1. What parameters you decided to use for the provided example dataset

I experimented on various parameters using few sets of images that are similar, completely different and only slightly different. Then I decide a score threshold based on these different parameters.

The various parameters that I experimented to choose the threshold for similarity score (**score threshold**), are **Image threshold** used in the function compare_frames_change_detection() and the **minimum contour area** which is to be provided as parameter to the function compare_frames_change_detection(). All scores below the threshold are considered as similar.

The table below show the various experimented performed using the different parameters.

Based on the table we can see that Image threshold parameter has a greater effect in removing similar images than minimum contour area. We decide the score threshold based on these two parameters and we can decide how many similar images need to be removed.

	Parameters		
IMAGES DELETED	***SCORE THRESHOLD***	***GRAY IMAGE THRESHOLD***	***MIN CONTOUR AREA***
405	60000	45	2000
412	65000	45	3000
443	40000	75	2000
449	26000	100	2000
449	24000	100	3000
449	8000	125	2000
446	100	125	3000
474	1000	200	2000

2. How you found these values

Initially when I executed the task using the score threshold based on first observation, the dataset after removing similar images still contained images that look very similar by has a very slight difference. I decided to then perform the process again by setting the score threshold above the range of above mentioned slightly similar images. Because while training using an ML model these images still look the same and can introduce bias to the model and affect its performance. I decided to choose a gray image threshold of 100 in the function compare_frames_change_detection() and a minimum contour area of 3000 because from the table above we can see that a higher image threshold deletes more images than required.

3. What amount of duplicates script found with these parameters

449 duplicate images are deleted from the folder of 484. Only 35 image were pending in the dataset when image threshold of 100 was used and min contour area as 3000.

4. What you would suggest improving to make data collection of unique cases

Better?

Unique images can make the performance of the ML model better. Whereas duplicate or similar images in the dataset introduces bias into the model and the model will not be able to generalize to new data during evaluation.

We can perform more effective techniques like hashing to remove exact duplicates and the use the contour area method to find more similar images. Visualisation of images with highest similarity score or in a sorted order can also be performed so that we can decide the threshold on which images are to be removed and which images need to be retained.

5. Any other comments about imaging_interview.py or your solution?

In my solution I can remove appending contours and returning it in the function compare_frames_change_detection(), because I would need only the similarity score. This can reduce the execution time when computing similar images for entire dataset.

Also since the provided dataset is small, the current image size is not much of a problem for the entire process, but when it comes to huge dataset, I can resize the images to a smaller size in the pre-processing step and then compute the similarity score. But it can affect the setting of threshold for the similarity score, as I have only less range to choose from.

The current computation complexity is O(n^2) as I am computing an image with every other image present in the dataset. I can reduce the complexity by performing a linear search which reduces the complexity to O(n), which can be useful for large datasets. I have also provided a function for performing the linear search deleteSimilarImagesFromTempLinearly() using the same threshold. It does not have the best performance as the initial method, but we can reduce the size of the dataset considerably using the linear search first and then perform an O(n^2) search.

Other python libraries like Structural Similarity Index (SSIM) can also be used which only takes images as the parameter. In this case we don’t have to experiment with any parameters as the entire comparison is performed by the library and we need to just decide the score threshold.