Journal of Visual Communication and Image Representation 52 (2018) 118–130

Contents lists available at ScienceDirect

Journal of Visual Communication and Image Representation

journal homepage: www.elsevier.com/locate/jvci

Edited nearest neighbour for selecting keyframe summaries of egocentric videos

Ludmila I. Kuncheva[a,*], Paria Yousefi[a], Jurandy Almeida[b]

[a] School of Computer Science, Bangor University, Dean Street, Bangor, Gwynedd, Wales LL57 1UT, UK
[b] Institute of Science and Technology, Federal University of São Paulo – UNIFESP, São José dos Campos, São Paulo 12247-014, Brazil

ARTICLE INFO

Keywords:
Keyframe summary
Nearest neighbour classifier
Instance selection
Egocentric video
Feature representations

ABSTRACT

A keyframe summary of a video must be concise, comprehensive and diverse. Current video summarisation methods may not be able to enforce diversity of the summary if the events have highly similar visual content, as is the case of egocentric videos. We cast the problem of selecting a keyframe summary as a problem of prototype (instance) selection for the nearest neighbour classifier (1-nn). Assuming that the video is already segmented into events of interest (classes), and represented as a dataset in some feature space, we propose a Greedy Tabu Selector algorithm (GTS) which picks one frame to represent each class. An experiment with the UT (Egocentric) video database and seven feature representations illustrates the proposed keyframe summarisation method. GTS leads to improved match to the user ground truth compared to the closest-to-centroid baseline summarisation method. Best results were obtained with feature spaces obtained from a convolutional neural network (CNN).

1. Introduction

Keyframe selection is now an established way of summarising video data [7,38,50]. The result is a compact and diverse collection of frames which covers the content of the video. The large and still growing number of methods and approaches to keyframe selection can be explained with the variety of applications, video types, purposes and criteria for building a video summary [38]. This variety also makes it difficult to create a comprehensive taxonomy of these approaches [44]. Summaries of videos and photo streams, both in their static version (keyframes) or dynamic version (video skims) can serve at least the following purposes [5,6,38,50]:

- Easy browsing, navigating and retrieval of a video from a repository [1,20,14] or on the Web [2,18,56].
- Concise representation of the storyline of a TV episode [3], sports, news, rushes, documentaries, etc.
- Summarising daily activities captured by an egocentric or life-logging camera [6,10,37], including identifying frames which look like intentionally taken photos [59].
- Memory reinforcement [6,15,25,31].
- Motion capture and retrieval used in many areas such as gaming, entertainment, biomedical and security applications [28].
- Recording cultural experience [52].
- Summarising and annotating surveillance videos [9].

Depending on the type, the length of a video may range from less than a minute to several hours, and the shot lengths can vary dramatically within. This suggests that one-fits-all methods for keyframe selection may not be as successful as tailor-made ones. Nonetheless, there is consensus among the researchers that a keyframe-based video summary should be 'concise', 'informative', should 'cover' the content of the video, and should be 'void of redundancies'. While the interpretation of these categories is domain-specific, they are valid across different video types and applications.

Driven by these desiderata, here we cast the keyframe selection problem as prototype selection (instance selection) for the nearest neighbour classifier. We assume that the video has been segmented into units such as shots, scenes, or events. Any segmentation method can be used for this task. Our approach can be formulated as follows: Select the smallest number of keyframes which allows for the best discrimination between the units. In this paper we assume that the frames can be represented as points in an n-dimensional space $\mathbb{R}^n$. The quality of the discrimination between units is defined as the estimated generalisation accuracy of the nearest neighbour classifier (1-NN) using the selected frames as the reference set, where each unit is treated as a class. This approach will automatically address some of the desirable properties of a video summary:

(a) The approach ensures that the units of interest are all distinguishable from one another, which implies *diversity and coverage* of the

* Corresponding author.
E-mail addresses: l.i.kuncheva@bangor.ac.uk (L.I. Kuncheva), paria.yousefi@bangor.ac.uk (P. Yousefi), jurandy.almeida@unifesp.br (J. Almeida).

https://doi.org/10.1016/j.jvcir.2018.02.010
Received 16 March 2017; Received in revised form 15 January 2018; Accepted 13 February 2018
Available online 15 February 2018

representing keyframes. This is different from the current approaches in that in our approach the importance of the individual frames is determined implicitly, in relation to all the frames in the collection.

(b) *Anomalies*, which are not mere artefacts, will be captured as they will be strong candidates for discriminating between different events.

While the proposed approach does not explicitly maximise the aesthetic quality [59] or memorability [26] of each image, it is designed to tell the story *as a whole*.

The rest of the paper is organised as follows. Section 2 reviews related work. A taxonomy of the edited nearest neighbour methods is presented in Section 3. Our Greedy Tabu Selection method (GTS) is explained in Section 4. An experiment with four egocentric videos from the UTE data base [32] is reported in Section 5. Section 6 offers our conclusions and some future research directions.

2. Related work

Let $\mathbf{V} = \langle f_1,...,f_N \rangle$ be the video to be summarised, and f_i be the frames arranged according to time. The task is to select a collection of keyframes, usually ordered by time tag, such that

$$\mathbf{f}^* = \langle f^*_{j_1},...,f^*_{j_K} \rangle = \arg\max_{j_1 j_2 ... j_K} J(\mathbf{f}), \tag{1}$$

where $J(\mathbf{f})$ is a criterion function evaluating the merit of keyframe selection $\mathbf{f}$. Sometimes the number of frames K is also a part of the criterion, and is derived through the optimisation procedure.

The criterion function J is rarely defined in mathematical terms; it is more often a domain-specific interpretation of the desirable properties such as coverage, conciseness, informativeness, diversity, etc.

2.1. Keyframe selection from events/segments

Keyframe selection has been approached from at least two perspectives. In the first perspective, the video is split into *units*, typically ordered (from smallest to largest) as:

$$\text{frames} \rightarrow \underbrace{\text{shots} \rightarrow \text{scenes/events} \rightarrow \text{clips}}_{\text{units}} \rightarrow \text{video}$$

In the standard video structure, shots are regarded as the primitive unit of meaning [50]. Truong and Venkatesh report back in 2007 that the task of independent segmentation of a video into *shots* has been declared a "solved problem" by NIST TRECVID benchmark. However, the task of segmenting an unedited video, especially an egocentric video, into contextually meaningful parts is much more difficult and far from over, as witnessed by a host of a later-date publications: [4,23,37,44,47].

After segmentation, each unit (event) gives rise to one or more keyframes. The keyframes are pooled, and the final collection is often analysed in order to prune irrelevant or redundant keyframes. Similarity to frames already selected within the event, and dissimilarity to keyframes in other events have been among the most popular pair of criteria [11,36,50,55]. Other criteria include visual and temporal attention [16,43], utility [54], and quality [27] of the individual frame. Such criteria usually include a similarity term which enforces diversity or temporal distance with keyframes selected already.

2.2. Keyframe selection from the entire video

By selecting keyframes from shots or other units *independently*, we lose sight of the whole video. Diversity between the selected keyframes is often compromised on the larger scale, requiring post-processing to eliminate irrelevant and redundant keyframes. One way to combat this problem is to take the video as a whole. The shot-based methods optimising a "quality" function with a penalty for high similarity between the selected keyframes, can be applied straightforwardly [17,22,34,43,54]. Possible solutions to the optimisation problem represented by Eq. (1) are sought through greedy procedures [22,35], dynamic programming [33,54], or 0/1 knapsack optimisation [23].

Consider representation of the frames in some n-dimensional feature space $\mathbb{R}^n$. The frames are grouped into one or more clusters, and representative keyframes are elected from each cluster [45,42,46,61]. Most clustering procedures are iterative (and agglomerative), whereby the clusters are grown from single frames, and new clusters are seeded when a frame happens to be too far from the current clusters. Usually the representative keyframe for a cluster is chosen to be the one closest to the cluster centroid in the feature space. Selecting non-central keyframes to capture cluster variability has also been explored [18]. Note that clustering can be applied to a single event/segment as well to the whole video. When applied over the whole video, temporal relationship between the clusters is not enforced, and some events may lose their identity. This can happen when events distant in time have similar representations, and will warrant a single representative frame. Such an approach will not be useful if the goal of the summary as memory aid.

Nonetheless, clustering approaches over the whole video have proven successful [21,40,51,60]. Keyframes are selected from the clusters and often post-processed. Such a 'monolithic' approach gives better control over handling the balance between diversity and representativeness.

We propose to look at the keyframe selection task from a different angle. Assume that the events are classes, and the task is to select keyframes which best discriminate between them. The classes don't have to be a particular activity, scenario or place. The term "class" here represents the video content in the event's time span. The solution will automatically (and implicitly) maximise both representativeness and diversity. Using a representation of the data in $\mathbb{R}^n$, and labels corresponding to the events, we can solve the problem by choosing from the rich variety of prototype/instance selection methods [19,58].

2.3. Discrimination-based extraction of keyframes

In our case, the labels are defined by the segmentation. The idea closest to the one we propose is to include a discriminative component in the quality measure. Cooper and Foote [11] propose three variants of a quality measure for a frame f. One of these is derived from the linear discriminant analysis (LDA).

Suppose that the video has been segmented into units $U_1,...,U_K$, where the frames are indexed as follows:

$$U_i = \langle f_{i,1}, f_{i,2}, ... f_{i,k_i} \rangle.$$

A feature extraction function is used to transform all the frames into feature vectors. Then the quality measure is the negative Mahalanobis distance from the frame data point to its class mean

$$Q(f) = -(F(f)-\mu_i)^T W^{-1} (F(f)-\mu_i), \quad f \in U_i,$$

where

$$\mu_i = \frac{1}{k_i} \sum_{j=1}^{k_i} F(f_{i,j})$$

is the mean of unit U_i, and W is the pooled covariance matrix

$$W = \frac{1}{N-1} \sum_{i=1}^{K} \sum_{j=1}^{k_i} (F(f_{i,j})-\mu_i)(F(f_{i,j})-\mu_i)^T,$$

where N is the number of frames in the video. The frame with the highest quality for U_i will be the one closest to the mean. We can simplify the measure and use Euclidean distance in Q. The result is the widely-used baseline methods for keyframe selection where all frames