Dedication...

Dissertation Presentation

Automatic Summarisation of Casually Captured Videos

Supervisor

Dr.S.Balasubramanian

Candidate

Varshaneya V
Regd No: 15013

March 28, 2017

What is video summarisation?

Video summarisation methods attempt to abstract the main occurrences, scenes, or objects in a clip in order to provide an easily interpreted synopsis.

Current instances of video summarisation

- Previews of movies, TV episodes.
- Summaries of documentaries, home videos.
- Highlights of games.
- Interesting events in surveillance videos (major commercial and security applications).

Techniques of video summarisation

- *Keyframe summaries based on clustering*
  - K-means clustering
  - Delaunay clustering
- *Skims summaries*
  - Summary based on interestingness
  - Summary based on interestingness, representativeness and uniformity

Techniques of video summarisation

Keyframe summaries

Keyframe summaries

Method to create feature vectors

- Frames are sampled.
- HSV histograms generated for each of the sampled frame.
- Feature vectors are created by applying principle component analysis to the 'hue' component of the image histograms.

Techniques of video summarisation

Keyframe summaries

K-means clustering

Keyframe summaries

K-means clustering

- Cluster the feature vectors by k-means clustering algorithm.
- The median of the clusters formed, will be keyframes which summarise the video.
- The parameter k has to be specified by the user and hence this is not suitable for batch processing.
- There are no closed form formulae for determining k for an unknown video.
- Redundant and background frames in the summary.
- Temporal coherence is not maintained.

Keyframe summaries

Summaries generated by k-means clustering algorithm

Test videos used were taken from Open-Video project. The summary for a couple videos from open-video database are shown in Figure below.

Figure : Summary generated by k-means algorithm with $k = 6$ for the video "America_ New_ frontier_ Segment 10.mpeg".

Figure : generated by k-means algorithm with $k = 5$ for the video "America_ New_ frontier_ Segment 3.mpeg". Note that there is a background frame and a repetition in the summary

Techniques of video summarisation

Keyframe summaries

Delaunay clustering

Keyframe summaries

Delaunay diagram

Delaunay triangulation for a set P of points in a plane is a triangulation $DT(P)$ such that no point in P is inside the circumcircle of any triangle in $DT(P)$.

- Delaunay triangulation on a set of points gives the delaunay diagram.
- Delaunay diagram is also the dual of Vorornoi diagram.

Keyframe summaries

Delaunay clustering

- Delaunay diagram is constructed for the feature vectors.
- Short and separating edges in the diagram are identified.
- Separating edges are removed so that clusters are formed.
- No user intervention as there is no parameter tuning and is suitable for batch processing.
- Temporal coherence is not maintained.

Keyframe summaries

Formulae used in delaunay clustering

- The mean length of edges incident at each point p_i is given by

$$localMeanLength(p_i) = \frac{\sum_{j=1}^{d(p_i)} \|e_j\|}{d(p_i)}$$

where $d(p_i)$ denotes the number of edges incident at the point p_i and $\|e_j\|$ denotes the length of each edge e_j.

- The local standard deviation of length of edges incident at point p_i is given by

$$localStandardDeviation(p_i) = \sqrt{\frac{\sum_{j=1}^{d(p_i)} localMeanLength(p_i) - \|e_j\|^2}{d(p_i)}}$$

- The global standard deviation is denoted by

$$globalStandardDeviation = \frac{\sum_{i=1}^{N} localStandardDeviation(p_i)}{N}$$

where N is total number of points.

Keyframe summaries

Formulae used in delaunay clustering contd...

- A short edge or intra-cluster edge is defined as

$$shortEdge(p_i) = \{e_j : \|e_j\| < localMeanLength(p_i) - globalStandardDeviation\}$$

- A separating edge or inter-cluster edge is defined as

$$separatingEdge(p_i) = \{e_j : \|e_j\| > localMeanLength(p_i) + globalStandardDeviation\}$$

Keyframe summaries

Metrics for evaluation of keyframes generated by delaunay clustering

- Significance factor for each keyframe gives a score to it corresponding to the size of the cluster it came from.

$$significanceFactor(I) = \frac{C_I}{\sum_{j=1}^{k} C_j} \quad (1)$$

  where C_I is number of frames in cluster I and k is total number of frames in the video.
- Compression factor for each video is an indication of the reduction in size with the summarised content as compared to the original set of frames.

$$compressionFactor = \frac{k}{N} \quad (2)$$

  where k is number of key-frames and N is the total number of frames processed.
- Overlap factor quantifies the extent of overlap between the summary generated by the algorithm and one generated by user.

$$overlapFactor = \frac{\sum_{k \in commonKeyFrameCluster} C_k}{\sum_{j=1}^{k} C_j} \quad (3)$$

Keyframe summaries

Evaluation of metrics for some of the summaries generated by delaunay clustering

Video name	Number of clusters	Significance factor	Compression factor	Overlap factor
America's new frontier segment 4	4	0.173 0.3946 0.3054 0.1081	0.1081	100
America's new frontier segment 10	4	0.1432 0.334 0.3963 0.1245	0.083	100
America's new frontier segment 3	6	0.1157 0.088 0.1065 0.1389 0.1528 0.0139	0.3704	100
The voyage of Lee segment 15	3	0.4802 0.4626 0.0573	0.1322	94.27

Keyframe summaries

Summaries generated by delaunay clustering algorithm

Figure : Summary generated by delaunay clustering for the video "America_ New_ frontier_ Seg10.mpeg"

Figure : Summary generated by delaunay clustering for the video "America_ New_ frontier_ Segment 3.mpeg".

Keyframe summaries

Comparison

Figure : The summaries to the left are from delaunay clustering and to the right are from k-means clustering for the video “America_ New_ Frontier_ Seg10.mpg”.

Figure : The summaries to the left are from delaunay clustering and to the right are from k-means clustering for the video “America_ New_ Frontier_ Seg4.mpg”.

Keyframe summaries

Discussion

- Summaries of k-means and of delaunay clustering are quite similar to each other when value of k is equal to number of clusters generated by delaunay clustering.
- This validates the keyframes generated by delaunay clustering.
- This is a proof of correctness of the approach towards generating keyframe summaries using delaunay clustering.

Techniques of video summarisation

Skims summaries

Skims summaries

Motivation

- Keyframes are still images - “motion” part of summary is lost.
- Computer does not what is “interesting” to a user.
- Temporal coherence needs to be maintained.

Skims summaries

Learning "interestingness"

- Used ground truth from "SumMe" dataset, which consists of 25 videos and 15 summaries for each video generated by users from different age and gender groups.
- Interestingness criteria used are:
  - Aesthetics - contrast, distribution of edges and colourfulness
  - Face and person
  - Attention - static and temporal saliencies.
  - Complexity
- Weights were regressed by least squares method (for the above criteria) from the scores given in the ground truth.

Skims summaries

Learning "interestingness" contd...

- Suppose we have N frames in a video, the score of interestingness for a frame is calculated by the formula:

$$i_k = w_0 + \sum_{i=1}^{N} w_i \cdot u_i + \sum_{i=1}^{N} \sum_{j=i+1}^{N} w_{i,j} \cdot u_i u_j \tag{4}$$

  where u_i is the score of feature i and w_i is the weight corresponding to feature i.
- The total interestingness $I(S_i)$ score of each of the superframe S_i is the sum of interestingness score of each of the frame in it given in formula:

$$I(S_i) = \sum_{k=n}^{m} i_k \tag{5}$$

Techniques of video summarisation

Skims summaries

Summary based on interestingness.

Summary based on interestingness

Generating summary

- Video is segmented using "superframe segmentation".
- 0-1 knapsack optimisation is done on superframes with the value of the superframe to be its interestingness score $I(S_i)$ and its weight to be the number of frames in it i.e. its length $\|S_i\|$.

$$\max_x \sum_{i=1}^{n} x_i I(S_i) \tag{6}$$

subject to

$$\sum_{i=1}^{n} x_i \|S_i\| \leq L_s \tag{7}$$

where $x \in \{0, 1\}$ with $x_i = 1$ indicating that the superframe i is selected and L_s is the desired length of the summary.

Summary based on interestingness

Results

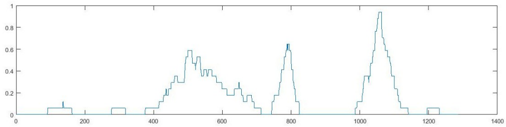

Figure : Plot of interestingness score vs frame number as given in the ground truth of video "Cooking.mp4".

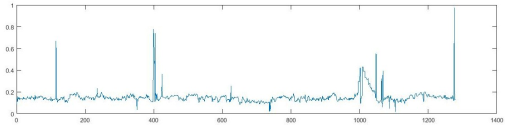

Figure : Plot of interestingness score vs frame number as obtained from learnt weights for the video "Cooking.mp4".

Summary based on interestingness

Running time

Video name	Video time (sec)	Summarisation time (Hr)
nasaani.mpg	30	0.39
America_ New_ Frontier_ Seg4.mpg	123	1.62
America_ New_ Frontier_ Seg10.mpg	161	2.15
games.avi	135	2.46
drone.mp4	253	3.75

Table : Summarisation time for videos using only interestingness criteria.

Techniques of video summarisation

Skims summaries

Summary based on interestingness, representativeness and uniformity.

Skims summaries

Creating summaries by jointly optimising multiple objectives

- The objectives of a good summary are interestingness, representativeness and uniformity.
- The objectives are modelled using sub-modular functions which have diminishing returns property.
- Video sumarisation modelled as a subset selection problem.

Creating summaries by jointly optimising multiple objectives

Given a video V and a budget B, let Y_V denote all possible solutions of $y \subseteq V$ given the constraint $|y| \leq B$. We need to find a subset $y^* \in Y_V$ that maximises the objective function o.

$$o(x_V, y) := w^T f(x_V, y) \tag{8}$$

where the entries of w are positive.

Creating summaries by jointly optimising multiple objectives

Modelling video summarisation

So the task of video summarisation is to select a summary y^* such that

$$y^* = \arg\max_y o(x_V, y) \tag{9}$$

where $y \in Y_V$ and x_V are all the features extracted from the video. $o(x_V, y)$ is defined as a linear combination of submodular objectives:

$$f(x_V, y) = [f^{int}(x_V, y), f^{rep}(x_V, y), f^{uni}(x_V, y)]^T \tag{10}$$

where f^{int}, f^{rep} and f^{uni} represents interestingness, representativeness and uniformity respectively.

Creating summaries by jointly optimising multiple objectives

Formulation of interstingness

- Interestingness objective f^{int} is given by:

$$f^{int}(x_V, y) = \sum_{k \in \cup s, s \in y} I(k) \tag{11}$$

- In the case of non-overlapping segments the formula becomes:

$$f^{int}(x_V, y) = \sum_{s \in y} I(x_s) \tag{12}$$

Creating summaries by jointly optimising multiple objectives

Formulation of representativeness

Representativeness is to find best k segments to represent a video is known as the the k-medoids problem. The k-medoid objective can be reformulated as a submodular objective as follows:

$$f^{rep}(x_V, y) = L_r(x^r, \{p'\}) \text{ - } L_r(x^r, y \cup \{p'\}) \quad (13)$$

$$L_r(x^r, y) = \sum_{i \in V} \min_{s \in y} \|x_i^r - x_s^r\|_2^2 \quad (14)$$

where x_i^r is the deep feature for i^{th} frame, x_s are the deep features used to represent a segment and p' is a phantom examplar.

Creating summaries by jointly optimising multiple objectives

Formulation of uniformity

Uniformity ensures that there are no abrupt jumps in the summary at the same time maintaining the temporal coherence.

$$f^{uni}(x_V, y) = L_r(x^u, \{p'\}) \text{ - } L_r(x^u, y \cup \{p'\}) \tag{15}$$

$$L_r(x^u, y) = \sum_{i \in V} \min_{s \in y} \|x_i^u - x_s^u\|_2^2 \tag{16}$$

where x^u are frame numbers and x_s is mean frame number used to represent a segment and p' is a phantom examplar

Creating summaries by jointly optimising multiple objectives

Learning weights

Given T pairs of videos and their summaries (V, y_{gt}), the weights in the vector w needs to be learnt. So the following large-margin formulation must be optimised:

$$\min_{w \geq 0} \frac{\sum_{t=1}^{T} L_t(w) + \frac{\lambda \|x\|^2}{2}}{T} \tag{17}$$

where $L_t(w)$ is the generalized hinge loss of training example t given by

$$L_t(w) = \max_{y \subseteq Y_V^{(t)}} (w^T f(x_V^{(t)}, y) + l_t(y)) - w^T f(x_V^{(t)}, y_{gt}^{(t)}) \tag{18}$$

Here superscript (t) is used refer to both features and subsets of video t.

$$l_t(y) = \frac{1}{B}\dot{(}\|y\| - \|y \cap y^{(t)}\|), \tag{19}$$

$l_t(y)$ is a count of how many of the candidate summary y are not represented in the ground truth, normalized by the maximal length of the summary.

Creating summaries by jointly optimising multiple objectives

Generating summaries

- Given pairs of videos and their user created summaries as training examples, python implementation of "gm_ submodular" package was used for learning the weights for submodular objectives.
- Once the weights were found, optimisation was done using MATLAB® toolbox *Submodular Function Optimisation* which is an implementation of lazy-greedy algorithm for submodular function optimisation.
- When unknown video is input, this method creates summaries that are interesting, representative and uniform.
- Reading and writing videos were done in MATLAB®.

Weights for different objectives

Objective	Weight
Interestingness	0.98619
Representativeness	0.00002
Uniformity	0.01379

Table : Weights obtained for different objectives

Creating summaries by jointly optimising multiple objectives

Running time

Video name	Video time (sec)	Summarisation time (Hr)
nasaani.mpg	30	0.69
America_ New_ Frontier_ Seg4.mpg	123	1.74
America_ New_ Frontier_ Seg10.mpg	161	2.34
games.avi	135	6.06
drone.mp4	253	10.55

Table : Summarisation time for videos using optimisation of submodular mixtures.

Skims summary

Disscussion

- Submodular optimisation is twice slower than the knapsack optimization.
- Superframe segmentation is highly resource intensive (requires a lot of RAM) - solution: resize the frames before passing frames to superframe segmentation.
- Weights learnt from the ground truth show that 98% importance is given to interestingness and importance to representativeness and uniformity is very meagre.

Skims summary

Disscussion contd...

Then why do we have to formulate summarisation in the way we did???

- This formulation leads us to the conclusion that humans prefer *interesting* summaries.
- Take look at the summary for "games.avi" video generated by both the methods.
- Can we try out other objectives???

Skims summary

f-measure and recall values for different objectives

Method of generating summaries	f-measure (%)	recall (%)
Random	18.95 ± 0.06	43.72 ± 0.14
Interestingness	20.3 ± 0.06	59.62 ± 0.19
Uniformity	17.96 ± 0.08	38.04 ± 0.22
Representativeness	19.04 ± 0.04	46.58 ± 0.11
Combination of 3 objectives	22.31 ± 0.08	58.43 ± 0.21

Table : f-measure and recall values for different methods of summarisation.

Video Summarisation

Conclusion

- It is very hard to exactly define what a good summary is for all.
- An ideal summarisation algorithm must be very adaptive to the preference of each individual.
- Length of summary also plays a very important role.
  - Long summary focuses on representing various content.
  - Short summary focuses more on interestingness.
- The idea of interestingness varies from domain to domain and from person to person.
- *A summary is a "good summary" only in the eyes of a particular user.*

Future work

- Personalised video summarisation.
- Ego-centric video summarisation.
- Using audio clues for improving the video summaries.

Bibliography

1 Tommy Chheng. Video summarization using clustering. Department of Computer Science, University of California, Irvine, pages 1–7.

2 Yelana Yesha Padmavathi Mundur, Yong Rao. Keyframe-based video summariza- tion using delaunay clustering. International Journal on Digital Libraries, 6(2): 219–232, April 2006.

3 Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. Cre- ating summaries from user videos. ECCV, 2014.

4 Luc Van Gool Michael Gygli, Helmut Grabner. Video summarization by learning submodular mixtures of objectives. Computer Vision and Pattern recognition, pages 3090–3098, 2015.

5 Harry Agius Arthur G. Money. Video summarisation: A conceptual framework and survey of the state of the art. Elseivier, pages 121–144, April 2007. URL www.sciencedirect.com.

6 Geoffrey E. Hinton Alex Krizhevsky, Ilya Sutskever. Imagenet classification with deep convolutional neural networks. NIPS.

7 Mehmood I. Wook Baik S Ejaz, N. Efficient visual attention based framework for extracting key frames from videos. Signal Processing: Image Communication, 2013.

8 Jia Li James Z. Wang Ritendra Datta, Dhiraj Joshi. Studying aesthetics in photo- graphic images using a computational approach. ECCV, 2006.

9 Feng Jing Yan Ke, Xiaoou Tang. The design of high-level features for photo quality assessment. Computer Vision and Pattern recognition, 2006.

10 Jones M Viola P. Robust real-time face detection. IJCV, 2004.

Bibliography contd...

11 H. Lin and J. Bilmes. Learning mixtures of submodular shells with application to document summarization. In Uncertainty in Artificial Intelligence (UAI), 2012.

12 Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093, 2014.

13 Oriol Vinyals Judy Hoffman Ning Zhang Eric Tzeng Trevor Darrell Jeff Donahue, Yangqing Jia. Decaf: A deep convolutional activation feature for generic visual recognition. arXiv, pages 1–10, october 2013.

14 Andreas Krause. Sfo: A toolbox for submodular function optimization. Journal of Machine Learning Research, 2010.

15 Daniel Golovin Andreas Krause. A survey on submodular function maximization. 2012.

Questions???