Edited nearest neighbour for selecting keyframe summaries of egocentric videos

Kuncheva, Ludmila; Yousefi, Paria; Almeida, Jurandy

Journal of Visual Communication and Image Representation

DOI:
10.1016/j.jvcir.2018.02.010

Published: 01/04/2018

Peer reviewed version

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

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).

Keywords: Keyframe summary, nearest neighbour classifier, instance selection, egocentric video, feature representations.

1. Introduction

Keyframe selection is now an established way of summarising video data [6, 37, 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 [37]. This variety also makes it difficult to create a comprehensive taxonomy of these approaches [43]. Summaries of videos and photo streams, both in their static version (keyframes) or dynamic version (video skims) can serve at least the following purposes [4, 5, 37, 50]:

- Easy browsing, navigating and retrieval of a video from a repository [1, 19, 13] or on the Web [2, 17, 56].
- Concise representation of the storyline of a TV episode [48], sports, news, rushes, documentaries, etc.
- Summarising daily activities captured by an egocentric or lifelogging camera [5, 9, 36], including identifying frames which look like intentionally taken photos [59].
- Memory reinforcement [5, 14, 24, 30].
- Motion capture and retrieval used in many areas such as gaming, entertainment, biomedical and security applications [27].
- Recording cultural experience [52].
- Summarising and annotating surveillance videos [8].

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 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

*Corresponding Author
Email addresses: l.i.kuncheva@bangor.ac.uk (Ludmila I Kuncheva), paria.yousefi@bangor.ac.uk (Paria Yousefi), jurandy.almeida@unifesp.br (Jurandy Almeida)

different events.

While the proposed approach does not explicitly maximise the aesthetic quality [59] or memorability [25] 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 [31] is reported in Section 5. Section 6 offers our conclusions and some future research directions.

2. Related work

Let $\mathbf{V} = \langle f_1, \ldots, f_N \rangle$ 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

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

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.

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: [3, 22, 36, 43, 46].

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 [10, 35, 50, 55]. Other criteria include visual and temporal attention [15, 42], utility [54], and quality [26] of the individual frame. Such criteria usually include a similarity term which enforces diversity or temporal distance with keyframes selected already.

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 [16, 21, 33, 42, 54]. Possible solutions to the optimisation problem represented by Eq. (1) are sought through greedy procedures [21, 34], dynamic programming [32, 54], or 0/1 knapsack optimisation [22].

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 [44, 41, 45, 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 [17]. 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 [20, 39, 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 [18, 58].

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 [10] 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 $U_1, \ldots, U_K$, where the frames are indexed as follows:

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

A feature extraction function $F(f)$ 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 in the unit are regarded as one cluster, and the frame closest to the centroid is taken as the keyframe for this unit.

A remark: credit apportionment problem. The difficulty in evaluating keyframe summaries has often been noted [25, 38, 50, 37]. Added to this, there is a marked need for credit apportionment analysis. Many studies propose a whole pipeline, from extracting specific features, through tailor-made segmentation, cleaning of redundant/irrelevant/low-quality frames, and leading to the keyframe selection method proposed within the study. While the overall quality of the summary is typically judged by user studies, it is not clear which element of the proposed methodology is responsible for the results. It stands to reason that the components of a video summarisation pipeline should be evaluated separately. Hence, we focus on a keyframe selection method which can be coupled with any segmentation approach and feature space. Our approach requires only that each frame is represented as a point in some n-dimensional space, regardless of what the dimensions mean and how the feature values are calculated.

3. An edited nearest neighbour approach to keyframe selection

Data editing has been a long-standing theme in pattern recognition. Following the two classical methods: Condensed Nearest Neighbour (CNN) [23] and Edited Nearest Neighbour (ENN) [57], a large number of data editing approaches and methods have been proposed and periodically summarised [7, 12, 18, 49, 58].

3.1. Motivation

The following example illustrates the rationale behind our proposal. Suppose that you have recorded your day in a set of 4 events as shown in Figure 1: (1) Met Mary, (2) Looked at the door, (3) Met Mary again, (4) Looked at the door again. Each row with frames corresponds to one event (from left to right).

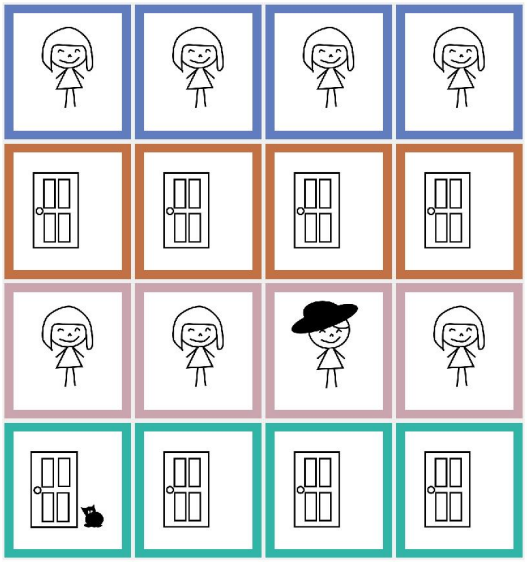
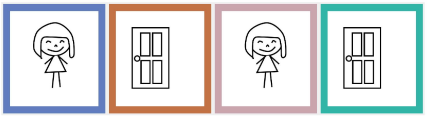
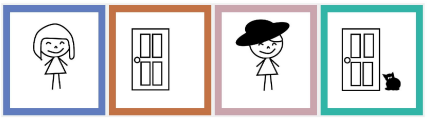


Figure 1: Example: A day with 4 events.

The standard approach which selects the frame closest to the cluster centroid will pick a frame with Mary (without the hat) for both events 1 and 3, and a frame with the door (without the cat) for both events 2 and 4, as shown in Figure 2 – Summary 1. If, however, you need to tell the story about your day to a friend, you will likely pick the frames with the hat and the cat to distinguish events 1 from 3 and 2 from 4 (Summary 2 in Figure 2).



Summary 1 (traditional): Closest to class centroid.



Summary 2 (proposed): Edited nearest neighbour.

Figure 2: Two keyframe summaries of the 4 events in the example in Figure 1.

Admittedly, a diversity-wise selection method may also be expected to recover the different frames for events 3 and 4. However, we re-position this task as an edited nearest neighbour problem, which will not require manual setting of the balance between diversity and representativeness.

3.2. Problem formulation

Let **V** be the video of a temporally ordered collection of N frames that is to be summarised. We shall assume that the

segmentation of the video into units has been done so that the frames are labelled into K segments, $U_1, \ldots, U_K$, which we will treat as classes. We assume that n features have been extracted from each frame so that the video is represented as a data set of size $N \times n$, with another vector containing the N class labels. Then the problem is to select a subset of frames $S \subset \mathbf{V}$ such that the nearest neighbour classifier (1-NN) has as high as possible resubstitution accuracy using S as the reference set and U_i as the class labels. Our hypothesis is that such a keyframe selection will work well for at least the following reasons:

- This approach ensures that S will contain frames which describe their own classes as accurately as possible (coverage/representativeness/relevance) while accounting for the differences between the classes (diversity).
- The frames are chosen collectively, in relation to one another, which counteracts redundancy, and contributes towards "story telling".

In pattern recognition and machine learning, this task is known under different names: instance selection, prototype selection (extraction, generation, replacement), and editing for the nearest neighbour classifier, among others.

4. Greedy Tabu Selector (One-per-Class): 1-nn editing for keyframe selection

4.1. The algorithm details

The proposed algorithm is detailed as Algorithm 1. As the algorithm is applicable to any data type (not only video data), we use the universal pattern-recognition/machine-learning terminology:

• instance (=prototype) = frame,
• class = unit/event, obtained through segmentation of the video (hence no additional annotation is needed), and
• selected subset of prototypes = keyframe summary.

The input is a data set X labelled in c classes. The algorithm starts by identifying the instance closest to the class centroid for each class. These c instances are taken together to be the first candidate reference set of prototypes S. (This corresponds to a keyframe summary often used as a baseline in comparative studies.) The set is subsequently modified in the following way. The nearest neighbour classifier (1-nn) is applied on X using S as the reference set. All classes are declared 'available' at the beginning. A 'privileged' class is chosen among the available classes as the one with the worst proportion of correctly labelled instances. It is subsequently made unavailable for the next t iterations, where t is the 'tabu' parameter, $0 < t < c$. The prototype for the privileged class, say x^j, is marked for replacement. All remaining instances from class j are taken in turn to replace x^j in S, and the resubstitution error of 1-nn is calculated for each new version of S. Suppose that the reference set with the smallest error was S', when x^j in S was replaced by x^{j*}. The 1-nn error with S' as the reference set is compared with the error with S. If the new error is smaller, the replacement is made permanent by setting $S \leftarrow S'$. Otherwise, no change is made to S , and the algorithm continues by selecting a new privileged class from the available classes.

Algorithm 1: Greedy Tabu Selector (One-per-Class)

Input: Data set $X = \{x_1, \ldots, x_N\} \subset \mathbb{R}^n$ and the correspodnig labels into classes $\{1, 2, ..., c\}$. Tabu parameter, an integer t, $0 < t < c$.

Output: Selected set of prototypes $S \subset X$ with cardinality $|S| = c$, containing one instance from each class.

1 **for** $i \leftarrow 1, \ldots, c$ **do**
2 Find the centroid of class i and identify the instance x^i from this class closest to the centroid.
3 Construct the initial set of prototypes: $S \leftarrow \{x^1, \ldots, x^c\}$.
4 Set all classes as 'available'.
5 Initialise the minimum-error holder: $E_{\min} \leftarrow 1$.
6 Initialise the 'no-change' counter: $w \leftarrow 0$.
7 **while** $w < c$ **do**
8 Among the 'available' classes, find the class with the highest proportion of misclassified instances, say class j.
9 Replace *temporarily* the current instance $x^j \in S$ with each of the remaining instances from class j, one at a time. Identify the instance x^{j*} which gives the minimum resubstitution error E.
10 Mark class j as 'not-available' for another t iterations.
11 **if** $E < E_{\min}$ **then**
12 $E_{\min} \leftarrow E$.
13 Replace x^j permanently: $S \leftarrow S \setminus \{x^j\} \cup \{x^{j*}\}$.
14 $w \leftarrow 0$.
15 **else**
16 $w \leftarrow w + 1$
17 Return S.

The stopping condition of the algorithm is implemented as follows. A counter w of steps without changes is initially set to 0. This counter is incremented any time a privileged class is checked but no change to S is made (the 'else' statement in lines 15 and 16 in Algorithm 1). The counter is reset to 0 every time a change in S occurs. If there have been c steps without a change, the greedy approach cannot improve any further on the 1-nn resubstitution error, the search is terminated, and S is returned.

Note that, after the first t iterations, the choice will be only among the available $c - t$ classes. Therefore, if we set $t = c - 1$, the classes will be ordered during the first pass through all of them, and checked in this order thereafter.

4.2. Greedy Tabu Selector for the cartoon example

Consider applying the Greedy Tabu Selector to the example in Figure 1. To quantify the frame data, we introduce 4 binary

features: (1) Mary present, (2) hat present, (3) door present, and (4) cat present. The labelled data is shown in Table 1.

Table 1: Cartoon example data

Frame	Features (1)	(2)	(3)	(4)	Labels
1.	1	0	0	0	1
2.	1	0	0	0	1
3.	1	0	0	0	1
4.	1	0	0	0	1
5.	0	0	1	0	2
6.	0	0	1	0	2
7.	0	0	1	0	2
8.	0	0	1	0	2
9.	1	0	0	0	3
10.	1	0	0	0	3
11.	1	1	0	0	3
12.	1	0	0	0	3
13.	0	0	1	1	4
14.	0	0	1	0	4
15.	0	0	1	0	4
16.	0	0	1	0	4

Set $t = c - 1 = 3$. At the initialisation step the Greedy Tabu Selector will pick frames $S = \{1, 5, 9, 14\}$, leading to 50% re-substitution error. The first privileged class will be class 3. After replacing frame 9 with frame 11, the error drops to 43.75%. Class 3 is banned from checking again in the next 3 steps. The next privileged class is 4, and frame 14 is replaced with frame 13, leading to error rate 37.50%. Class 1 and class 2, which are still available are checked next, and no change to S is made. At this step, class 3 becomes available again, and the check reveals that no improvement of the error is achieved. Class 4 becomes available next, and again, no improvement is possible. As there have been 4 steps ($w = 4$) with no change to S, the best version is returned: $S = \{1, 5, 11, 13\}$, which corresponds to the desired summary shown in Figure 2 (Summary 2).

4.3. An example with generated data

Figure 3 shows the scatterplot of a 2D data set labelled in three classes, shown with different markers and colours. The Greedy Tabu Selector was applied to the dataset. The migration of the prototypes in the original set (instances closest to the class centroids) is marked by lines. The final prototypes returned by the algorithm are circled. The error rate at the start is 22.28%, and the one at the end, with the selected set of three prototypes, is 17.89%, which demonstrates that substantial improvement on the error can be achieved with a minimal-size set of prototypes obtained through a simple greedy approach.[1]

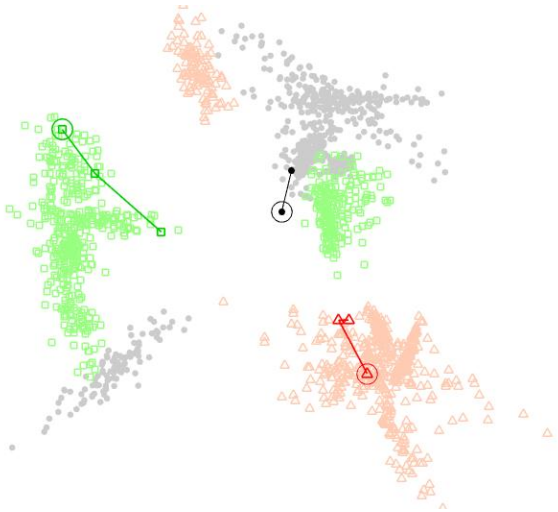


Figure 3: An example of 2D data labelled in three classes, shown here with different markers and colours. The migration of the prototypes in the original set is marked by lines. The final set of prototypes selected through the Greedy Tabu Selector algorithm are circled.

5. Experimental evaluation

5.1. The challenge of egocentric video data

We chose to examine our method on egocentric videos because, as we demonstrate in this section, they offer an extra degree of challenge for the task of keyframe selection [37]. To this end, we selected three videos of different categories: a video professionally prepared as educational material, a third-person casual video, and an egocentric video. For the purposes of this illustration, we took four units (shots/segments/events) from each one.[2] The videos were as follows: Educational material[3], video number 21: "The Great Web of Water-segment 01", a third-person casual video[4], "Jumps", and sub-sampled egocentric video[5], video P01. Figures 4 – 6 show the results of applying the Closest-to-Centroid and the GTS method to the three videos.

The top plots (subplots (a)) in the three figures show a montage of 10 frames uniformly spaced within each event. Each row corresponds to an event. In addition, the events are colour-coded by the frame borders. The colours are also carried forward in the scatterplots (c) and (d).

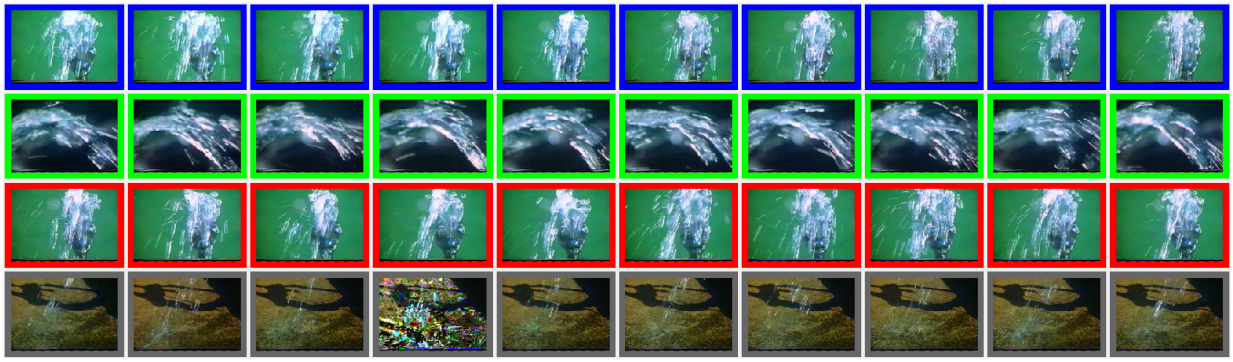
[1] MATLAB code for the GTS and the CC algorithms, as well as the data and code and this example are stored in GitHub `https://github.com/LucyKuncheva/1-nn-editing`.

[2] We shall term the units of interest 'events'.

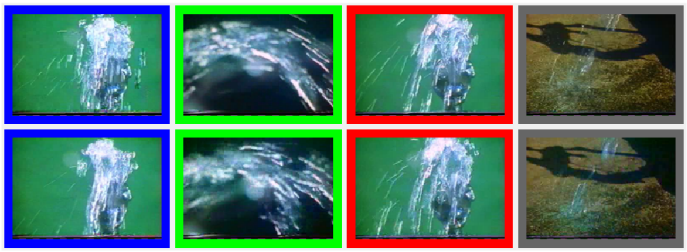
[3] VSUMM[13]:`https://sites.google.com/site/vsummsite/download`

[4] SUMME[22]:`https://people.ee.ethz.ch/~gyglim/vsum/`

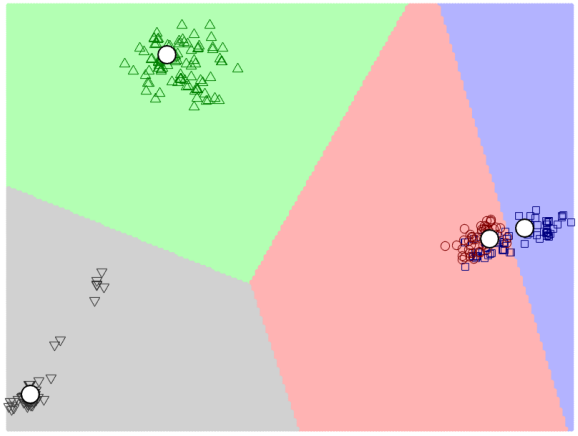
[5] UTE[31]:`http://vision.cs.utexas.edu/projects/egocentric/`



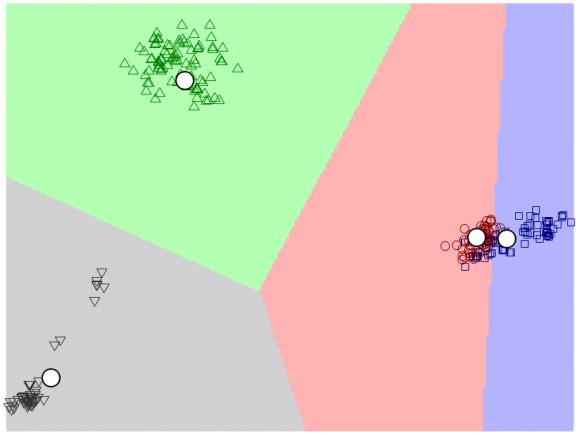
(a) Montage of uniformly spaced frames from the four events (shots in this case).



(b) Summaries of the four events. Top row: closest-to-centroid; bottom row GTS summary.



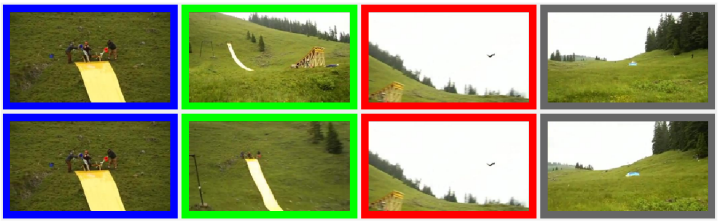
(c) Classification regions for the close-to-centroid method
1-nn error rate 7.4%



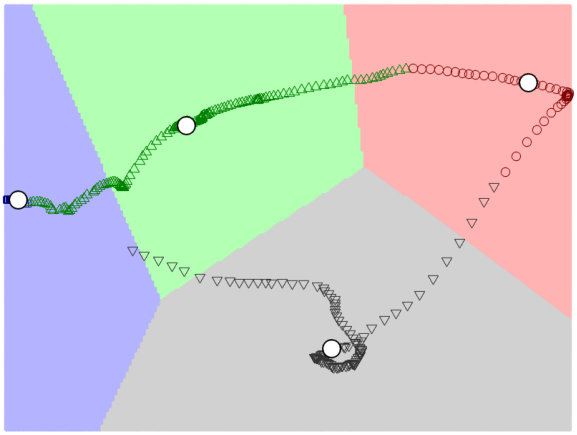
(d) Classification regions for the GTS method
1-nn error rate 4.1%

Figure 4: Educational video: Keyframe selection through Closest-to-Centroid (CC) and Greedy Tabu Search (GTS) for a part of video #21 from the VSUMM collection, RGB space.

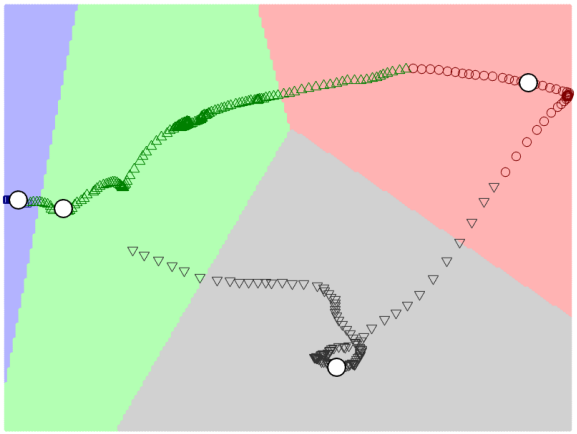
(a) Montage of uniformly spaced frames from the four events (segments in this case).



(b) Summaries of the four events. Top row: close-to-centroid; bottom row GTS summary.

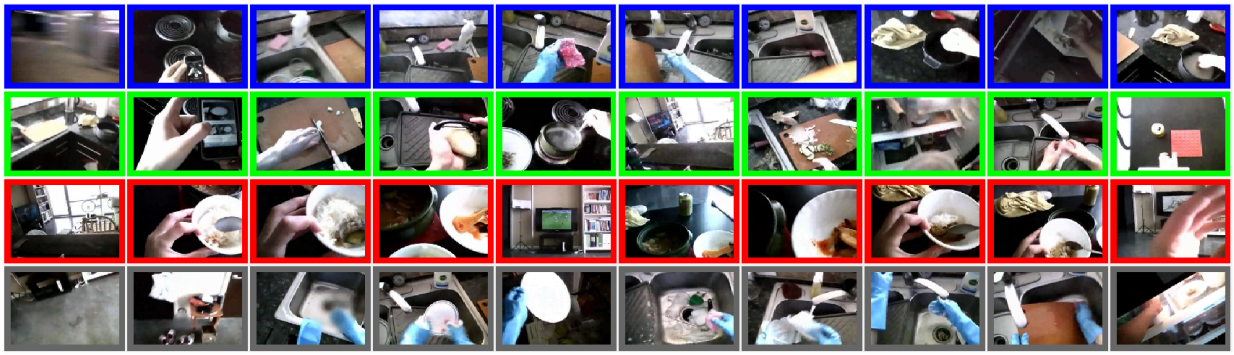


(c) Classification regions for the close-to-centroid method
1-nn error rate 9.3%



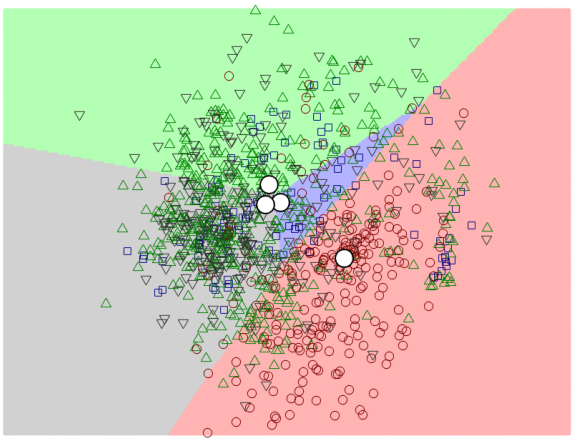
(d) Classification regions for the GTS method
1-nn error rate 5.5%

Figure 5: Third Person Video: Keyframe selection through Closest-to-Centroid (CC) and Greedy Tabu Search (GTS) for a part of video "Jumps" from the SUMME collection, RGB space.

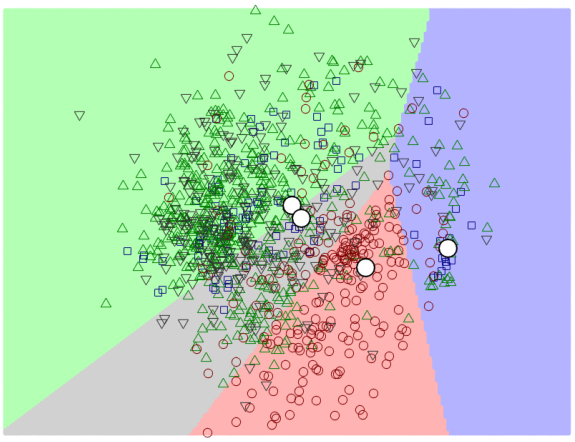


(a) Montage of uniformly spaced frames from the four events (events in this case).

(b) Summaries of the four events. Top row: close-to-centroid; bottom row GTS summary.



(c) Classification regions for the close-to-centroid method
1-nn error rate 55.2%



(d) Classification regions for the GTS method
1-nn error rate 40.1%

Figure 6: Egocentric Video: Keyframe selection through Closest-to-Centroid (CC) and Greedy Tabu Search (GTS) for a part of video P01 from the UTE collection, RGB space.

Subplot (b) in all three figures contains two 4-frame summaries. One frame has been selected from each event. The top row is the result of the Closest-to-Centroid method, and the bottom row is the result of the proposed GTS method. Note that, for the purposes of this illustration, in both methods we used the simple RGB feature space RGB^{cm} described in detail later in Section 5.2.

Finally, subplots (c) and (d) give the classification regions for the 4 events (treated as classes) for the two summaries. The scattered points correspond to frames of the video. Different events (classes) are denoted by different marker shapes and colours. The four selected frames are marked with large open-circle markers in each plot. The classification regions are shaded with the colour of the event. They are calculated *only* in the 2d projection space obtained as the first two principal components of the RGB^{cm} space. Shown in the subplot caption are the error rates obtained with the nearest neighbour classifier using the selected 4 frames as the reference set.

The figures demonstrate the dramatic differences between the types of videos. Non-egocentric videos are likely to have a much simpler structure in that the units of interest are represented by visually similar frames, as can be seen in Figures 4 and 5. The events are clearly distinguishable in all subplots. This is especially visible in the scatterplots (c) and (d). Conversely, these subplots in Figure 6 reveal that the classes are highly overlapping. This fact is also supported by a visual inspection of the frame montage for the four events. We can broadly label the events in this figure as: (1) Preparing the kitchen, (2) Cooking, (3) Eating, (4) Washing up. Because of the overlap, the Closest-to-Centroid summary picks similar frames as shown in the top row of subplot (b) in Figure 6. Our GTS method manages to 'disentangle' the events to some extent, as demonstrated by the differences between the keyframes in the bottom row of the same subplot.

Compare now the differences between (c) and (d) in the three figures. The regions for the egocentric video change the most, suggesting that GTS has a much stronger effect for this type of video. Another indication of the suitability of GTS for egocentric video is the reduction of error rate. The error rates for the educational video and the third-person video were not very large to begin with. This means that many similar frames can be chosen as the summary, and the summary will still be good. For these two types of video, GTS makes a small improvement on the error rate, but the two rival summaries CC and GTS are not really distinguishable. This is not the case for the egocentric video. The two summaries are indeed different, and the proposed method leads to a more diverse and meaningful summary.

Hence, while many keyframe selection methods may give equivalent results for the first two video types, egocentric videos are significantly more complicated. This explains the abundance of criteria, approaches and methods for summarisation of this video type. As a byproduct of GTS, we have a standalone measure of the merit of a keyframe summary: the classification accuracy achieved by using this summary as the reference set for the nearest neighbour classifier. Lower error will mean that the keyframes are representative of the events they are meant to summarise, and diverse enough to allow for these events to be distinguishable.

5.2. Feature representations

We examined 7 feature spaces explained in Table 2.

Table 2: Feature spaces

Level	Information	Notation	Size
Low	colour	RGB	54
	colour	HSV	144
	texture	LBP	59
	shape	HOG	864
High	people and objects	CNN	4096
	people and objects	CNN90	84,89, 86, 74
	semantic	SEM	1001

Note: The number of retained principal components was different for the four videos, as listed in the last column for CNN90.

From the colour type, we chose RGB and HSV. The RGB are colour moment features extracted as follows: each frame was divided uniformly into a 3-by-3 grid of blocks and then we computed the mean and the standard deviation for each block and each colour (9 blocks $\times$ 3 colours $\times$ 2 statistics = 54 features). For the HSV space, the frame was split again into 9 block, and a 16-bin colour histogram was computed from the hue value (H) of the HSV colour space.

From the texture type, we used the local binary patterns (LBP) [40], and for the shape type, the histogram oriented gradients (HOG) [11].[6]

The high-level feature spaces were calculated using a Convolutional Neural Networks (CNN) from the MATLAB Toolbox MatConvNet [53]. The 4096 deep features were extracted right before the classification (soft-max) layer, from the response of the Fully Connected layers (FC7) of the CNN. The runner-up in ILSVRC 2014, known as VGGNet architecture [47], was chosen to train the network. This network contains 16 hidden (Conv/FC) layers.

We subsequently performed PCA on the CNN feature space and retained the components which preserve at least 90% of the variability of the data in the CNN space. This feature space is denoted as CNN PCA(90%) or just CNN90. Different number of components were retained for each video; these numbers are shown in the last column of Table 2.

The last feature space in our collection is semantic labelling (SEM) obtained from the VGGNet classification(soft-max) layer. The output layer of 1000 probability estimates was taken as the feature space, and augmented by one variable to account for people being present in the frame. A non-zero value of this variable means that one of following is detected in the frame: a face, a human figure, or an upper body.[7] The value was rescaled to the magnitude of the largest posterior probability among the 1000 CNN outputs.

[6] For both feature spaces we used the respective functions in the MATLAB Computer Vision toolbox.

[7] The detection was done by the respective MATLAB functions included in the Computer Vision toolbox.

5.3. Experimental protocol

Data. We chose the UTE (UT egocentric) dataset to demonstrate the work of the Greedy Tabu Selector. The UTE dataset [31] contains 4 long videos (each lasting about 3-4 hours) of subjects, performing their daily activities such as driving, shopping, attending lectures and eating. Videos were recorded at 25 frames/second with 350 × 480 resolution per frame. The data set is challenging because it contains a variety of daily activities with frequent illumination changes, camera view shifts, and motion blur.

Method. The proposed Greedy Tabu Selector assumes that the video has already been segmented into units (events). For this experiment, each video was segmented by a subjective opinion. For each video and feature representation, we applied the Greedy Tabu Selector, and calculated the 1-nn resubstitution error. While minimising the error rate is used as a criterion enforcing coverage and diversity, it does not automatically imply high visual quality of the summary or adequate semantic content. We assume that by minimising the error, the obtained summary will be closer to a user-selected summary of the events. Here we rely on the hypothesis that a user would naturally select visually diverse frames, as in our cartoon example in Section 4.2. To evaluate this part, we created a user ground truth summary for each video. To quantify the similarity between the summaries obtained from GTS and GT, we used a well-known measure based on the H-histogram [13], as detailed below. For comparison, we calculated the same values for the Closest-to-Centroid (CC) summary, which we treat as the baseline. An improvement on CC will demonstrate the effectiveness of the edited 1-nn for extracting keyframe summaries.

Matching procedure. Our matching procedure is intended to pair two frames *for the same event* with respect to their visual appearance.

Let f_1 and f_2 be the frames being compared. A 16-bin histogram of the hue value is calculated for each frame. The bin counts are normalised so that the sum is 1 for each histogram. Let $B_j = \{b_{j,1}, \ldots, b_{j,16}\}$ be the normalised histogram for f_j, $j = 1, 2$. The L_1 distance is calculated by

$$D_H = \sum_{i=1}^{16} |b_{1,i} - b_{2,i}|.$$

The two frames are considered matching if $D_H < \theta_H$, where $\theta_H \in [0, 2]$ is a threshold.

Finally, the F-measure is calculated using the number of matches. As both compared summaries have the same number of frames, the F-measure reduces to the proportion of matching frames. Our previous study probing various feature spaces singled out this matching measure as the best one among the alternatives [29], which seems to be in agreement with current practices [13].

The value of the F-measure depends on θ_H. The GTS summary itself depends on the tabu parameter t. We experimented with

- $\theta_H \in \{0.2, 0.3, 0.4, 0.5, 0.6\}$, and
- $t \in \{c - 3, c - 2, c - 1\}$, where c is the number of events.

5.4. Results

The first visual observation during our experiment was that the CC summaries were already an excellent match to the ground truth, as also reported in our earlier publication [28]. In many cases, inspecting the event in the video together with the three visually different summaries (user-GT, CC and GTS) leaves doubts as to which of the three summaries represents the event in the best way. Typically, the GTS frames gave a more diverse visual account of the storyline of the video.

Table 3 shows the F-values and the classification error (in parentheses) for the 4 videos for $\theta_H = 0.6$, and for the three values of the tabu parameter t. We use the following notations: $F(GTS, GT)$, abbreviated as F_{GTS} is the F-value for the comparison of the GTS summary and the user-GT summary. Similarly, $F(CC, GT)$, abbreviated as F_{CC} is the F-value for CC and the user-GT. E denotes the starting resubstitution error obtained with CC as the reference set, and $E_{\min}$ is the resubstitution error with the GTS summary.

Next we examine the effect of parameters θ_H and t. We note that large values of θ_H are more "liberal", and lead to declaring more matches for the same summaries, which results in higher F-values. For the purpose of supporting our point, we look to demonstrate that the F-value for the GTS summary is larger than the F-value for the CC summary. This will indicate that the GTS summary is closer to the ground truth (GT) chosen by the user. Thus, we calculated

$$\Delta F = F_{GTS} - F_{CC},$$

and note that high positive values of ΔF are desirable.

Figure 7 shows ΔF as a function of θ_H for the three values of the tabu parameter t and the 7 feature spaces. Each plot contains the curves for all 7 feature spaces plotted in grey. The curve for the feature space in the title of the plot is shown in black. This allows for an instant comparison of the feature space with the remaining ones. For reference, we plot the 0-line (red) in each plot. If the black curve runs above the 0-line, ΔF is positive, and GTS improves on CC for the respective feature space.

One conclusion from the results so far is that different feature spaces behave differently. It can be observed that HOG, and CNN offer improvements on the baseline for almost all parameter combinations. While CNN and HOG are not affected much by the value of t, RGB and SEM prefer the GTS summaries obtained with tabu parameter $t = c - 3$. The PCA selection and the reduction of the dimensionality does not seem to pay off; the values for CNN90 are lower than those for CNN. The least successful feature spaces in our experiment were LBP and HSV.

To evaluate visually the improvement of GTS over CC for each video, we identified the parameter combination and feature space which lead to the largest ΔF. The results are shown in Figures 8–11. Each figure contains the three summaries: user-GT, CC and GTS. The matches for CC-GT and GTS-GT found by our matching procedure are highlighted by the colour of the rim.[8]

[8] A full set of figures for $\theta = 0.6$, all videos and all feature spaces is shown in the Supplementary material.

Table 3: F-values and classification error (in parentheses, both shown in %) for the 4 videos for $\theta_H = 0.6$, and for the three values of the tabu parameter t. The entries in the boxes highlight the cases where GTS is strictly better than CC ($F_{GTS} > F_{CC}$), and the underlined values, the cases where GTS is strictly worse.

Tabu parameter $t = c - 1$

Feature	Video P01 (10 events)				Video P02 (12 events)				Video P03 (9 events)				Video P04 (10 events)			
space	F_{CC}	E	F_{GTS}	$E_{\min}$	F_{CC}	E	F_{GTS}	$E_{\min}$	F_{CC}	E	F_{GTS}	$E_{\min}$	F_{CC}	E	F_{GTS}	$E_{\min}$
RGB	40	(68)	40	(54)	25	(54)	25	(43)	33	(79)	67	(58)	40	(66)	30	(40)
HSV	50	(49)	30	(38)	58	(48)	50	(38)	78	(62)	44	(45)	50	(56)	30	(33)
LBP	50	(66)	60	(51)	50	(57)	50	(44)	89	(70)	33	(50)	20	(55)	30	(36)
HOG	20	(67)	60	(54)	33	(76)	25	(54)	44	(80)	44	(49)	30	(65)	40	(40)
CNN	50	(47)	70	(34)	42	(27)	58	(20)	56	(64)	78	(29)	30	(40)	50	(20)
CNN90	50	(46)	40	(30)	42	(27)	42	(19)	67	(59)	56	(29)	30	(37)	50	(19)
SEM	40	(67)	40	(50)	42	(56)	33	(45)	44	(72)	33	(37)	30	(58)	10	(48)

Tabu parameter $t = c - 2$

Feature	Video P01 (10 events)				Video P02 (12 events)				Video P03 (9 events)				Video P04 (10 events)			
space	F_{CC}	E	F_{GTS}	$E_{\min}$	F_{CC}	E	F_{GTS}	$E_{\min}$	F_{CC}	E	F_{GTS}	$E_{\min}$	F_{CC}	E	F_{GTS}	$E_{\min}$
RGB	40	(68)	30	(55)	25	(54)	25	(43)	33	(79)	56	(60)	40	(66)	30	(41)
HSV	50	(49)	50	(39)	58	(48)	67	(40)	78	(62)	67	(50)	50	(56)	20	(34)
LBP	50	(66)	50	(54)	50	(57)	50	(45)	89	(70)	56	(54)	20	(55)	40	(36)
HOG	20	(67)	60	(58)	33	(76)	25	(69)	44	(80)	44	(73)	30	(65)	30	(44)
CNN	50	(47)	70	(34)	42	(27)	58	(20)	56	(64)	56	(45)	30	(40)	50	(21)
CNN90	50	(46)	60	(34)	42	(27)	33	(20)	67	(59)	67	(44)	30	(37)	50	(19)
SEM	40	(67)	60	(51)	42	(56)	25	(47)	44	(72)	44	(49)	30	(58)	10	(46)

Tabu parameter $t = c - 3$

Feature	Video P01 (10 events)				Video P02 (12 events)				Video P03 (9 events)				Video P04 (10 events)			
space	F_{CC}	E	F_{GTS}	$E_{\min}$	F_{CC}	E	F_{GTS}	$E_{\min}$	F_{CC}	E	F_{GTS}	$E_{\min}$	F_{CC}	E	F_{GTS}	$E_{\min}$
RGB	40	(68)	50	(55)	25	(54)	58	(44)	33	(79)	56	(61)	40	(66)	40	(46)
HSV	50	(49)	40	(39)	58	(48)	67	(40)	78	(62)	89	(51)	50	(56)	20	(34)
LBP	50	(66)	50	(55)	50	(57)	58	(45)	89	(70)	56	(58)	20	(55)	40	(41)
HOG	20	(67)	50	(60)	33	(76)	25	(69)	44	(80)	44	(73)	30	(65)	20	(45)
CNN	50	(47)	70	(35)	42	(27)	42	(22)	56	(64)	67	(45)	30	(40)	30	(24)
CNN90	50	(46)	50	(36)	42	(27)	42	(21)	67	(59)	56	(44)	30	(37)	20	(23)
SEM	40	(67)	70	(52)	42	(56)	42	(46)	44	(72)	56	(50)	30	(58)	30	(48)

The figures show that our matching algorithm has flaws. Some matches are missed, and some of the found matches are not convincing. Nonetheless, in the absence of a perfect matching algorithm, or one which the community agrees upon, an imperfect algorithm applied across all feature spaces, videos and parameter choices will have to suffice. Our results are in agreement with the general view that high-level feature spaces (CNN, SEM) lead to better summaries. For these spaces, we were able to improve on CC by applying the proposed GTS method.

Assume that the the F-value is a reasonably faithful estimate of the quality of the GTS summary. It would be reassuring if the resubstitution error rate correlated with F. Table 4 shows the correlation between F and E for the best-scoring feature space in our experiment, CNN. To calculate each coefficient, for each video and each t, we concatenated F_{CC} and F_{GTS} for the 5 values of θ_H for each video, thus obtaining a vector $\mathbf{f}$ with 10 values.The same was done for E and $E_{\min}$ to obtain vector $\mathbf{e}$.

Table 4: Correlation coefficients between F-values and the error rate E for the CNN feature space for the 4 videos and the three tabu parameter values.

	$t = c - 1$	$t = c - 2$	$t = c - 3$
P01	−0.2040	−0.2040	−0.1601
P02	−0.2261	−0.2261	0.1048
P03	−0.3246	−0.2010	−0.1448
P04	−0.6509	−0.1843	−0.3592

The entries in the table are the Pearson correlation coefficients between 10-element vectors for F and for E.

The negative values in the table (lower error, higher match) support our overarching hypothesis that classification error can be linked to the interpretability and usefulness of the summary.

GTS has a single tuning parameter, t. In our experiment the results were not significantly different across the values of t which we examined. We propose that for an egocentric video

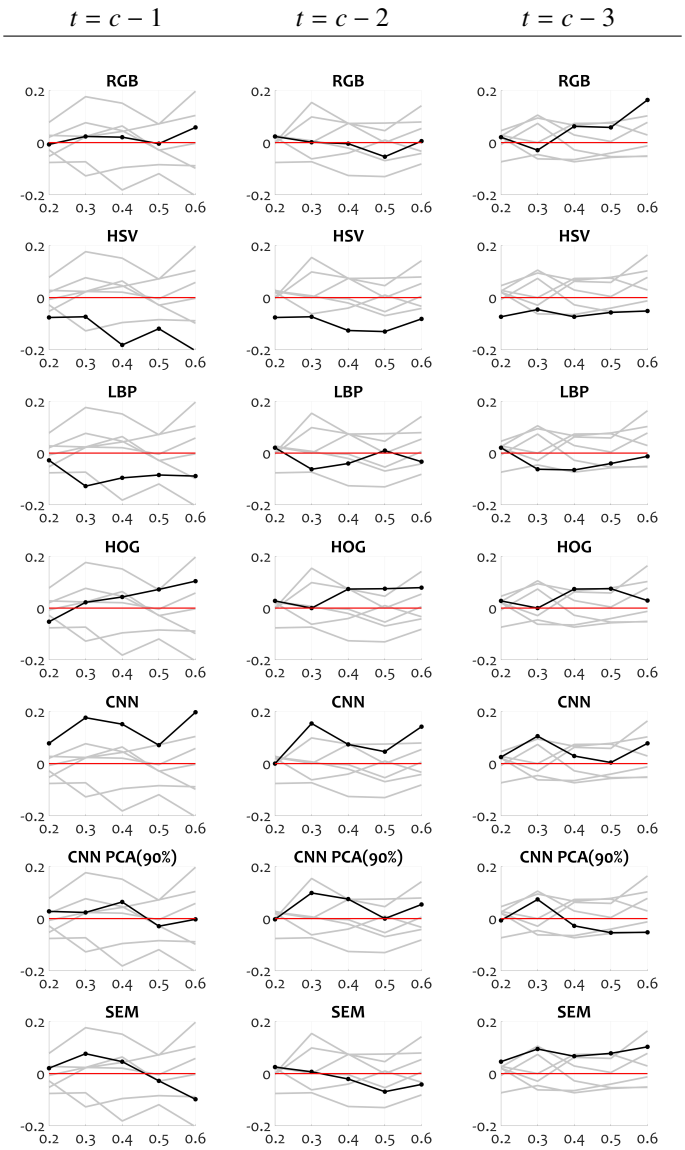


Figure 7: Improvement ΔF for the three values of the tabu parameter t and the 7 feature spaces.

split into 9-12 events, $t = c - 1$ is a good choice, based on the correlation between F and and E in Table 4.

We note that overtraining, which is a major concern in pattern recognition, is not an issue here. Generalisation accuracy of the edited 1-nn classifier is not a quantity of interest because the aim is to minimise the error on the *training* data, given an extremely limited budget of one frame per event.

6. Conclusion

In this study we relate the keyframe selection for video summarisation to prototype (instance) selection for the nearest neighbour classifier (1-nn). Drawing upon this analogy, we propose a Greedy Tabu Selection (GTS) method for extracting a keyframe summary. It is assumed that the video has already been split into units (segments or events), and each such unit is regarded as a class. Our hypothesis is that better 1-nn classification accuracy of the video using the selected set of keyframes as the reference set (resubstitution accuracy) is linked to a better summary.

We compared 7 feature representations including low level features (colour, texture, shape) and high-level features (people and objects). According to our results, the CNN feature space was consistently better than the alternatives. Applying GTS on the CNN space led to better summaries than the baseline ones, obtained through the closest-to-centroid (CC) method.

The difficulties in evaluating summaries for egocentric videos come from several sources. First, because of the intrinsic diversity of each event, many selections of representative frames, which may be visually quite different, could be equally good summaries of the video. Thus a comparison with a single user summary may score low potentially good automatic summaries. Second, the CC baseline is often an excellent summary already, and improvements on that summary may be difficult to rank. This holds in general, not only for the present study. Many times, authors of new video summarisation methods choose baselines which are not very competitive (random, uniform, mid-event), and still, the results from user studies are less impressive than expected. Perhaps this difficulty in distinguishing between summaries within a narrow margin for improvement, combined with the subjective uncertainty involved in any such evaluation are the reason for the lack of large-scale experimental comparisons of video summarisation methods.

There are several interesting directions for further research. First, with a larger budget (more than one frame allowed for each segment), new, more accurate variants of the GTS can be developed. Second, combination of feature spaces can be explored to find even better summaries. While concatenation of feature spaces is a straightforward solution, classifier ensembles may be more effective. Finally, the error-rate criterion for selecting the frames can be combined with quality-enforcing criteria to boost the aesthetic quality of the summary in addition to diversity and coverage. Last but not least, we remark that a lot of effort in developing new summarisation methods may be fruitless without a standard, widely accepted method for comparing keyframe summaries.

Acknowledgements

This work was done under project RPG-2015-188 funded by The Leverhulme Trust, UK. Also, we are grateful to the São Paulo Research Foundation – FAPESP (grant #2016/06441-7).

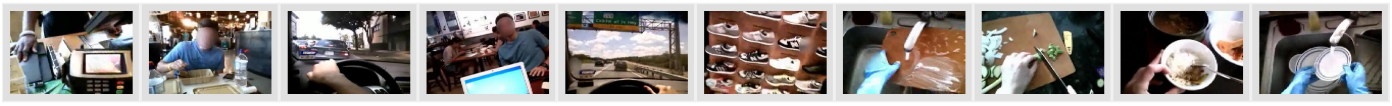
References

[1] Almeida, J., Leite, N. J., Torres, R. S., 2012. VISON: VIdeo Summarization for ONline applications. Pattern Recognition Letters 33 (4), 397–409.

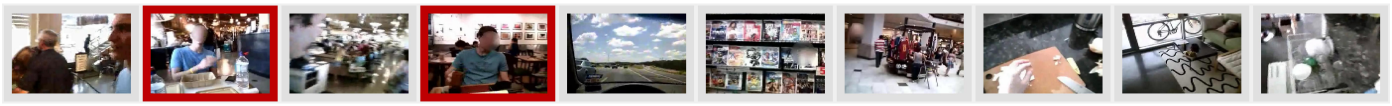
[2] Almeida, J., Leite, N. J., Torres, R. S., 2013. Online video summarization on compressed domain. Journal of Visual Communication and Image Representation 24 (6), 729–738.

[3] Apostolidis, E., Mezaris, V., 2014. Fast shot segmentation combining global and local visual descriptors. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing ICASSP. No. May. pp. 6583–6587.

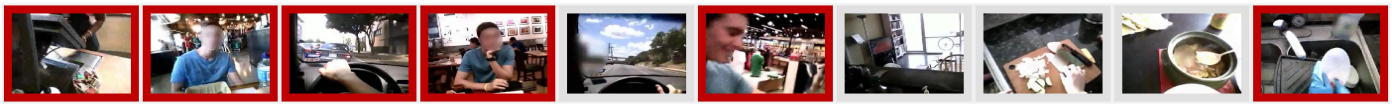
[4] Bambach, S., 2015. A survey on recent advances of comp. vision algorithms for egocentric video. arXiv:1501.02825.



(a) Ground truth

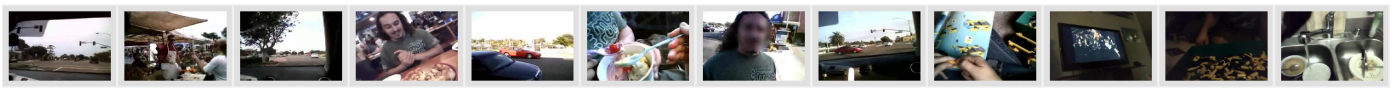


(b) Closet-to-Centroid (CC) summary. Matches with GT are highlighted.

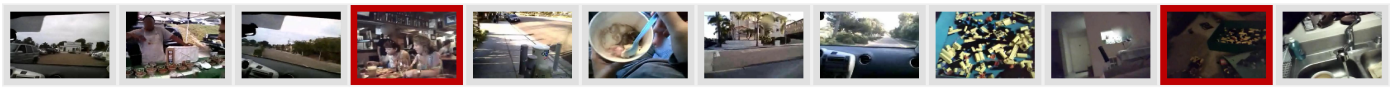


(c) Greedy Tabu Search (GTS) summary. Matches with GT are highlighted.

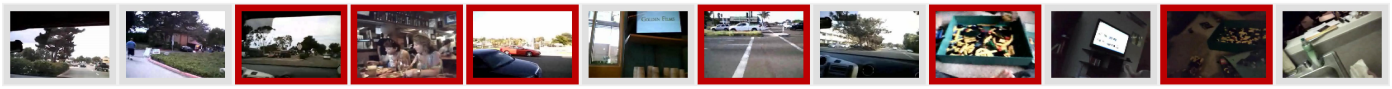
Figure 8: **Video P01.** Summaries: GT, CC and GTS with highlighted matches. $\Delta F = 0.40$ for $\theta_H = 0.6$, $t = c - 1$, space HOG.



(a) Ground truth



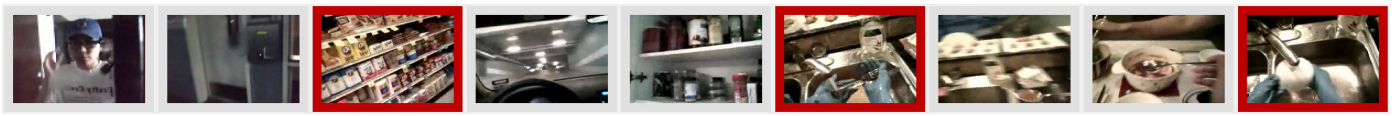
(b) Closet-to-Centroid (CC) summary. Matches with GT are highlighted.



(c) Greedy Tabu Search (GTS) summary. Matches with GT are highlighted.

Figure 9: **Video P02.** Summaries: GT, CC and GTS with highlighted matches. $\Delta F = 0.33$ for $\theta_H = 0.5$, $t = c - 3$, space RGB.

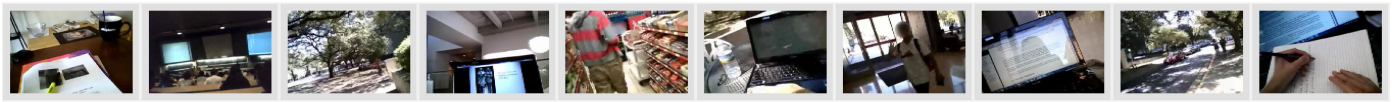
(a) Ground truth



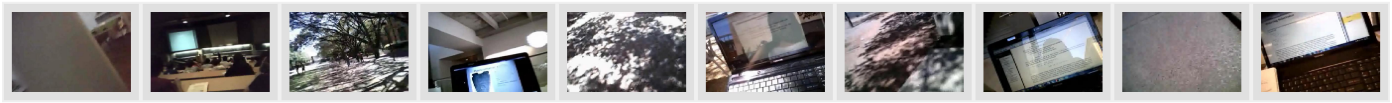
(b) Closet-to-Centroid (CC) summary. Matches with GT are highlighted.

(c) Greedy Tabu Search (GTS) summary. Matches with GT are highlighted.

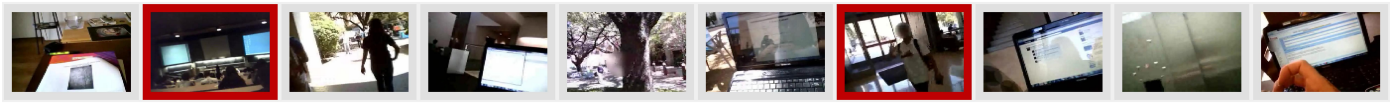
Figure 10: **Video P03.** Summaries: GT, CC and GTS with highlighted matches. $\Delta F = 0.33$ for $\theta_H = 0.6$, $t = c - 1$, space RGB.



(a) Ground truth



(b) Closet-to-Centroid (CC) summary. Matches with GT are highlighted.



(c) Greedy Tabu Search (GTS) summary. Matches with GT are highlighted.

Figure 11: **Video P04.** Summaries: GT, CC and GTS with highlighted matches. $\Delta F = 0.20$ for $\theta_H = 0.2$, $t = c - 1$, space CNN.

[5] Bolaños, M., Dimiccoli, M., Radeva, P., 2017. Toward storytelling from visual lifelogging: An overview. IEEE Transactions on Human-Machine Systems 47 (1), 77–90.

[6] Bolaños, M., Mestre, R., Talavera, E., Giró i Nieto, X., Radeva, P., 2015. Visual summary of egocentric photostreams by representative keyframes. In: Proc. IEEE Int. Multimedia and Expo Workshops. pp. 1–6.

[7] Brighton, H., Mellish, C., 2002. Advances in instance selection for instance-based learning algorithms. Data mining and knowledge discovery 6 (2), 153–172.

[8] Chao, G. C., Tsai, Y. P., Jeng, S. K., 2010. Augmented keyframe. Journal of Visual Communication and Image Representation 21 (7), 682–692.

[9] Chowdhury, S., McParlane, P., Ferdous, M. S., Jose, J., 2015. My day in review: Visually summarising noisy lifelog data. In: Proc. 5th ACM Int. Conf. on Multimedia Retrieval. pp. 607–610.

[10] Cooper, M., Foote, J., 2005. Discriminative techniques for key frame selection. In: Proc. IEEE Int. Multimedia and Expo Workshops (ICME). Vol. 2. pp. 0–3.

[11] Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). Vol. 1. pp. 886–893.

[12] Dasarathy, B. V., 1991. Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques. IEEE Comp. Society Press.

[13] De Avila, S. E. F., Lopes, A. P. B., Da Luz, A., De Albuquerque Araújo, A., 2011. VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method. Pattern Recognition Letters 32 (1), 56–68.

[14] Doherty, A. R., Byrne, D., Smeaton, A. F., Jones, G. J. F., Hughes, M., 2008. Investigating keyframe selection methods in the novel domain of passively captured visual lifelogs. In: Proc. 2008 Int. Conf. on Content-based Image and Video Retrieval CIVR. pp. 259–268.

[15] Ejaz, N., Mehmood, I., Baik, S. W., 2013. Efficient visual attention based framework for extracting key frames from videos. Signal Processing: Image Communication 28 (1), 34–44.

[16] Ejaz, N., Tariq, T. B., Baik, S. W., 2012. Adaptive key frame extraction for video summarization using an aggregation mechanism. Journal of Visual Communication and Image Representation 23 (7), 1031–1040.

[17] Furini, M., Geraci, F., Montangero, M., Pellegrini, M., 2010. STIMO : STIll and MOving Video Storyboard for the Web Scenario. Multimedia Tools and Applications 46 (1), 47–69.

[18] Garcia, S., Derrac, J., Cano, J., Herrera, F., 2012. Prototype selection for nearest neighbor classification: Taxonomy and empirical study. Pattern Analysis and Machine Intelligence, IEEE Transactions on 34 (3), 417–435.

[19] Gianluigi, C., Raimondo, S., 2006. An innovative algorithm for key frame extraction in video summarization. Journal of Real-Time Image Processing 1 (1), 69–88.

[20] Gong, Y., Liu, X., 2003. Video summarization and retrieval using singular value decomposition. Multimedia Systems 9 (2), 157–168.

[21] Gygli, M., Grabner, H., Gool, L. V., 2015. Video summarization by learning submodular mixtures of objectives. In: Proc. IEEE Conf. on Comp. Vision and Pattern Recognition (CVPR). pp. 3090–3098.

[22] Gygli, M., Grabner, H., Riemenschneider, H., Van, L., 2014. Creating summaries from user videos. In: Proc. European Conf. on Comp. Vision (ECCV) 2014. Vol. 8695 LNCS. pp. 505–520.

[23] Hart, P., 1968. The condensed nearest neighbor rule. IEEE Transactions on Information Theory 16, 515–516.

[24] Harvey, M., Langheinrich, M., Ward, G., 2016. Remembering through lifelogging: A survey of human memory augmentation. Pervasive and Mobile Computing 27, 14–26.

[25] Isola, P., Xiao, J., Parikh, D., Torralba, A., Oliva, A., 2014. What makes a photograph memorable? IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (7), 1469–1482.

[26] Jinda-Apiraksa, A., Machajdik, J., Sablatnig, R., 2013. A Keyframe Selection of Lifelog Image Sequences. Proceedings of MVA 2013 IAPR Int. Conf. on Machine Vision Applications, 33–36.

[27] Kim, M. H., Chau, L. P., Siu, W. C., 2012. Keyframe selection for motion capture using motion activity analysis. In: Proc. IEEE Int. Symposium on Circuits and Systems (ISCAS). pp. 612–615.

[28] Kuncheva, L. I., Yousefi, P., Almeida, J., 2017. Comparing keyframe summaries of egocentric videos: Closest-to-centroid baseline. In: Proceedings of the 7th International Conference on Image Processing Theory, Tools and Applications (IPTA 2017).

[29] Kuncheva, L. I., Yousefi, P., Gunn, I. A. D., 2017. On the evaluation of video keyframe summaries using user ground truth. ArXiv:1712.06899.

[30] Le, H. V., Clinch, S., Sas, C., Dingler, T., Henze, N., Davies, N., 2016. Impact of video summary viewing on episodic memory recall: Design guidelines for video summarizations. In: Proc. 2016 CHI Conf. on Human Factors in Computing Systems. pp. 4793–4805.

[31] Lee, Y. J., Ghosh, J., Grauman, K., 2012. Discovering important people and objects for egocentric video summarization. Proc. IEEE Comp. Society Conf. on Comp. Vision and Pattern Recognition, 1346–1353.

[32] Lee, Y. J., Grauman, K., 2015. Predicting important objects for egocentric video summarization. arXiv:1505.04803 abs/1505.04803.

[33] Li, Y., Merialdo, B., 2011. Multi-video summarization based on OBMMR. In: Proceedings - International Workshop on Content-Based Multimedia Indexing. No. Di. pp. 163–168.

[34] Lidon, A., Bolaños, M., Dimiccoli, M., Radeva, P., Garolera, M., i Nieto, X. G., 2015. Semantic summarization of egocentric photo stream events. arXiv:1511.00438.

[35] Liu, G., Wen, X., Zheng, W., He, P., 2009. Shot boundary detection and keyframe extraction based on scale invariant feature transform. In: Proc. 8th IEEE/ACIS Int. Conf. on Comp. and Information Science (ICIS). pp. 1126–1130.

[36] Lu, Z., Grauman, K., 2013. Story-driven summarization for egocentric

video. In: Proc. IEEE Comp. Society Conf. on Comp. Vision and Pattern Recognition. pp. 2714–2721.

[37] Molino, A. G. D., Tan, C., Lim, J. H., Tan, A. H., 2017. Summarization of egocentric videos: A comprehensive survey. IEEE Transactions on Human-Machine Systems 47 (1), 65–76.

[38] Money, A. G., Agius, H., 2008. Video summarisation: A conceptual framework and survey of the state of the art. Journal of Visual Communication and Image Representation 19 (2), 121–143.

[39] Mundur, P., Rao, Y., Yesha, Y., 2006. Keyframe-based video summarization using delaunay clustering. Int. Journal on Digital Libraries 6 (2), 219–232.

[40] Ojala, T., Pietikäinen, M., Mäenpää, T., 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (7), 971–987.

[41] Omidyeganeh, M., Ghaemmaghami, S., Shirmohammadi, S., 2011. Video keyframe analysis using a segment-based statistical metric in a visually sensitive parametric space. IEEE Transactions on Image Processing 20 (10), 2730–2737.

[42] Peng, J., Xiao-Lin, Q., 2010. Keyframe-based video summary using visual attention clues. IEEE Multimedia 17 (2), 64–73.

[43] Potapov, D., Douze, M., Harchaoui, Z., Schmid, C., Sep 2014. Category-specific video summarization. In: Proc. European Conf. on Comp. Vision (ECCV). Zurich, Switzerland.

[44] Priya, G. L., Domnic, S., 2014. Shot based keyframe extraction for ecological video indexing and retrieval. Ecological Informatics 23, 107–117.

[45] Sidiropoulos, P., Mezaris, V., Kompatsiaris, I., 2013. Enhancing video concept detection with the use of tomographs. In: Proc. 2013 IEEE Int. Conf. on Image Processing, (ICIP). pp. 3991–3995.

[46] Sidiropoulos, P., Mezaris, V., Kompatsiaris, I., Meinedo, H., Bugalhol, M., Trancoso, I., 2011. Temporal video segmentation to scenes using high-level audiovisual features. IEEE Transactions on Circuits and Systems for Video Technology 21 (8), 1163–1177.

[47] Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[48] Tapaswi, M., Bauml, M., Stiefelhagen, R., 2014. StoryGraphs: Visualizing character interactions as a timeline. In: Proc. IEEE Comp. Society Conf. on Comp. Vision and Pattern Recognition. pp. 827–834.

[49] Triguero, I., Derrac, J., Garcia, S., Herrera, F., 2012. A taxonomy and experimental study on prototype generation for nearest neighbor classification. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 42 (1), 86–100.

[50] Truong, B. T., Venkatesh, S., 2007. Video abstraction. ACM Transactions on Multimedia Computing, Communications, and Applications 3 (1), 3–es.

[51] Uchihashi, S., Foote, J., Girgensohn, A., Boreczky, J., 1999. Video Manga: generating semantically meaningful video summaries. In: Proc. 7th ACM Int. Conf. on Multimedia (Part 1). pp. 383–392.

[52] Varini, P., Serra, G., Cucchiara, R., 2015. Personalized egocentric video summarization for cultural experience. In: Proc. 5th ACM on Int. Conf. on Multimedia Retrieval. pp. 539–542.

[53] Vedaldi, A., Lenc, K., 2015. Matconvnet – convolutional neural networks for matlab. In: ACM International Conference on Multimedia (ACM-MM'15). pp. 689–692.

[54] Vermaak, J., Perez, P., Blake, A., Gangnet, M., 2002. Rapid summarisation and browsing of video sequences. In: Procedings of the British Machine Vision Conf. 2002. pp. 40.1–40.10.

[55] Vila, M., Bardera, A., Xu, Q., Feixas, M., Sbert, M., 2013. Tsallis entropy-based information measures for shot boundary detection and keyframe selection. Signal, Image and Video Processing 7 (3), 507–520.

[56] Wang, M., Hong, R., Li, G., Zha, Z. J., Yan, S., Chua, T. S., 2012. Event driven web video summarization by tag localization and key-shot identification. IEEE Transactions on Multimedia 14 (4 PART1), 975–985.

[57] Wilson, D., 1972. Asymptotic properties of nearest neighbor rules using edited data. IEEE Transactions on Systems, Man and Cybernetics SMC-2, 408–421.

[58] Wilson, D. R., Martinez, T. R., 2000. Reduction techniques for instance-based learning algorithms. Machine learning 38 (3), 257–286.

[59] Xiong, B., Grauman, K., September 2014. Detecting snap points in egocentric video with a web photo prior. In: Proc. Ecuropean Conf. of Comp. Vision (ECCV). Vol. 8693 LNCS. pp. 282–298.

[60] Yu, X. D., Wang, L., Tian, Q., Xue, P., 2004. Multilevel video representation with application to keyframe extraction. In: Proc. 10th IEEE Int. Multimedia Modelling Conf. pp. 117–123.

[61] Zhuang, Y., Rui, Y., Huang, T. S., Mehrotra, S., 1998. Adaptive key frame extraction using unsupervised clustering. In: Proc. Int. Conf. on Image Processing (ICIP). Vol. 1. pp. 866–870.