The paper applies the analogy of text retrieval for allocate an outlined object in a video. A collection of “visual words” are pre-computed descriptors in the subsets of video. And an object is like a sequence of query words. For showing the result, tf-idf and document ranking are both applied.

Noise and unstable region are removed if the regions are not tracked more than three frames. But the problem could be what velocity we should use for extract the frames of video. The frame rate is data-driven when we query a specific object. If we are querying a fast moving object, it might disappear in three frames. So the frame rate would affect the accuracy of matching.

A visual word represents a cluster of similar region for computation efficiency. The weight of region is calculated similar to tf-idf. So if a region appears in the low occurrences documents, then that region might contain interesting and useful information. Also by the idea of tf-idf, we can build a stop-list. A stop-list contains the very common regions that appear in many frames, such as background regions, and those regions give us little information and reduce the accuracy of matching. The result of matching is highly dependent on the relevant images for query image. If the number of relevant image is low, then it indicates the matching might be accurate.

The method also contains the spatial information of each match. Because we assume that the image are smooth. Each corresponding region of query region contains will votes by the number of matches. So that we can get a best match.