

# Modeling and Detecting Human Action using Animated Pose Templates

Keerthika.S<sup>1</sup> Kirubakaran.B<sup>2</sup>

<sup>1,2</sup>Assistant Professor

<sup>1,2</sup>Department of Electronics and Telecommunication Engineering

<sup>1,2</sup>N.S.N College Of Engineering and Technology Karur

**Abstract**— This research paper pose templates are used to identify short term, long term, and contextual action from a video scene. Each pose template consists of two components: 1) shape template represented by Histogram of Oriental Gradient (HoG). 2) The motion template represented by Histogram of Optical-Flows (HoF). Totally five videos are taken in training phase. The images are converted into frames. Neural Network is used to identify these actions from the videos. The images are divided into back view and front view. The back view image is subtracted from the original image. Then from the front view image we extract the actions of the person. The HoG is based on identifying the edges of the image. The HoF is based on threshold value. The threshold value is the grade value. A shape template may have more than one motion template represented by or -node. Therefore, each action is defined as a mixture (Or-node) of pose templates in an and-Or tree structure. While this pose template is apposite for detecting short-term action snippets in two to five frames, we spread it in two ways: 1) For long-term actions, we activate the pose templates by adding temporal constraints in a Hidden Markov Model (HMM), and 2) for contextual actions which are detected by the complete set by using SIMULATION and area is obtained by using the MATLAB 2013 software.

**Key words:** HOG-Histogram of gradient, HOF-Histogram of optical flow, HMM-Hidden markov Model

## I. INTRODUCTION

Human action recognition has increasing research interest in recent years motivated by a range of applications from video surveillance, human computer intercommunication, content-based video retrieval. But real world human understanding presents lot of challenges like recognizing the action, localizing the action, interpreting the action. Action understanding from real-world videos, which commonly contain heavy clutter and background motions, remains a hard problem.

Actions are the building blocks for activities and events. They are based upon the agents involved and the time to do the activity. But it have complexity in time and space.

- 1) In space, actions are defined by body parts, body poses, or by human-scene interaction.
- 2) In time, actions are defined by single frame, or by two to five frames, or by apart from future.

To overcome the short comings of the existing method and to cover the time complexity neural network algorithm has been used for recognizing short term, long term, and contextual actions. Short term action refers to action performed by two to five frames. Complex actions composed of more frames. By clustering method these complex actions will be clustered and composed.

A shape template having a root window (bounding box) covering a number of deformable parts whose appearance is modeled by the HOG features. A motion template specifying

the motion of the parts by the Histogram of Optical-Flows (HOF) features. Long term or continuous actions are represented by sequence of moving pose templates and animated pose templates.

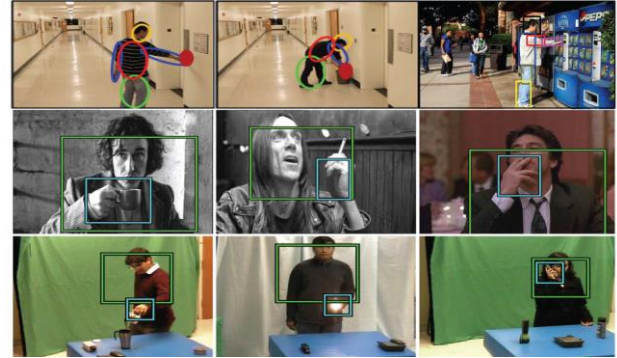


Fig. 1: Actions as interactions between an agent and contextual objects. Three examples from three action data sets that we use in the experiments.

## II. MOTIVATION

The motivation of this research paper is to identify persons from the video images by using pose templates. Several methods were used but they does not provide optimized result. The existing system is a complex and time consuming system. The person is identified by using background subtraction method. Human action recognition has attracted increasing research interest in recent years motivated by a range of applications from video surveillance, human-computer interaction, to content-based video retrieval. Building a robust system for real-world human action understanding presents challenges at multiple levels: 1) localizing the actions of interest; 2) recognizing the actions; and 3) inter- prettying the interactions between agents and contextual objects. Recent research has made major progress on classifying actions under idealized conditions in several public data sets, such as the KTH data set and Weizmann data set. Where each video clip contains one person acting in front of static or uncluttered background with one action per video clip. State-of-the-art methods have achieved nearly 100 percent accuracy on these two data sets. On the other hand, action understanding from real-world videos, which commonly contain heavy clutter and background motions, remains a hard problem. In general, actions are building blocks for activities or events, and thus are simpler than the latter in terms of the number of agents involved and the time duration, but still have diverse complexities in space and time:

- 1) In space, actions can be defined by body parts, such as waving and clapping, by human poses, such as walking, or by the human-scene interactions, such as making coffee in an office and washing dishes in a kitchen. In the last case, the whole image provides contextual information for action recognition.

- 2) In time, actions can be defined in a single frame, such as sitting and meeting, two to five frames such as pushing a button and waving hand which are also called action snippets, or a longer duration say in 5 seconds, such as answering a phone cell

### III. OVERVIEW OF THIS METHOD

Motivated by the various shortcomings in existing methods and the need to cover actions of different space-time complexity, we present an APT model for recognizing short term, long term, and contextual actions from real-world videos. In the following, we overview our model and algorithm in comparison to the existing methods: Short-term actions as moving pose templates (MPTs). Short term actions or the so-called action snippets refer to actions observed in three to five frames (0.2-0.3 seconds of video), and they contain rich geometry, appearance, and motion information about the action. Fig. 1 shows two examples for clapping and drinking. A more complex action is often composed of a sequence of action snippets. Three instances and each instance have three snippets. By clustering these snippets, which implicitly aligns them in time, we learn a dictionary of moving pose templates: one for each snippet A moving pose template consists of two components. A shape template having a root window (bounding box) covering a number of deformable parts whose appearance is modeled by the HOG features. Like the DPM model for human detection the geometric relations between the parts are included in a Gaussian distribution.

A motion template specifying the motion of the parts by the Histogram of Optical-Flows (HOF) features. We compute motion velocities of ingredient by the Lucas-Kanade algorithm to avoid the complexity of establishing temporal correspondence of parts between frames, since tracking form parts in cluttered video is a notoriously hard problem. The same shape template may have different motion templates, for example, in the clapping action, the raised arms could be moving upwards or downwards. In comparison with the popular STIP representations, the moving pose templates represent the human geometry, appearance, and motion jointly and clear. Thus, it is a Stronger model.

Long-term actions as animated pose templates. Long-term and continuous action, such as walking and running can be represented by a sequence of moving pose templates and we call it the animated pose template. The term “animation” has been used in motion picture as a technique of rapidly displaying a sequence of images to create an illusion of continuous movement. The famous example is the galloping horse created by Muybridge and is shown in fig 3 in this example, it is relatively easy to track the rider and the body of the horse over frames; however, it is hard or impossible to track (establishing correspondence between) the fast moving legs. The displacement of the horse legs is so large between two consecutive frames that conventional optical flow algorithms will fail.



Fig. 2: Action snippets contain rich appearance, geometry, and motion information in three frames: 1) the static pose, 2) the short-term motion velocities (illustrated by blue arrows).

#### A. Animated Pose Template:

The formulation of the animated pose template model in three incremental; moving pose templates for short-term action snippets; APTs to account for long-term transitions between the pose templates and APT augmented with contextual object.

#### B. Moving Pose Templates:

Each action is composed of a sequence of key poses or action snippets, and the number of poses depends on the complexity of the action. Fig. 2 displays three poses for a hand-clapping action from the MSR data set.

The term “animation” has been used in motion picture as a technique of rapidly displaying a sequence of images to create an illusion of continuous movement. It is relatively easy to track the rider and the body of the horse over frames; however, it is hard or impossible to track (establishing correspondence between) the fast moving legs. The displacement of the horse legs is so large between two consecutive frames that conventional optical flow algorithms will fail. Studied the issue of intractability as a measure for the entropy of the posterior probability of velocity. They display that motion in a video can be partitioned as trackable motion, for which motion correspondence can be computed reliably, and intractable motion, for which we need to compute a reduced or projected representation, such as a histogram of velocity in an area. The intractability is affected by factors, such as object density, image scaling (sampling rate in space and time), and stochasticity of the video.

Our animated pose template is a generative model based on the moving pose templates (or snippets). The shape templates between consecutive action snippets are considered trackable. So we will track the bounding boxes for the root node and its parts over time by Hidden Markov Model (HMM) model. The HMM model captures the spatial constraints on the movement of bounding boxes between frames and the transition between the type of pose templates (label of index in the pose template dictionary). The details inside each part are considered intractable, and thus, we calculate the histogram of the flow without pixel level correspondence.

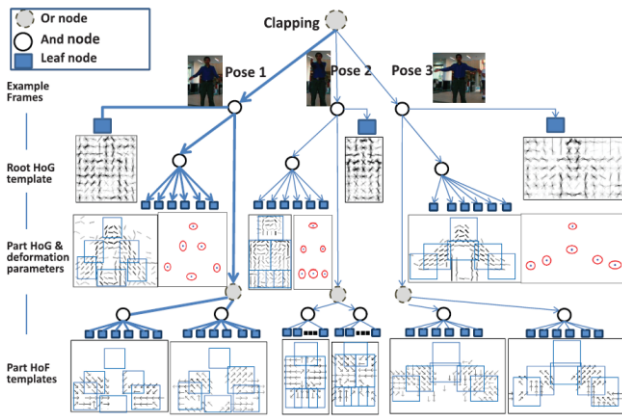


Fig. 3: A hand-clapping action includes six moving pose templates and is represented by a 2-level AND-OR tree structure.

Each moving pose consists of a shape template and a motion template. The shape template has one “root template” with HOG features for the entire bounding box at a coarse level and some (six in this case) “part templates” with finer scale HOG features. These part templates can deform with respect to the root template governed by 2D Gaussian functions whose mean and variance are illustrated by ellipses. A figure template can be associated with some (two in this case) motion templates for different movements of the parts and the motion of parts is represented by HOF features.

Each pose is represented by a shape template (denoted by ST) and one of the two motion templates (denoted by MT) depending on whether the arms are swinging upwards (or forwards) or downwards (or backwards). Thus, this one generates a total of six moving pose templates organized in a hierarchical AND-OR tree structure.

$$\Omega_{mpi} = \{MPT_i = (ST_i MT_i); i = 1, \dots, n\}$$

Each shape template ST<sub>i</sub> consists of a root template ST<sub>i0</sub> for the coarse level human figure and m templates ST<sub>ij</sub>; j = 1... m for body parts:

$$ST_i = (T_{i0}, T_{i1}, \dots, T_{im})$$

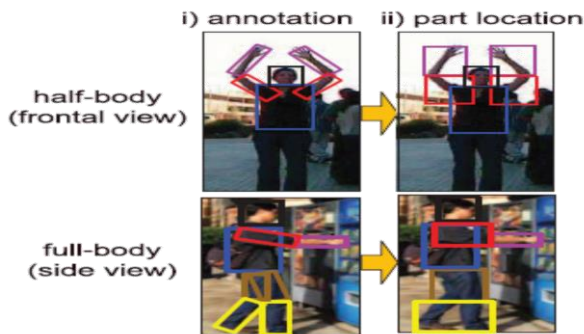


Fig. 5:

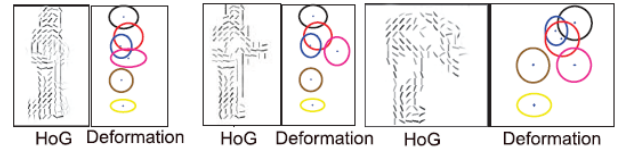
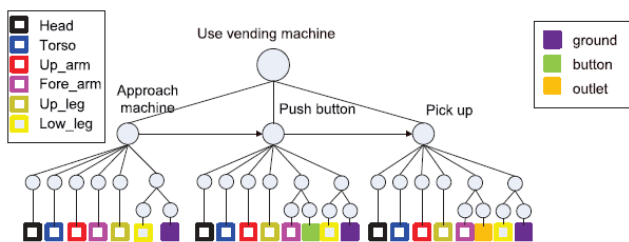


Fig. 6:

(a) The APT includes three sequential steps (or MPTs): walking on the “ground” to approach the vending machine, pushing “button” at the vending machine, and picking up the merchandise at the “outlet.” The open squares are the body parts and we add three new nodes in solid squares for the contextual objects: ground (in purple), button area (green), and outlet (yellow). These objects have spatial relations with the low-leg, forearm nodes. In the second row, we display the learned HOG templates parts and their typical deformations by the ellipses. The bottom row of the figure shows the actual detection results of form parts on a motion sequence. (b) The semantic maps for contextual objects generated by detected actions and person’s body parts. Different colors indicate body parts and the objects that city hall with them. The purple region is where feet are detected, and thus implies a standing point on the ground. The green region is where forearms are detected in a “pushing button” MPT, and thus implies a button. The yellow region is where forearms are detected while the person is in a picking-up MPT, and thus implies the outlet.

#### IV. COFFEE AND CIGARETTE DATA SET

The Coffee and Cigarette data set is collected mostly from the movie “Coffee and Cigarettes” and some training data are from a different movie named Sea of Love and a controlled lab video. It has 11 short stories each with different scenes and actors. This data set focuses on two action classes: “drinking” and “smoking.” For the drinking actions, there are 106 samples for training and 38 for testing. For the smoking action, there are 78 samples for training and 42 for testing. Performance evaluations on the MSR data set. (a) Detection performance of three action snippets using MPT model. The “boxing” class is the best because some poses of the “waving” and “clapping” classes are easily confused with each other. (b) Performance comparison in terms of the area under Precision-Recall curve against the amount of annotated key frames used to initialize training. Here, 100 percent means that all the training frames are key frames. (c) Performance against the number of poses used for each class. Comparisons between using full model and using only “shape” or “motion” parts are also included. There are three numbers in each cell, which represent, from left to right, “clapping,” “waving,” and “boxing,” respectively. Detection performance of action snippets on the coffee and cigarette data set. We only show the bounding boxes for the human head and the boxes on the contextual objects: the hand holding a cup or cigarette. We manually choose and annotate 40 key frames (evenly spaced in time) with six upper-body parts and one contextual object (i.e., the hand holding a mug or a cigarette).



Fig. 6: Contextual objects in action recognition.



Fig. 7: Detection performance of action snippets on the coffee and cigarette data set.

### V. CMU HUMAN-OBJECT INTERACTION DATA SET

The CMU human-object interaction is comprised of 60 videos with 10 actors performing six different actions, that is drinking from a cup, spraying from a spray bottle, answering a phone call, making a phone call, pouring from a cup, and lighting a flash light the videos are split into five color of the action-object, which requires manual pixel wise segmentation of the object at training time. It is also interesting to see that APT with latent parts method performs very poorly on this data set (even worse than MPT). This is in fact understandable because, while C&C data set is mainly about localization, CMU data set is a classification test. Therefore, it is much more important to distinguish the subtle difference between classes. This confirms our intuition that modeling contextual objects is the key for solving such a problem.

#### A. UCLA Contextual Action Detection Data Set:

Our data set consists of videos of 10 scenes taken from everyday living places such as campus plaza, food court, office, corridors, and so on. Each of these videos contains about a dozen instances from the following event list:

- 1) Purchase from a vending machine,
- 2) Use an elevator,
- 3) Throw trash into a can,
- 4) Use a water dispenser,
- 5) Pick up newspapers from a paper-stand, and
- 6) Sit down on a chair then get up and leave.

Most of these categories involve multiple action phases and involve contextual objects. Illustrate a snapshot of the data set. All the events are annotated with six body parts: "head," "torso," "upper arm," "lower arm," "upper leg," and "lower leg." What made this data set different special is that its contextual objects are static in the background. Even though it is very hard to directly detect some contextual objects such as "vending machine button," we can exploit the fact that these objects can be represented as hot zones within the scene. Therefore, we do not directly annotate the

contextual objects for this data. Instead, we divide actions in each video into two halves and use the first half to learn a semantic map of hot-zones, that is, to build a 2D histogram for each part of interest (in this paper, we consider two parts lower arm and lower leg). Since these semantic maps are learned from the "ground truth," they are very accurate. Even if without ground truth, we can imagine ways to automatically learn



Fig. 8: Snapshots from the UCLA action data set.

Event	APT-FULL	APT <sub>w</sub> /latent parts
Vending machine	82%	43%
Elevator	92%	67%
Throw trash	86%	58%
Water dispenser	87%	62%
News-stand	89%	74%

Table 1: Detection Performance on the UCLA Data Set

Applying these semantic maps, we then use the second half of our data for testing. To minimize the effect of over fitting, we apply a fivefold cross validation by randomly choosing different combinations of training and testing actions. The average detection precision is measured for six event classes as shown in latent parts method does not use the contextual information, it is much worse than the full APT model.

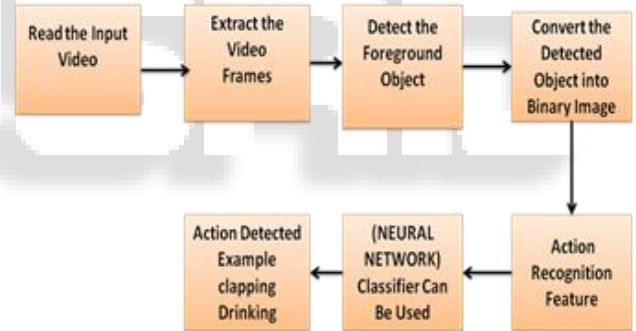


Fig. 9: Block Diagram of Modelling and Detection of Action

Optical flow or optic flow is the pattern of apparent motion of objects, surfaces, and edges in a visual scene caused by the relative motion between an observer (an eye or a camera) and the scene. The concept of optical flow was introduced by the American psychologist James J. Gibson in the 1940s to describe the visual stimulus provided to animals moving through the world. James Gibson stressed the importance of optic flow for affordance perception, the ability to discern possibilities for action within the environment. Followers of Gibson and his ecological approach to psychology have further demonstrated

The role of the optical flow stimulus for the perception of movement by the observer in the world, perception of the shape, distance and movement of objects in the world and the control of locomotion. Recently the term optical flow has been co-opted by roboticists to incorporate related techniques from image processing and control of navigation, such as motion detection, object segmentation, time-to-contact information, focus of expansion calculations, luminance, motion compensated encoding, and stereo disparity measurement.

Object recognition includes the process of determining the object's identity or location in space. The problem of object or target recognition starts with the sensing of data with the help of sensors, such as video cameras and thermal sensors, and then interpreting these data in order to recognize an object or objects. We can divide the object-recognition problem into two categories: the modeling problem and the recognition problem

Image segmentation is the process of partitioning a digital image into disjointed, meaningful regions. The meaningful regions may represent objects in an image of three-dimensional scene, regions corresponding to industrial, residential, agricultural, or natural terrain in an aerial recognizance application, and so on. A region is a connected set of pixels and the objects are considered either four-connected, if only laterally adjacent pixels are considered, or they can be eight-connected, if diagonally adjacent pixels are also considered to be connected.

## VI. SIMULATION RESULTS AND DISCUSSIONS



Fig. 10: Simulated result video vip car and speed detected using matlab

## VII. DISCUSSION AND FUTURE WORK

Human actions are complex patterns and most of the current data sets are quite constrained and there is still a long way to go before robust and general vision system can work on generic scenes. Our model is limited and, thus, can be extended in the following aspects. First, it is two-dimensional and thus view-dependent. For different views, more pose templates are needed. Second, it does not have rich appearance model to account for human clothes at high resolution. The HOG feature for each body part needs more than one templates to account for the intraclass variations. We plan to address the above two problems by using the And-Or graph representation, where different views are modeled by Or-nodes, and each node in the And-Or graph terminates in low resolution. Third, we should also learn the action and contextual objects in 3D model, for example, using Kinect as training data. This will help the action recognition to new scenes for robust performance. Fourth, we are connecting the action recognition with long term event recognition with goal and intent reasoning

## REFERENCES

- [1] Bobick. A And Davis. J “The Recognition Of Human Movement Using Temporal Templates,” (Mar. 2001). Ieee Transation. Pattern Analysis And Machine Intelligence.
- [2] Dolla ´R, Rabaud. V Cottrell. G And Belongie.S “Behavior Recognition Via Sparse Spatio-Temporal Features,”(2005) Proceeds. Ieee Int’l Conference. Computer Vision Workshop Visual Surveillance And Performance Evaluation Of Tracking And Surveillance (Vs-Pets).
- [3] Essa. I And Pentland.A “Coding, Analysis, Interpretation, And Recognition Of Facial Expressions,”( July 1997). Ieee Transation. Pattern Analysis And Machine Intelligence, Vol. 19, No. 7, Pp. 757-763.
- [4] Felzenszwalb. P Girshick. R Mcallester. D And Ramanan. D “Object Detection With Discriminatively Trained Part-Based Models,” (Sept. 2010) Ieee Transation. Pattern Analysis And Machine Intelligence, Vol. 33, No. 9, Pp. 1627-1645.
- [5] Joachims.T Finley.T And Yu.C “Cutting-Plane Training Of Structural Svms,”Machine Learning, (2009) Proceeds. Ieee Conference. Computer Vision And Pattern Recognition (Cvpr).Vol. 77, No. 1, Pp. 27-59.
- [6] Kovashka. A And Grauman,K “Learning A Hierarchy Of Discriminative Space-Time Neighborhood Features For Human Action Recognition,”(2010) Proceeds. Ieee Conference. Computer Vision And Pattern Recognition (Cvpr).
- [7] Kovashka. A And Grauman. K “Learning A Hierarchy Of Discriminative Space-Time Neighborhood Features For Human Action Recognition,”(2010) Proceeds. Ieee Conference. Computer Vision And Pattern Recognition (Cvpr).
- [8] Laptev. I Marszalek. M Schmid.C And Rozenfeld.B “Learning Realistic Human Actions From Movies,”(2008) Proceeds. Ieee Conference. Computer Vision And Pattern Recognition (Cvpr).
- [9] Marszalek. M Laptev. L And Schmid. C “Actions In Context,” (2009). Proceeds. Ieee Conference. Computer Vision And Pattern Recognition (Cvpr).
- [10] Schindler. K And Gool . L.V “Action Snippets: How Many Frames Does Human Action Recognition Require?”(2008).Proceeds. Ieee Conference. Computer Vision And Pattern Recognition (Cvpr).
- [11] Yang.W Wang.Y And Mori.G “Recognizing Human Actions From Still Images With Latent Poses,”(2001). Proceeds. Ieee Conference. Computer Vision And Pattern Recognition (Cvpr), Pp. 2030-2037.
- [12] Yang. Y And Ramanan. D “Articulated Pose Estimation With Flexible Mixtures-Of-Parts,” (2011). Proceeds. Ieee Conference. Computer Vision And Pattern Recognition (Cvpr).