Efficient Temporal Mean Shift for Activity Recognition in Video

Carnegie Mellon

Yan Ke[1], Rahul Sukthankar[2,1], Martial Hebert[1]
[1]Carnegie Mellon University, [2]Intel Research Pittsburgh
{yke, rahuls, hebert}@cs.cmu.edu

Intel Research

Abstract

We propose a temporal mean shift algorithm that clusters spatio-temporal regions in video by exploiting the temporal nature of video. Extracting spatio-temporal regions is often one of the first pre-processing steps in an activity recognition system. Our key contribution is the insight that mean shift clustering can exploit the fact that there is typically very little change between successive video frames. Most of the pixels, and therefore the clusters, shift only slightly from frame to frame. Since mean shift is an iterative procedure, fewer iterations are required for convergence if the initial search is already close to the local optimum. Our temporal mean shift algorithm exploits the temporal similarity between successive frames by initializing the search using the modes found in the previous frame.

Standard Mean Shift Clustering

- Finds local modes in the clusters of points
- Iterative gradient ascent procedure
- Let f_i be a d-dimensional point in a set of n feature points $f_1 \cdots f_n$. We calculate

$$y_{(i,1)} = f_i \qquad (1)$$

$$y_{(i,j+1)} = \frac{\sum_{k=1}^{n} f_i g(||\frac{y_{(i,j)} - f_k}{h}||^2)}{\sum_{k=1}^{n} g(||\frac{y_{(i,j)} - f_k}{h}||^2)}, \qquad (2)$$

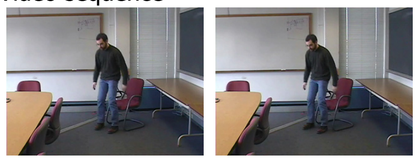
where $j = 1, 2, \ldots$, with kernel g, typically a Gaussian kernel, and bandwidth h.

- The mode is the limit of the series, where $y_{(i,j)}$ converges to a fixed point.
- The inner loop of mean shift requires a range-search of neighbors f_k near $y_{(i,j)}$, which is particularly time-consuming in high dimensions.
- Previous methods in optimizing standard mean shift focused on reducing the time needed for range-search of neighbors.

Temporal Mean Shift Clustering

- We observe that in most videos, there is very little change between successive frames.

Video sequence



Segmented image

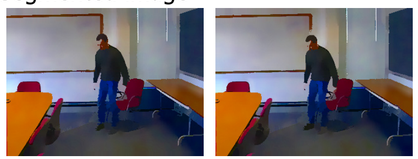
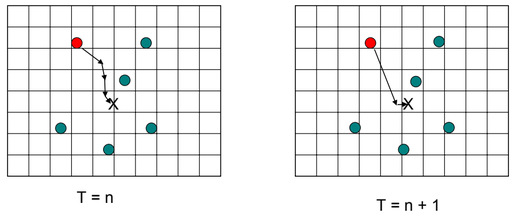


Illustration of standard vs. temporal mean shift

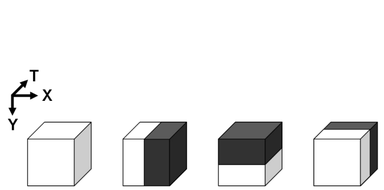


- Let f_i^t denote a feature point at pixel (x_i, y_i, t), and let y_i^t denote the mode found using mean shift, starting at (x_i, y_i, t).
- We expect the difference in features, $|f_i^{t+1} - f_i^t|$, to be small, and this will cause the difference in mode locations, $|y_i^{t+1} - y_i^t|$, to be small as well.
- Therefore, we initialize the search $y_{(i,1)}^{t+1}$ to be y_i^t, which reduces the number of iterations needed for convergence to the new mode.

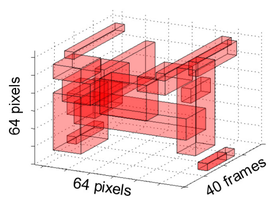
- Other considerations:
  - Pixels with large changes in appearance from the previous frame are not initialized with the mode from the previous frame.
  - To avoid long running accumulated errors, we restart the search every n frames.

Applications in Activity Recognition

- The clusters can be used to identify actions and events in video.
- In ICCV '05, we introduced a set of volumetric features for detecting visual events.
- We can train the volumetric features with these clusters to quickly detect them in video.

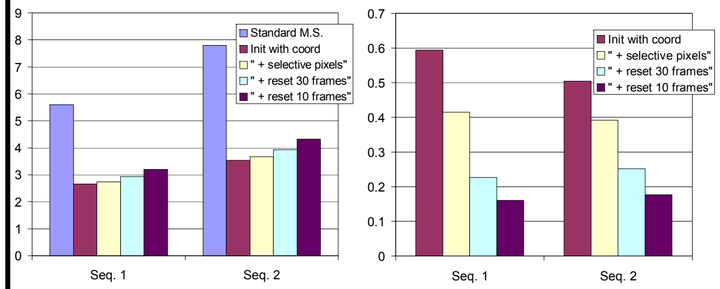


Spatio-temporal volumetric features



Example features trained on a hand wave action.

Results



- Segmentation on the LUV representation of the image.
- Error measured as the difference between our temporal mean shift and the standard mean shift segmentations, normalized to the average distance between the image color and the cluster center using standard mean shift.

Original Image | Standard M.S. | Init with coord | + selective pixels | + reset 10 fr.

Conclusions

- Introduced a novel temporal mean shift for clustering of images in video, based on the key observation that the pixels, and therefore clusters do not change much from frame to frame.

References

- D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *PAMI*, 2002.
- B. Georgescu, I. Shimshoni, and P. Meer. Mean shift based clustering in high dimensions: A texture classification example. *ICCV*, 2003.
- P. Indyk and R. Motwani. Approximate nearest neighbor – towards removing the curse of dimensionality. *In Symposium on Theory of Computing*, 1998.
- Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. *ICCV*, 2005.