1. **Summary papers**
   Vision-Language Pre-training: Basics, Recent Advances, and Future Trends (2022) (link)
    a. Survey paper of techniques which involve both vision (image or video) and natural language

2. **Papers for online object detection in videos**
   1. Online Video Object Detection using Association LSTM (2017) (link)
   2. Wanderlust: Online Continual Object Detection in the Real World (2021) (link)
   3. Video representation learning through prediction for online object detection (2022) (link)
   4. Voting for Voting in Online Point Cloud Object Detection (2015) (link)
   5. FedVision: An Online Visual Object Detection Platform Powered by Federated Learning (2020) (link)
   6. LiDAR-based Online 3D Video Object Detection with Graph-based Message Passing and Spatiotemporal Transformer Attention (2020) (link)
    a. Fist paper proposes the Association LSTM which unlike a regular LSTM is able to learn the association between frames in a video
    b. The third paper proposes a real-time object detection technique which first pre-trains the model on the task of predicting the feature map of the next frame given the frame up till that point
    c. Fourth and sixth papers propose a system for laser-based object detection
    d. The fifth paper proposes a federated learning technique which could be useful if multiple robots are deployed and are using a shared object detection model

3. **Papers on retrieving objects from an image given a natural language description**
   1. Modeling Context in Referring Expressions (2016) (link)
   2. Generation and Comprehension of Unambiguous Object Descriptions (2016) (link)
   3. Natural Language Object Retrieval (2016) (link)
   4. Grounding of Textual Phrases in Images by Reconstruction (2017) (link)
   5. Conditional Image-Text Embedding Networks (2018) (link)
    a. Papers explores comprehending natural language which refers to a given image
    b. Natural language can describe an object in an image and the system needs to recognise the object and find it in the image (usually by selecting the bounding box around it)
    c. Second paper gives a dataset of objects with referring expressions (link)
    d. The fourth paper proposes a visual attention-based LSTM model to find objects in images by only attending to the relevant portion of the image. They extend this system to be able to learn without grounding supervision (no bounding boxes around images). They do this by using an auto-encoder-like model which takes an image and a natural language description of an object and then finds the relevant portion of the image and finally tries to caption this portion to match the original natural language description

2. **Papers for object detection in videos**
   Context Matters: Refining Object Detection in Video with Recurrent Neural Networks (2016) (link)

3. **Papers for real-time text detection in videos**
   1. An End-to-end Video Text Detector with Online Tracking (2021) (link)
   2. Real-Time Scene Text Detection with Differentiable Binarization (2020) (link)
   3. MSER-based Real-Time Text Detection and Tracking (2014) (link)
      a. The first paper proposes a system in which text detection and text tracking in videos are separated into two different tasks
      b. The second paper proposes a binarization (assigning pixels to a pixel group) module which works together with a segmentation network to produce efficient text recognition (efficient enough to be used in real-time)

4. **Papers proposing datasets of images with the objects in them annotated**
   Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models (2015) (link)
      a. Gives a dataset with 158K captions, 244K co-references of the same object in an image and 276k manually annotated bounding boxes for mentions of objects
      b. The dataset can be used for finding natural language mentions of an object in an image
      c. Dataset at: link

5. **Papers on robot navigation through natural language commands**
   1. Robot Language Learning, Generation, and Comprehension (2015) (link)
   2. Learning perceptually grounded word meanings from unaligned parallel data (2014) (link)
   3. Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments (2018) (link)
      a. The first paper covers three tasks relating to remote control of a robot through natural language
         i. Acquisition: The robot is driven through various paths and natural language descriptions of the paths are given. Through this the robot must learn the correspondence between natural language tokens and objects in the floorplan known to the robot
         ii. Generation: The robot is driven through various paths and it must generate natural language descriptions of the paths it is driven on
         iii. Comprehension: Given a natural language description of a path, the robot must generate a path in the floorplan for it to follow
      b. The second paper covers the acquisition and comprehension tasks

6. **Papers on custom architectures for video tasks**
   Long-term Recurrent Convolutional Networks for Visual Recognition and Description (2015) (link)
      a. Paper proposes a Long-term Recurrent Convolutional Network (LRCN) which combines spatial depth of CNNs and temporal depth of RNNs
      b. The LRCN can be used for tasks like video activity recognition

7. **Papers on gaining visual knowledge with weakly supervised learning**
   1. Learning Everything about Anything: Webly-Supervised Visual Concept Learning (2014) (link)
   2. Webly Supervised Learning of Convolutional Networks (2015) (link)
      a. The first paper proposes a framework for automated training of models to recognize a wide range of visual phenomena in images (like actions, interactions, attributes). Images can contain 4 types of concepts: object, scene, event and action
      b. Given a term, eg. 'horse', the system goes through online resources such as Google books and learns a vocabulary of all the words under the umbrella term 'horse'
      c. When the system discovers a term, it uses a web image search with the term to retrieve relevant images. It then uses a weakly supervised object localization technique to find the object in the image. Finally the model uses the term it discovered along with the object-annotated images to learn
      d. The second paper proposes a method of training CNNs using two step curriculum learning, i.e. they first train the CNNs on easy images and then adapt this pre-trained CNN to harder images
      e. They also propose a similar framework for training R-CNNs for object localization

8. **Papers on weakly supervised object localization**
   1. Multi-fold MIL Training for Weakly Supervised Object Localization (2014) (link)
   2. On learning to localize objects with minimal supervision (2014) (link)
      a. The papers propose different techniques to learn object localizers (which draw bounding boxes around objects in an image) using weak supervision
      b. Weak supervision in this context means that images are annotated only by what object is present but the bounding box is not given

9. **Papers on pre-training vision models to transfer them to downstream tasks**
   Learning Transferable Visual Models From Natural Language Supervision (2021) (link)
      a. Paper proposes that pre-training vision models with an NLP-Vision task will enable it to be transferred to another vision related task
      b. According to the paper the model is able to learn good image representations during the pre-training which enables zero-shot transfers of the model to other tasks
      c. The pre-training task is to match captions with the correct corresponding images
      d. Some example downstream tasks given by the paper are OCR, action recognition in videos and object classification
      e. The pre-trained model (trained on 400 million images) is available to the public

10. **Papers on finding a moment in video based on a natural language description**
    Proposal-free Temporal Moment Localization of a Natural-Language Query in Video using Guided Attention (2020) (link)
    a. Technique to find the moment (set of frames) in a video which matches a given natural language description of an action