Ideas:

You tube video clip search using Visual Question Answering or Action Recognition (going to specific instance of Videos.)

Action Recognition,

Visual Question Answering. (<http://vqa.cloudcv.org/>)

<https://tryolabs.com/blog/2018/03/01/introduction-to-visual-question-answering/>

Tensorflow object detection:  
<https://towardsdatascience.com/analyse-a-soccer-game-using-tensorflow-object-detection-and-opencv-e321c230e8f2>

Matterport Mask RCNN:  
<https://github.com/matterport/Mask_RCNN>

**Main Idea:**

By looking at a video, intelligently scanning and pinpointing the frames related to the search query.  
Example: In soccer video, penalty being shot.  
In cooking video, vegetables being cut.

So ultimately, it is a Visual Question Answering System for each frame in a video, and we can search for the frame corresponding to the answer.  
But, some frames may not at all be relevant to the question. You may have to generate rules based on question, so that only relevant frames are captured for the question.

Following gives good introduction to VQA:  
<https://tryolabs.com/blog/2018/03/01/introduction-to-visual-question-answering/>

**What is Scene Recognition:**

Scene recognition is the task of getting to know where the scene in the picture is happening, what is it about, what’s happening there, etc. We can have many adjectives for scenes too. It’s approached as a classification problem.

Scene recognition can be used for relevance detection.  
What can be our assumptions?

**Steps thought of now:**

1. ​Understand the most important part of the question and decipher what type of answer is it asking (count, color, activity). This may be done by pre-trained LSTM networks / bag of words model.
2. Find out the relevant frames in the video about the question (if question about classroom is asked, find the frames with classroom in it).  
   **Key frame identifying in videos (relevant frames)**.
3. From these frames, answer the question through VQA pipeline (detecting objects if count is asked, answering about color if asked, etc.).

**Assumptions and Dilemma:**

1. As of now, I may take dataset with different kinds of scenes, so that I can at-least apply the pipeline of scene recognition. Because of this, I can do the Step 2 mentioned above in easier way. Then I will do Step 3.
2. What if scenes are very close by throughout the video? What is the second best category you would choose? (Maybe the environment, or elements of scenes, or what is happening in that scene.)
3. What kind of questions would be asked? I can ask like show me the frames where some scene / activity is happening (like dribbling goal and celebration), or I can ask some specific information from the scene (like frame where champagne bottle was thrown up while celebrating, and which champagne bottle it was). In the second example of champagne bottle, I should find relevant frames as the ones where there is celebration and where champagne bottle is thrown. If scene level information is contains that much of information, then it’s good enough, else we may have to take the pipeline of Visual Question Answering. In this pipeline, step 1 would be to find scenes with celebration, and second would be – is champagne bottle thrown here (may also have to see if there is champagne bottle first of all. This may require object detection too).

**Dillemas:**

1. What kind of dataset do I start with? Do I start with only scene dataset first, with just the task of scene recognition and make it more complicated, or do I start with VQA dataset, or do I do a combination training of both datasets for this kind of goal?
2. If I am approaching to do a paper from this apart from self learning, what should be the approach, which domain to apply on, and what are the minimum results expected (qualitatively and quantitavely)?

Choose video of one kind (like surveillance, sports, soccer, cooking, chemistry lab dataset (like when some test tube got broken), driving, self attached camera video. etc.)

Search terms:  
video understanding

video indexing

video captioning

action recognition

content based video retreival

**Video Question Answering** (TVQA dataset)   
Key frame identifying in videos (relevant frames).

GTEA dataset, Epic kitchen dataset

thecvf.com

**Immediate tasks:**

1. Fix a domain from above mentioned ones
2. Read research papers on Videos in that domain  
   Key note frame detection.  
   Video question answering papers  
   What they are doing in GTEA, and what not being done in Epic, and what is being done in Epic. (action and scene selection)  
   **Dhruv Batra**, **Devi Parikh, Tamara Berg** (professors at Georgia Tech)
3. Search for relevant datasets, or see if you can collect it.

**Some relevant papers:**

<https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=410146>

**Plan:**  
1. Dataset, how to collect and all.

**Datasets popular for egocentric vision:**

Epic-kitchens, VLOG, GTEA

**GTEA Dataset:**

This dataset is mostly on cooking videos from Georgia tech university. This dataset is mostly being used for Action Recognition in first person video / egocentric vision, detecting hands, gaze anticipation.  
Some work also has been done in interaction of hand grasps with objects in questions. These are forms of action recognition which we can use for defining actions.

**EPIC Kitchens Dataset:**

This dataset comes from University of Bristol. Similar to GTEA dataset, action recognition, human behavior prediction (trajectory forecasting, action forecasting), TFGAN (text to video synthesis), Skill determination, recognition and localization of food in cooking video, Deep diary lifelogging, have been carried out.

**Similar work done in some areas:**Action Recognition, Visual Question Answering, Video Question Answering, Object Detection (maybe), Video Indexing.

**Video Indexing:**It’s basically indexing videos such that they can be searched easily from databases by queries.

**Visual Question Answering:**Visual Question Answering means answering abstract questions about image. It was significantly explored by Georgia tech university (Dhruv Batra, Devi Parikh) and Virginia Tech university.

**Video Question Answering:**The problem statement which we are doing.  
Another paper:  
**IQA: VQA in Interactive Environment (Joseph Redmon)** deals with VQA on egocentric videos.

**Next Steps:**

1. VQA paper and what dataset and what questions were used in that.
2. In GTEA and Kitchens dataset, look at dataset and frame questions (what kind of questions to be answered)
3. Look at Dhruv Batra’s work.
4. Read Video Question Answering papers
5. Search Video Question Answering with GTEA / Epic (No results afaik)

**VQA Original paper by Dhruv Batra:**

**TGIF QA Paper:**

Tasks encompassed in this paper:  
A) *State transition* (what happened before X)  
 What does he do after sitting?  
 Facial expression transition, action transition (sitting to standing), object properties (empty to full)  
  
B) *Repetition Count of Action* (How many times is the man shaking the hand)  
  
C) *Which actions repeated x times* (What is being repeated 5 times?)

**Some Notes on Our problem**

We also have to recognize actions and state transitions, but we have to recognize interaction / action with an object. (Eg: taking out eggs from fridge, opening microwave / fridge, cutting tomatoes, peeling off fruit / vegetable skin with peeler, etc.)

**Things to study now:**

Natural Language Processing  
Deep learning (hyperparameter tuning, vanishing gradients, activations, batch norm, momentum, grid search)  
Neural network embeddings in CNN and word representations

**GOAL:**

What’s the problem statement? What kind of questions has Visual Question Answering has answered, and what kind of questions do we want Video Question Answering to answer? (Minimum expectations)