio-temporal data. However, the issue is they have many more parameters than 2D Conv Nets, making them hard to train. In addition, they are prevented from enjoying the benefits of ImageNet pre-training.

Temporal modelling is one of the key variables in designing deep networks. One straightforward way is to directly apply 3D convolution to several consecutive frames.

* **Space-time Networks**

Space-time networks extend 2D Conv Nets by capturing temporal information using 3D convolutions.

The C3D network contains 5 convolution layers, 5 max pooling layers, 2 fully connected layers, and a softmax loss layer, subject to the machine memory limit and computation affordability. Their work demonstrated that C3D learns a better feature embedding for videos. Results showed that C3D method with a linear classifier can outperform or approach the state-of-the-art methods on a variety of video analysis benchmarks including action recognition and object recognition.

The 3D CNN network architecture has 5 hardwired kernels including gray, gradient-x, gradient-y, optflow-x and optflow-y resulting in 33 feature maps.

One limitation of 3D Conv Nets is that they typically include very short temporal intervals, such as 16 frames, thereby failing to capture long-term temporal information. To address this problem, the work on “Long-term temporal convolutions for action recognition,” increases the temporal extent in the 3D convolutions, and empirically illustrates significant improvement in the recognition performance.

* **Multi-Stream Networks**

Multi-stream networks utilize multiple convolutional networks to model both appearance and motion information in action videos. To overcome performance issues compared to shallow implementations, recent work’s architecture contains two separate streams, a spatial Conv Net and a temporal Conv Net. The former one learns actions from still images, and the later one performs recognition based on optical flow field.

An improvement was proposed, which used the two-stream network to obtain multi-scale convolutional feature maps and pooled the feature maps together with the detected trajectories to compute Conv Net responses centered at the trajectories.

In order to improve interactions between two streams, Feichtenhofer et al. proposed a series of spatial fusion functions that make channel responses at the same pixel position be in the same correspondence.

Such a strategy bridges the gap between the two streams and allows information transfer in learning spatiotemporal features.

* **Hybrid Networks**

Hybrid networks employ addition of RNNs to aggregate temporal information for e.g. two-stream CNN is used to extract motion features from video frames, and then fed into a bi-directional LSTM to model long term temporal dependencies.

**ACTION PREDICTION AND MOTION PREDICTION**

Action prediction tasks can be roughly categorized into two types, short-term prediction and long-term prediction.

* **Early Action Classification**

Action prediction approaches aim at recognizing unfinished action videos. Integral and dynamic bag of word approaches are used for action prediction. Different from the aforementioned methods, a research work studied the action prediction problem in a first-person scenario, which allows a robot to predict a person’s action during human-computer interactions.

The work in one such paper proposed a new monotonically decreasing loss function in learning LSTMs for action prediction. Inspired by that, we adopted an autoencoder to model sequential context information for action prediction. Our method learns such information from fully observed videos and transfer it to partially observed videos.

* **Intention Prediction**

Pier et al. addressed the problem of goal inference and intent prediction using an and-or-graph method, based on stochastic context sensitive grammar. It models all possible parse graph of a single event of model object interactions.

Probabilistic Suffix Tree (PST), which captures variable Markov dependencies between action primitives in a complex action. They proposed an anticipatory temporal conditional random field (ATCRF) to model three types of context information, including hierarchical structure of action primitives.

**Motion Trajectory Prediction**

Motion trajectory prediction is accomplished by the use of state-of-the-art semantic scene understanding combined with ideas from inverse optimal control (IOC) or inverse reinforcement learning. In this work, human motion is modeled as a sequence of decision-making process, and a prediction is made by maximizing the reward.

It captured the spatial positions of the neighboring trajectories of a person by a so-called social affinity map.

Recent advancements in deep networks has ushered in motion trajectory prediction problem solution using RNN/LSTM networks.

A single LSTM model was used to account for one single person’s trajectory, and a social pooling layer in LSTMs was proposed to model dependencies between LSTMs and preserve the spatial information.

An encoder-decoder framework was proposed for path prediction in more natural scenarios where agents interact with each other and dynamically adapt their future behaviors. Past trajectories are encoded in a RNN and then future trajectory hypotheses are generated using another decoder implemented by a separate RNN.

The proposed Deep IOC is used to rank all the possible hypotheses. Using an adversarial loss, the approach can potentially learn the distribution of multiple socially acceptable trajectories, rather than learning the average trajectories in the training data.

**FUTURE DIRECTIONS**

* **Benefitting from image models**

This line of work concentrates on leveraging deep network works on image data for video data so as to transfer knowledge from image models to video models.

* **Interpretability on temporal extent**

Temporality in video data frames decides on importance of each frame towards prediction of future events/action. This involves breaking the task into simple primitives over period of time and how they contribute to classification problem. And the machine would learn salient signals on why some actions can be predicted at earlier stage compared to others.

* **Learning from multi-modal data**

Humans consume multi modal data through five senses. This multi modal data can be used to build better understanding of visual field thereby providing a richer semantic meaning to the scene. It could be understood as assigning a noun, verb, preposition & adjective to an action so as to provide more descriptive info such as human action strength or flow thus enabling fine grained action understanding.

* **Learning long-term temporal correlations**

Multi modal data can also be used to understand long temporal relationship between visual entities present in data. Long term sequential correlation enumerates the ordering in which actions occur, strikingly similar to how our brain stores action data.

* **Physical aspect of actions**

Motion prediction usually concentrates on higher level action outcomes whereas finding relationship between primitive constituent actions to physical aspects would help provide fine grained action recognition by providing spatial and human-object as well as object – object interactions.

* **Learning actions without labels**

With the prevailing scarcity of availability of high-quality labeled video data for action recognition and prediction , development of a robust and efficient action recognition and prediction approaches that can learn automatically from unlabeled data would be an important milestone.

**CONCLUSIONS**

The availability of big data and powerful models diverts the research focus about human actions from understanding the present to reasoning the future. We have presented a complete survey of state-of-the-art techniques for action recognition and prediction from videos. These techniques became particularly interesting in recent decades due to their promising and practical applications in several emerging fields focusing on human movements. We investigate several aspects of the existing attempts including handcrafted feature design, models and algorithms, deep architectures, datasets, and system performance evaluation protocols. Future research directions are also discussed in this survey.

References:

* [The Poisson Equation in Image & Shape Processing](https://www.cs.jhu.edu/~misha/Fall07/)
* [Gabor filter](https://medium.com/@anuj_shah/through-the-eyes-of-gabor-filter-17d1fdb3ac97)
* [Spatiotemporal Hessian Matrix](https://en.wikipedia.org/wiki/Corner_detection#Spatio-temporal_interest_point_detectors)