Storyline Reconstruction for Unordered Images

Final Paper

Sameedha Bairagi, Arpit Khandelwal, Venkatesh Raizaday

Introduction:

Storyline reconstruction is a relatively new topic and has not been researched extensively. The main objective is to take a stream of images as input and re-shuffle them in chronological order. The recent growth of online multimedia data has generated lots and lots of unstructured data on the web. Image streams are generated daily on websites like Flicker, Instagram etc. and almost 400 hours of video is uploaded on YouTube on a daily basis.

In this paper, we try and implement an algorithm which uses the property of videos of being temporally adept to sort a stream of unordered images. The basic process is as follows:

- Generate key frames/video summary of a video from multiple instances of the same category.
- Cluster these key frames on the basis of the action being performed in them.
- Create a graph from these clusters using temporal data from the videos.
- Take an input stream of images and assign each image to its most probable cluster.
- Use the graph to assign ordering to the images.

In the following sections, we will try and go deep into each of the step mentioned above and discuss multiple approaches we implemented to do the same.

Background and Related work:

We looked at multiple sources to understand video summary generation. In [2], the author converts the optical flow of a video to the Clifford Fourier domain where they correlating each optical flow feature with spatial-temporal maximum average correlation height filters. This generates a set of highly relevant frames. In [3], the author introduces the concept of superframes, which basically are set of frames which have more changes in optical flow then the frames surrounding them. We used [4] in which, the main goal is to select from a video the most “significant” frames in order to broadcast, without apparent loss of content by decreasing the potential distortion criterion. The source code was made available by the authors and we tweaked it a little to use it ourselves.

The field of storyline reconstruction is relatively new and not much work has been done in this. We have extensively learned about this topic from [1] and our algorithm is inspired by their work. According to the authors, the strength of images over videos lies in that images are more carefully taken so that they capture the subjects from canonical viewpoints in a more semantically meaningful way but still images are fragmentally recorded, and thus the sequential structure is often missing even between consecutive images in a single photo stream. On the other hand,

videos are motion pictures, which convey temporal smoothness between frames. However videos suffer from issues like redundant information, backlit subjects, noisy data etc. So for each video the author finds k-nearest photo streams using naïve bayes nearest neighbor. Then builds a similarity graph between the video and the stream of images. Then a bunch of optimization techniques are used to condense the graph and shorten distances between related nodes.

The data set we have used is the UCF-101 dataset which is a compilation of videos from 101 different categories. We intentionally choose categories which do not have repetition involved as then it becomes difficult to assign temporal labels to frames for such activities.

Methodologies:

We used a range of frameworks to tackle this problem. We used supervised, unsupervised and semi supervised methods. We used feature extraction techniques like HOG, SIFT, LBP, SURF points and deep features. In this section we will try and cover all our experiments one by one.

Unsupervised Clustering:

In this method, generate key frames from multiple videos and transform each image into HOG feature representation. Then create clusters of these images using k-means algorithm.

Result: The algorithm failed as the clusters were made according to videos the images belonged to that is all images from one video end up in the same cluster.

In another modification, assign clusters to images one at a time. Take the first video and assign all of its images to different clusters. Then take a new image and convert it to HOG feature representation. Calculate distance of this image from each cluster and assign to the cluster with minimum distance.

Result: The algorithm does not fare much better as the cluster with maximum background on show gets assigned almost 90% of new images.

Semi-supervised Clustering:

Approach 1: For this method, create initial clusters for categories by taking key frames from 5-6 videos and assign them manually. The new cluster centers now become the mean of the HOG features of images in each cluster center. Take an image, convert it to HOG representation and calculate its distance from each cluster center. Assign to cluster with minimum distance.

Result: The algorithm has the exact same results as the one discussed in the section above. The cluster with maximum background on show gets assigned almost 90% of new images.

As an increment to the previous approach we improve the feature representation of each image and change the distance function for clustering.

1. Each image is represented as a histogram of HOG Features calculated at 4 and 8 window sizes. Histogram intersection is used for calculating the distance. In addition to this each

image is represented as a tiny feature image [5], which is 32x32 RGB representation of an image. Both descriptors are weighted equally. The algorithm remains the same as the previous one.

Result: An overall accuracy of ~15% is achieved which is good given the below baseline results in the previous approach.

2. Each image is represented in deep feature space by using the Overfeat package [6]. The distance function is same as the one used in majority of our methods. Algorithm remains the same as in previous methods.

   Result: An overall accuracy of ~21% is achieved which is the best that we can achieve in a unsupervised or semi supervised scenario.

Approach 2:

This approach can be briefly explained in following steps -

1. Here, instead of manually clustering for each key frame, we just simply assign a label for key frame which represents its' sequential order of its frame number. This is repeated all videos for a particular activity.
2. The number of frames can vary in different videos. This means it is not necessary to have same number of frames. Since, videos have similar actions, we assume that for a large collection of videos, the system would eventually be inclined to identify the correct sequence.
3. The images are then fed to CNN (overfeat) to generate feature vector and then, training is done.

Results

The results show an overall improvement when compared to other approaches but are still lower than the results for manually clustered images. However, it has shown a way forward and has reduced the huge time needed to manually cluster images.

Supervised Classification:

We used Convolutional Neural Networks (Overfeat) to generate the feature vector. The basic idea is to have a predefined set of images for each key frame of the video that had been extracted earlier and assign a label to each of these clusters. We have trained an SVM classifier for this using the SVM Multiclass library. It was interesting to note that, this improved the results by a large margin. However, it is a cumbersome process to manually cluster all Key frames.

Results and Discussions:

The following table summarizes the results from previous sections. The accuracies are calculated by calculating the ratio of the total number of images correctly assigned to their clusters to the total number of images for testing.

S.No	Feature	Accuracy (%)
1.	HOG + Tiny Images	14.62
2.	Deep Features (Semi supervised)	21.02
3.	Deep Features (Supervised)	52.00

As the results show the deep features tend to be a more powerful feature representation for our problem. The supervised approach largely outperforms its competitors but at the same time it is not feasible to manually annotate data coming from the internet as the categories are endless and the data is never ending.

The unsupervised approaches have accuracies below the baseline (<8%), it is largely due to the fact that the images with similar background have a higher similarity score. Also in HOG feature space background is represented in similar fashion and since the categories in our dataset have extensive backgrounds the results are never based on the activity being performed.

Accuracies are not high enough to give substantial results. There can be ways of improving these results and we will discuss this in the next section.

Conclusion:

The field of storyline reconstruction is an interesting one and not much research has been done in the field. We tried multiple approaches to come up with good results but hit unforeseen snags in each approach. None of the approaches are perfect and have a large scope for improvement. Some of the ideas are as follows:

- Weighted Distance Calculation: One of the major drawbacks of unsupervised approach was large backgrounds and small foregrounds. To solve this problem we can divide the image into cells and give higher weights to some specific cells.
- The specific cells can be found out by calculating top-k SIFT points in an image and giving importance to the cells that these SIFT points are located.
- Another approach for human specific activities could be to locate the person and create a bounding box around him and use this bounding box for distance function.
- Using L1 Normalized 3-Level Spatial Pyramid Histogram to create improved feature vector from HOG features.
- In the approach of Semi-Supervised Learning using the normalized Frame Number as class.

References:

1. Kim, G., Sigal, L., Xing, E.P.: Joint summarization of large-scale collections of web images and videos for storyline reconstruction. In: CVPR (2014)
2. Mikel Rodriguez, "CRAM: Compact Representation of Actions in Movies", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Fransico, 2010.
3. Creating Summaries from User Videos by Michael Gygli, Helmut Grabner, Hayko Riemenschneider and Luc Van Gool published in ECCV 2014, Zurich.
4. C. Panagiotakis, N. Ovsepian and E. Michael, Video Synopsis based on a Sequential Distortion Minimization Method, International Conference on Computer Analysis of Images and Patterns, 2013.
5. A. Torralba, R. Fergus, and W. T. Freeman. 80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition. IEEE PAMI, 30:1958–1970, 2008.
6. http://cilvr.nyu. edu/doku.php?id=software:overfeat:start