There were a total of 7 main frames detected from the video. From each of the main frames, the highest probabilities resulted in the following information concerning the overall video:

people - 0.97585094

box - 0.9641634

crowd - 0.9976672

woman - 0.96527994

one - 0.99124813

one - 0.99289644

paper - 0.9346962