Video Google: A Text Retrieval Approach to Object Matching in Videos

Authors: Josef Sivic and Andrew Zisserman
Year: 2003
Journal: Proceedings Ninth IEEE International Conference on Computer Vision.

Report:

1. **Abstract:**

"We describe an approach to object and scene retrieval which searches for and localizes all the occurrences of a user outlined object in a video. The object is represented by a set of viewpoint invariant region descriptors so that recognition can proceed successfully despite changes in viewpoint, illumination and partial occlusion. The temporal continuity of the video within a shot is used to track the regions in order to reject unstable regions and reduce the effects of noise in the descriptors. The analogy with text retrieval is in the implementation where matches on descriptors are pre-computed (using vector quantization), and inverted file systems and document rankings are used. The result is that retrieval is immediate, returning a ranked list of key frames/shots in the manner of Google. The method is illustrated for matching on two full length feature films."

2. **Introduction:**

The article's objective is that to achieve proximity and the speed of object recognition through text retrieval approach. Text retrieval approach is a form of finding the documents based on the words with maximum approximation and speed, which is used by Google by this paper had been published. Identifying the object can always be difficult because that the contrast, viewpoint and brightness can be changed for every frame. So, the authors ought to use the approach for it's simplicity and best results compared to the previous works of object recognition. So, they have trained the model with a movie sequence of scenes and compared the whole movie for the object recognition with the sample, for required image detection.

3. **Content and Approach:**

- The text retrieval approach is one of the mainly used approach for document recognition and finding from the database.
- Here the images are documents, required picture is the word and total movie is the database in analogy.
- The text retrieval approach consists of the following procedure to extract the document:
  - Parsing the document into words.
  - Taking the stems of the words of all the documents.

    - Eliminate the common word such as"is" and "the" etc.,
    - Collecting the frequency of the stem words and creating a quantized vector of the document based on the frequency.
    - Inverted file approach is used to increase the efficiency of document search as defined below.
- **Viewpoint invariants** are resolved by combining the two approaches called as "Shape Adaptation"(SA) and "Maximally stable"(MS). They have their own features, where we combine them for the better results.
    - **Shape Adaptation:** This is process of creating the ellipses around the selected point. It iteratively creates the ellipses based on the center and the scaling of the interested image.
    - **Maximally Stable:** These are also determining areas based on the intensity threshold of the interested point of the image.
- The reduction and rejection of the noise is aggregated over a sequence of the frames of the scenes.
- The visual vocabulary is built by the vector quantization of the descriptors of clusters, which are analogous to the words of the document.
- The vocabulary is prepared from the subpart of the movie and analyzed with the remaining movie for the recogntion.
- The implementation of the vector quantization of the scene is done as follows:
    - The mean of the frames are calculated initially.
    - The unstable regions, which are less than 10% of occurence, are eliminated from the covariance matrix.
    - The Mahalanobis distance is calculated. This enables to decrease the noise of the process.
- K-means algorithm is used repetitively with different initializations and they have taken the best results out of all.
- They have used both SA and MS to cover the major part of the document which are independent regions of the scene.
- Normally in the text retrieval approaches the weightage is taken by the frequency of the words in the document. Here, they have selected the standard weightage known as "term frequency - inverted document frequency".
- **Stop list** is used to repress the very frequent visual vocabulary and ended in very good results.
- **Spatial consistency** is another procedure of comparing the ranking of the documents and search word to represent the relevancy of the document adopted by Google. This is also taken into practise in the approach.

4. **Results:**
    - The retrieval performance is measured by the rank of the matrix, which compares $N_{relevant}$ and the total size of the image.

- The graph is drawn between the rank and the scene numbers for the procedures with SA, MS and SA+MS independently.
- This graph had given very good reason to approve both combinedly to represent the performance of the approach.
- The table is drawn in between the ranks with different procedures, which doesn't show much difference.
- At some scenes, in between they have got huge rank, which are said to be failed cases. These are happened because of the enormous change in the viewpoint in consecutive frames.
- To be concluded, it had been best approach which had not resulted in false negatives and had very good precision.

5. Discussion and Thoughts:

- There are very limited assumptions taken throughout the advancement and also left a considerable part of work for the future.
- The unstable regions are eliminated, which are below three frames, which is a trivial assumption.
- A failure occurred in the consecutive frames which has enormous change in viewpoint, which has to be dealt with. It is not a negligible obstacle however.
- All the terms had been explained in very decent manner and to the required knowledge.
- Although the object recognition in the data base became easy with the process, the process of updating visual vocabulary is required. That had been left to the future work.

6. Conclusion:

- On the whole the paper seems to be more interesting and pertinent, which used the analogous approach of text retrieval.
- It also showed a great commitment of results and the approach seems to be more assured.
- It had been one of the best approach which had not resulted in false negatives however throughout the procedure.
- The article had been a perfect evolution for the object recognition through the new approach and did not limited to many assumptions and errors.