1. Inferring Human Intent from Video by Sampling Hierarchical Plans (2016) (https://dspace.mit.edu/bitstream/handle/1721.1/138080/IROS_2016_camera.pdf?sequence=2&isAllowed=y)
    a. They propose a system for inferring a human's hierarchical intent through video
    b. An example of hierarchical intent would be:
        i. Wants to make coffee
            1. Gets coffee beans
                a. Gets up
                b. Walks to cupboard
                c. Opens cupboard
                d. Reaches for beans
                e. …
            2. Gets water
                a. Steps similar to what's mentioned above
                b. …
            3. Heats water
            4. …
    c. The main component of the system is an And-Or graph which models the actions the system thinks the human wants to make. The And-Or graph's probabilities are updated continuously based on new frames that the system sees
    d. The systems need object detection and skeleton tracking

2. Activity Forecasting (2012) (https://link.springer.com/content/pdf/10.1007/978-3-642-33765-9_15.pdf)
    a. They propose a system to predict the future action of a person based on video
    b. The system uses the surrounding environment of the person to predict their future actions. Concretely, it uses semantic scene labeling to extract physical scene features such as pavement, grass, tree, building, etc. and uses them as influencing factors in a person's future action
    c. Specific task performed is trajectory-based activity analysis using visual input. The trajectory of pedestrians in public spaces are analyzed
    d. It does not use traditional motion-based approaches but rather find relations between features of the environment and pedestrian trajectories
    e. The system they propose retrieves a distribution over all possible future actions

3. First-Person Activity Forecasting with Online Inverse Reinforcement Learning (2017) (https://openaccess.thecvf.com/content_ICCV_2017/papers/Rhinehart_First-Person_Activity_Forecasting_ICCV_2017_paper.pdf)
    a. The proposed system predicts the long term goals of a person through a first person view of their actions

b. The long terms goals include what the person's next action is, what the person's destination is and what their final objective is
c. The system uses online learning as it must learn incrementally as new footage arrives
d. It learns (incrementally) spatial and semantic intentions (next action and destinations) by tracking goals that the person achieves and predicts the person's future among the set of goals it has seen

4. Where and Why Are They Looking? Jointly Inferring Human Attention and Intentions in Complex Tasks (2018) (https://openaccess.thecvf.com/content_cvpr_2018/papers/Wei_Where_and_Why_CVPR_2018_paper.pdf)
    a. The authors propose a system to infer where a person is looking (attention), why they are looking there (intent) and what they are doing (task) when given RGBD video
    b. They give a single hierarchical framework to represent all three
        i. A task is a sequence of intents which transition to each other
        ii. An intent is composed of the person's pose, attention and surrounding objects
    c. The systems uses a beam search algorithm to infer the three things

5. What Will I Do Next? The Intention from Motion Experiment (2017) (https://openaccess.thecvf.com/content_cvpr_2017_workshops/w1/papers/Zunino_What_Will_I_CVPR_2017_paper.pdf)
    a. The paper explores the debate of whether video-based prediction and classification of actions can replace 3D kinematics-based systems (3D kinematics being obtained through motion capture)
    b. A video-based system is proposed which is shown to be as effective as 3D kinematics
    c. Video sequences of people grasping a bottle are used to forecast the intent behind reaching for the bottle (to pass it, place it in a box, drink from it, etc.)
    d. The system does not use any contextual cues but rather relies solely on the movement of the person (this is because they are comparing to 3D kinematics, which has no contextual cues)

6. Max-Margin Early Event Detectors (2013) (http://www.ca.cs.cmu.edu/sites/default/files/MMED_IJCV14.pdf)
    a. The paper proposes a framework for training temporal event detectors to recognize partial events. Being able to recognize partial events before they have been completed means this system can perform early event detection
    b. Early detection means to detect an event after it has started but before it ends
    c. The system is based on a Structured Output SVM
    d. An interesting use of early detection mentioned in the paper is to use early detection in conversational robots. The robot would need early detection during a conversation to detect the facial expressions and hence emotional state of the person. This would enable it to produce apt responses
    e. EmotionNet Nano: An Efficient Deep Convolutional Neural Network Design for Real-time Facial Expression Recognition (2020) (https://arxiv.org/pdf/2006.15759.pdf) is a more recent paper which also proposes a system for recognition of facial expressions

7. Audio-Visual Understanding of Passenger Intents for In-Cabin Conversational Agents (2020) (https://aclanthology.org/2020.challengehml-1.7.pdf)
    a. The authors propose a multi-modal dialog system which will operate in autonomous vehicles
    b. The system understands user intent based on both spoken interactions and input from the vehicles vision and audio systems

8. MMIU: Dataset for Visual Intent Understanding in Multimodal Assistants (2021) (https://arxiv.org/pdf/2110.06416.pdf)
    a. A new dataset of images and corresponding natural language questions is introduced
    b. This is meant to model the visual context based questions that a multi-modal assistant can receive
    c. The paper also gives a multi-class classification model that uses visual features from the image and textual features from the question to infer what the intent of the question is
    d. The model provided by the paper does not significantly outperform text-only models. The authors attribute this to poor fusion of text and image modalities