Expert Systems With Applications 189 (2022) 116079

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Egocentric video summarisation via purpose-oriented frame scoring and selection

V. Javier Traver [a,*], Dima Damen [b]

[a] Institute of New Imaging Technologies, Av. Vicent Sos Baynat, s/n, Universitat Jaume I, Castellón, Spain
[b] Department of Computer Science, Merchant Venturers Building, Woodland Road, University of Bristol, Bristol BS8 1UB, UK

ARTICLE INFO

Keywords:
Video summarisation
Egocentric vision
General-interest purpose-oriented summarisation
Purpose-specific evaluation metrics

ABSTRACT

Existing video summarisation techniques are quite generic in nature, since they generally overlook the important aspect of what actual purpose the summary will be serving. In sharp contrast with this mainstream work, it can be acknowledged that there are many possible purposes the same videos can be summarised for. Accordingly, we consider a novel perspective: summaries with a purpose. This work is an attempt to both, call the attention on this neglected aspect of video summarisation research, and to illustrate it and explore it with two concrete purposes, focusing on first-person-view videos. The proposed purpose-oriented summarisation techniques are framed under the common (frame-level) *scoring and selection* paradigm, and have been tested on two egocentric datasets, BEOID and EGTEA-Gaze+. The necessary purpose-specific evaluation metrics are also introduced.

The proposed approach is compared with two purpose-agnostic summarisation baselines. On the one hand, a *partially* agnostic method uses the scores obtained by the proposed approach, but follows a standard generic frame selection technique. On the other hand, the *fully* agnostic method do not use any purpose-based information, and relies on generic concepts such as diversity and representativeness. The results of the experimental work show that the proposed approaches compare favourably with respect to both baselines. More specifically, the purpose-specific approach generally produces summaries with the best compromise between summary lengths *and* favourable purpose-specific metrics. Interestingly, it is also observed that results of the partially-agnostic baseline tend to be better than those of the fully-agnostic one. These observations provide strong evidence on the advantage and relevance of purpose-specific summarisation techniques and evaluation metrics, and encourage further work on this important subject.

1. Introduction

Video summarisation has been investigated for both structured third-person point of view (Money & Agius, 2008) and for unstructured, first-person (egocentric) perspective (del Molino, Tan, Lim, & Tan, 2017). A range of approaches has been explored, from supervised methods that learn from available ground-truth summaries produced by human subjects (Zhao, Li, & Lu, 2018) to unsupervised ones which rely on heuristics such as diversity, sparsity or representativeness (Mahasseni, Lam, & Todorovic, 2017; Zhou, Qiao, & Xiang, 2018a). Some forms of weaker supervision or self-supervision have also been explored (Cai, Zuo, Davis, & Zhang, 2018; Panda, Das, Wu, Ernst, & Roy-Chowdhury, 2017; Xiong, Kalantidis, Ghadiyaram, & Grauman, 2019). Recently, innovative proposals address the difficulty of having paired video-summary by learning from unpaired sets (Rochan & Wang, 2019), and the elusive but critical problem of summarisation evaluation is revisited (Abdalla, Menezes, & Oliveira, 2019; Kaushal, Kothawade, Tomar, Iyer, & Ramakrishnan, 2021; Otani, Nakashima, Rahtu, & Heikkila, 2019). Others integrate shot segmentation into the summarisation (Zhao et al., 2018), consider summarising 360° videos (Lee, Sung, Yu, & Kim, 2018), and multi-view videos (Hussain, Muhammad, Ding, Lloret, Baik, & de Albuquerque, 2021). A spatio-temporal U-Net has been proposed for summarisation via reinforcement learning (Liu, Meng, Huang, Vlontzos, Rueckert, & Kainz, 2021), while interactive summarisation (Jin, Song, & Yatani, 2017) aims at providing users with some control. An alternative to frame selection for summarisation can be fast forwarding the less relevant video segments (Silva, Ramos, Ferreira, Chamone, Campos, & Nascimento, 2018), which uses sparse coding techniques. Instead of using sparsity at frame-level, the similarity of temporally close frames can be exploited via block-sparsity (Ma, Mei, Wan, Hou, Wang, & Feng,

* Corresponding author.
E-mail addresses: vtraver@uji.es (V.J. Traver), dima.damen@bristol.ac.uk (D. Damen).

https://doi.org/10.1016/j.eswa.2021.116079
Received 4 May 2020; Received in revised form 7 September 2021; Accepted 10 October 2021
Available online 2 November 2021

2020). Methods of region proposals and action localisation have been leveraged for video summarisation (Zhu, Lu, Li, & Zhou, 2021).

Generative adversarial frameworks based on recurrent neural networks, and Long Short-Term Memory (LSTM) networks in particular, have been explored for video summarisation (Mahasseni et al., 2017; Yuan, Tay, Li, Zhou, & Feng, 2019). To facilitate the training of LSTM-based summarisation techniques, an embedding layer is learned to reduce the dimensionality of the video features (Zhao, Li, & Lu, 2021a). Similar ideas rely on comparing original videos and their summaries in terms of embeddings (Zhang, Grauman, & Sha, 2018) or classification (Zhou, Xiang, & Cavallaro, 2018b). Although general concepts such as diversity and representativeness can be good guidelines for summarisation, mostly for unsupervised approaches (Zhou et al., 2018a), one of their limitations is that the semantics are not properly accounted for. This issue has been addressed by minimising the distance between the textual description generated by a video describer and a human-provided sentence describing the video (Wei, Ni, Yan, Yu, Yang, & Yao, 2018). A similar idea was proposed earlier by mapping the outputs of two networks (a video network and a description network) to a common semantic space (Otani, Nakashima, Rahtu, Heikkilä, & Yokoya, 2016).

The task of video summarisation is related to that of video captioning (Zhang & Peng, 2019) where the main events need to be transcribed. However, these approaches are multi-modal, whereas we focus on single modality (vision) without language. Saliency estimation, either spatially (Xu, Gao, Zhang, Li, & de Albuquerque, 2021) or temporally (Traver, Zorío, & Leiva, 2021), address more basic video-related problems and therefore might serve as building blocks and bring insights to video summarisation. Another approach complements diversity and representativeness with the video reconstructiveness of the candidate summaries (Zhao, Li, & Lu, 2020), with reported competitive results with less training data and even under an unsupervised setting. The availability of multiple videos of the same concept has interestingly been leveraged through co-summarisation (Chu, Song, & Jaimes, 2015) and collaborative summarisation (Panda & Roy-Chowdhury, 2017).

Graph Neural Networks have been proposed in the last two years for video summarisation (Gao, Yang, Zhang, & Xu, 2020; Park, Lee, Kim, & Sohn, 2020; Wu, hua Zhong, & Liu, 2020; Zhao, Li, Lu, & Li, 2021b). For instance, to properly model long-term dependencies, video contents can be represented at two levels (Zhao et al., 2021b): local dependencies between frames are captured with LSTMs, while global relationships among shots are captured with a graph convolutional network. Attention mechanisms have also been explored (Huang, Murn, Mrak, & Worring, 2021; Ji, Zhao, Pang, Li, & Han, 2021). Deep-learning-based approaches to video summarisation have been reviewed recently (Apostolidis, Adamantidou, Metsai, Mezaris, & Patras, 2021).

Despite the progress in the field, the problem of more automatic (e.g. with less human involvement in annotation tasks), robust, and useful (for end users) summarisation has still many open issues. Concretely, most existing approaches assume that the produced summaries aim at some unique general purpose, which is often somehow too generic or ill-defined: for instance, when evaluating with subjects, they may be asked which summary provides a *better overall* summary (Lee & Grauman, 2015).

There has been some effort to provide personalised summaries (Varini, Serra, & Cucchiara, 2017) and query-based summarisation (Huang & Worring, 2020; Sharghi, Gong, & Shah, 2016; Sharghi, Laurel, & Gong, 2017; Xiao, Zhao, Zhang, Yan, & Yang, 2020). These methods rely on user-provided texts, which are used to find relevant frames/shots within the video to build the summary. The problem addressed in this paper differs from these approaches in several respects. First, unlike query-based methods, user input is not required. Second, there are an arbitrarily large number of queries to fit the individual needs or preferences, but we propose summaries which are not person-specific but purpose-specific. Therefore, there can be less possible purposes and each of them can benefit potentially many users in different situations or real-world tasks. Third, in these methods, the summarisation problem is posed similarly to a search or retrieval problem, whereas our purpose-oriented summaries can be arbitrarily complex and can hardly be expressed easily as a simple text query. For these reasons, we propose the *general-interest purpose-oriented* (GIPO from now on) summarisation problem, which represents an intermediate solution between the general-purpose summarisation methods (a single summary fits all users) and those personalised ones (one summary per user). We believe this novel view of the summarisation problem has been largely missing, yet called for.

Like ours, some work (Kanehira, Van Gool, Ushiku, & Harada, 2018) challenges previous research that assumes that just a single summary exists for a given video; but that proposal (Kanehira et al., 2018) relies on the concept of *viewpoint*, defined as a "particular aspect of a video the viewer focuses on". Viewpoints are implicitly defined by video-level similarities. Therefore, the crucial difference between viewpoints-based summarisation and our proposal is that their summariser is guided by the video *contents* rather than by their usage *purpose*. Recently, the concepts of diversity, representativeness and coverage are combined in a unified framework and applied for several summarisation models (Kaushal et al., 2019), depending on the nature of the video category. Beyond this concept, we believe that for videos of the same category, many possible summaries are possible. Certainly, people may want to process a video or its summary with a very particular task or goal in mind, while the possible sets of tasks can be general enough to be of interest to a broad audience. And this is exactly the motivation behind GIPO.

Let us illustrate the GIPO concept with examples of potential purposes valuable for human end users of egocentric videos. First, consider a video of someone performing some task when using some physical product. Two different purposes for this same video would be: one, observing only the skilful parts of the performance for someone learning to perform the task; and two, observing where customers hesitate or find difficulties so that designers/engineers of that product can identify where either redesign or further training or support is required. In the context of a cultural visit, summarisation may entail extracting video segments of the most enjoyed moments, either for reliving (if the viewer is the wearer), or for marketing/recommendation purposes (for users other than the wearer). These example summarisation purposes may partly relate to existing recent research problems (Doughty, Mayol-Cuevas, & Damen, 2019; Jang, Sullivan, Ludwig, Gilchrist, Damen, & Mayol-Cuevas, 2019; Ragusa, Furnari, Battiato, Signorello, & Farinella, 2019). Understandably, general-purpose summaries can hardly serve these specific purposes. In other words, the usage purpose should strongly determine the produced summary. This, in turn, calls for specialised purpose-oriented generation procedures. For the same video, multiple different summaries are possible, depending on the user goal.

In our understanding, this work is a first attempt to start filling this knowledge gap which addresses a relevant practical need and represents a largely unexplored interesting research theme. This general-interest purpose-oriented summarisation problem is illustrated here with two possible purposes. We focus on egocentric videos where wearers perform some tasks and the considered purposes match realistic user needs. On the one hand, we consider a *reviewing* purpose, where a user is interested in finding the (main) steps of the performed tasks; and a *browsing* purpose, where given a certain video collection, the user wants to quickly tell apart videos of different categories. In addition to the summary generation mechanism, evaluating the quality of the generated summaries and/or the summariser is another open problem in summarisation, which is particularly challenging and relevant for purpose-oriented summarisation. Therefore, purpose-tailored evaluation metrics are proposed here as well. These ideas are illustrated in Fig. 1.

In sum, although much research has been performed in the past on video summarisation, and notable advances have been achieved, most of the existing techniques assume a generic summarisation purpose.

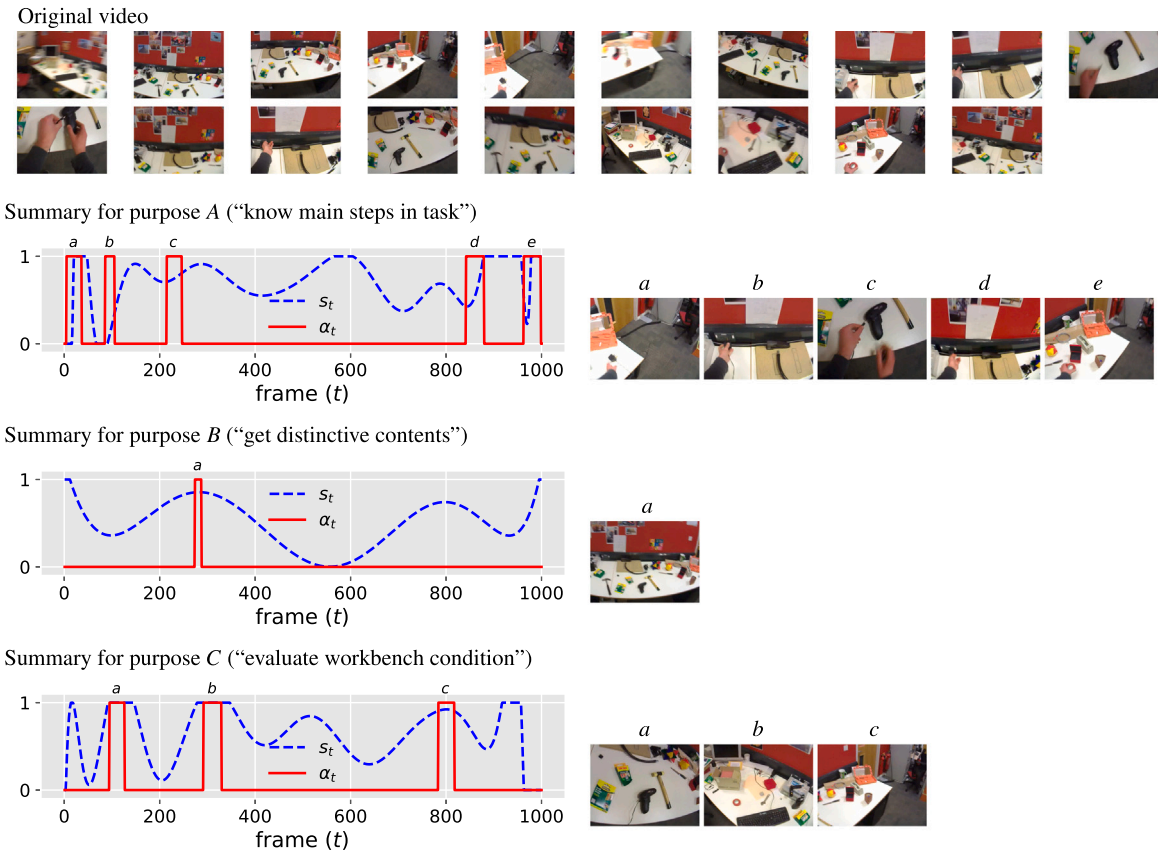

Fig. 1. Illustration of GIPO: for the same original video, many different summaries are possible according to their intended purpose. Therefore, the design of the summarisation method and the metrics to evaluate the quality of the summary must both vary accordingly. For instance, if purpose *A* required extracting the main steps of the task performed in the original video, it would be expected to be longer than a summary for purpose *B* which simply intends to assist the user in getting the gist of the video contents and can fit this purpose more compactly. It would thus be unfair to evaluate the quality of both summaries on the same grounds. Similarly, while the wearer's actions are important for purpose *A*, they are not for purpose *C* aimed at, say, helping the viewer evaluate the workbench conditions. The video and the summaries are simplified here as a few frames for clarity, but higher temporal sampling can actually be used. The summaries are generated by first estimating the scores (s_t, in dashed blue lines) and then selecting the frames (α_t, in red solid lines), and these results vary, for the same input video, as per summarisation purpose. Although this is a diagramatic illustration of purposes and summarisation outputs, purposes *A* and *B* are similar to Purposes 1 and 2 considered in this work. The frames correspond to one of the videos of BEOID, a dataset used in the experiments (Section 3.1).

However, in practice, video summaries should serve different end-user purposes, and this calls for purpose-aware summarisation methodologies. This contrasts with the dominant "one-size-fits-all" mainstream work in video summarisation research. Motivated by this gap, this work: (1) introduces a new research problem: general-interest purpose-oriented summarisation (GIPO); (2) addresses the GIPO problem via a frame scoring-and-selecting paradigm, and proposes purpose-specific evaluation metrics; (3) illustrates the proposed approach on two summarisation purposes and two egocentric video datasets; and (4) demonstrates the effectiveness of the purpose-aware approach by comparing its performance to partially and fully purpose-agnostic summarisation baselines.

2. Methodology

In the following, the GIPO problem is first introduced, and the two purposes used to illustrate it are motivated (Section 2.1). For each of these purposes, the tasks of frame scoring (Section 2.2.1), selection (Section 2.2.2), and evaluation (Section 2.3) are discussed.

2.1. Purposes

We first introduce more formally the problem of general-interest purpose-oriented (GIPO) video summarisation by comparing it to the generic summarisation problem. Given an input video V, generic summarisation approaches aim at producing a single video summary S based on common and general criteria C (e.g. diversity and representativeness) so as to satisfy a set of potential users $\mathcal{U}$ of such summary. In contrast, from the same input video V, GIPO is:

purpose-oriented since it is aimed at producing a different summary S_p for each different purpose p, guided by purpose-specific criteria C_p; and

general-interest since S_p is aimed at satisfying a potentially large subset of users $\mathcal{U}_p \subset \mathcal{U}$ interested in usage scenarios related to purpose p.

A third class of summarisation approaches, personalised summarisation, can actually be seen as a particular case of GIPO when purposes are user-specific. Additionally, it has been discussed above (Section 1) the difficulty of existing personalised approaches, such as query-based ones, to be cast under GIPO, and meet its goals.

Now, we explain the two purposes considered in this work and discuss some realistic usage scenarios behind them.

Purpose 1 (reviewing). For videos where the camera wearer is performing some specific tasks, the video consumer (the viewer) can be interested in getting the relevant *steps* of the performed tasks. More specifically, two examples of practical scenarios where such summaries can be useful are as follows. First, a task-domain expert user may want to *check* that the task has been carried out correctly. For instance, in

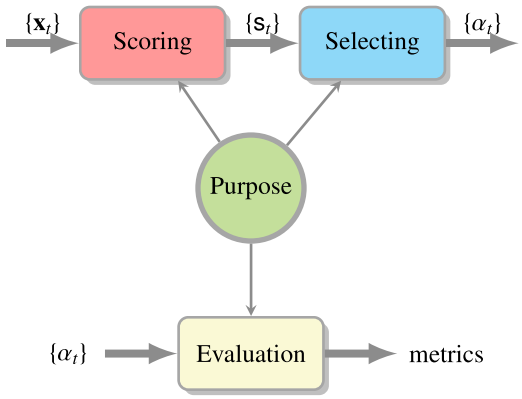

Fig. 2. Scoring and Selecting summarisation framework with GIPO. Frame-level descriptors $\mathbf{x}_t$ of an input video are first scored according to their general value (Section 2.2.1). The subsequent selection procedure considers the frame-level scores $\mathsf{s}_t \in [0,1]$ as input, and produces the selected frames $\alpha_t \in \{0,1\}$ as a summary (Section 2.2.2). For GIPO, the summarisation purpose (Section 2.1) conditions both the scoring and selecting algorithms, and how the quality of the produced summary is assessed (Section 2.3).

a training context, the summariser would assist an instructor to evaluate/verify the trainees' performance. Second, instead of having human experts preparing an explicit and laborious "how-to" user guide of a procedure, the relevant parts of this procedure could be automatically extracted from egocentric video recorded during sample executions, which relates to, and might support, guidance systems (Lu & Mayol-Cuevas, 2019). In both cases, the automatically generated summary would ideally allow their users carry out their supervision/learning tasks more efficiently and effectively than by watching or browsing the whole original videos.

Purpose 2 (browsing). Let us consider a scenario where a set of unlabelled videos of different contents can be available in a given repository, and the user wants to quickly find out distinctive parts of a particular video as compared to others. Performing this task by visually inspecting the full-length videos can be tedious, error-prone and very time consuming, whereas summaries that highlight the distinctive parts of the videos can potentially speed up the procedure. Another potential useful application of this purpose would be the automatic generation of video thumbnails to support (web) revisitation tasks (Leiva, Traver, & Castelló, 2013).

For an easy and quick reference throughout the paper, these purposes are referenced in Table 1.

Although specific usage scenarios have been suggested for each purpose separately, there is potential for their joint use in a given application, as follows. First, the browsing purpose would assist the user in locating some video(s) within a given collection. Next, the review purpose would support the user in watching parts within each of the previously located video(s).

2.2. Scoring and selection framework

The methods developed for both purposes adhere to a common approach of first scoring the frames and then selecting which frames will be part of the summary. Fig. 2 illustrates the respective roles of the scoring and selecting mechanisms for generating the summary from an input video, and that their particular definition vary according to the intended summarisation purpose under the GIPO framework.

2.2.1. Scoring

Both scoring methods discussed here share a frame-level n-dimensional feature vector $\mathbf{x}_t$ for the tth video frame. This feature vector $\mathbf{x}_t$ will alternatively be referred to as the (frame) descriptor. Details of the particular choice used here for $\mathbf{x}_t$ are given later (Section 3.3).

Each summarisation purpose requires a specific frame scoring method which captures the intuition of what frames are, approximately, good candidates for that purpose. The output of this first step is one score per frame at time t, $\mathsf{s}_t \in [0,1]$. These scores are, in turn, the input to the actual selection of frames, which is also purpose-specific, and refines and finally decides which frames to include in the summary.

Purpose 1. Inspired by past work (Zhou et al., 2018a), we use a recurrent neural network (RNN) whose input is a sequence of the frame-level features $\mathbf{x}_t$, and the output are the raw frame-level scores s'_t. A sigmoid $\sigma(z) = (1+\exp(-z))^{-1}$ is applied to these raw scores so that they later lie in $[0,1]$, resulting in the final scores $\mathsf{s}_t = \sigma(\mathsf{s}'_t)$. Then, the higher this score s_t, the more relevant the corresponding frame t is deemed. Previous work on supervised summarisation has used the ground-truth frame relevance according to human-generated summaries. In our case, we rely on action-based annotation available in the used datasets so that those frames in video segments annotated with actions are considered relevant. However, in our case, the frame "relevance" cannot be straightforwardly be interpreted in terms of its appropriateness for its inclusion in the summary, but simply as a first rough approximation which the selection strategy will refine. Fig. 3-A-a illustrates the scoring procedure for Purpose 1.

Purpose 2. As discussed above, considering the distinctive parts of a video compared to other videos is important for this purpose (browsing). Since distinctiveness relates to the classification ability, the scoring mechanism for this purpose is framed into a video classification problem.

For representing and classifying videos, the bag-of-words (BoW) method has widely been used in the past, for instance, for human action recognition (Agustí, Traver, & Pla, 2014; Niebles, Wang, Wang, & Fei-Fei, 2006). Since the BoW technique is relatively simple yet quite successful, it is adopted here. To that end, a vector quantisation of the frame-level features $\mathbf{x}_t$ is computed via clustering of a set of training videos. After this clustering into K codewords, each feature vector $\mathbf{x}_t$ is assigned a cluster index $c_t = \mathsf{C}(\mathbf{x}_t) \in \{1,\dots,K\}$, corresponding to a codeword. The BoW representation of a sequence of frames $\{\mathbf{x}_t\}_t$ is then the histogram $\{\mathbf{h}_k\}_{k=1}^K$ of the respective codeword indexes $\{c_t\}_t$. A video classifier is then trained using these histograms as the input feature vectors.

For frame scoring, we use the feature importance provided by the trained classifier. This way, we get $\mathsf{R}(k)$, the relevance of codeword $k, 1 \le k \le K$, for the K clusters. Specifically, since c_t is the index of the codeword corresponding to the descriptor $\mathbf{x}_t$, then the frame-level score s_t is obtained as the relevance of such codeword c_t, namely, $\mathsf{s}_t = \mathsf{R}(c_t)$.

This procedure is summarised in Fig. 3-B-a

2.2.2. Selecting

Once the scores s_t have been estimated for each frame t, we need to select the frames to build the summary α_t, with $\alpha_t = 1$ if frame t is selected, and $\alpha_t = 0$ otherwise. There are potentially many possible strategies to perform this selection, such as whether a budget for the summary length is required, whether a greedy approach is appropriate, whether temporal constraints on selected frames should be imposed, whether global or local policies are preferred, etc. Since the requirements for the desirable summary are certainly purpose-dependent, the frame selecting strategy is indeed in charge of instilling this knowledge in concrete algorithms.

2.2.2.1. Purpose 1. For defining the selection strategy for this purpose (reviewing), there are two key observations. First, we are generally more interested in selecting a few frames from many segments than many frames from a few segments. Second, the estimated scores tend to increase on the onset of an action, and it is precisely the beginning of an action (and possibly some previous frames providing some temporal context) that can naturally be expected to be more useful in practise for the summary users than some later part, when the action is already

Table 1
The two summarisation purposes considered in this work.

Purpose	Brief description	Potential supported user tasks
1	Review of task steps	Assessing trainees' performance; learning tasks ("how-to" guides)
2	Interactive video browsing	Finding relevant videos; (web page) revisitation aid

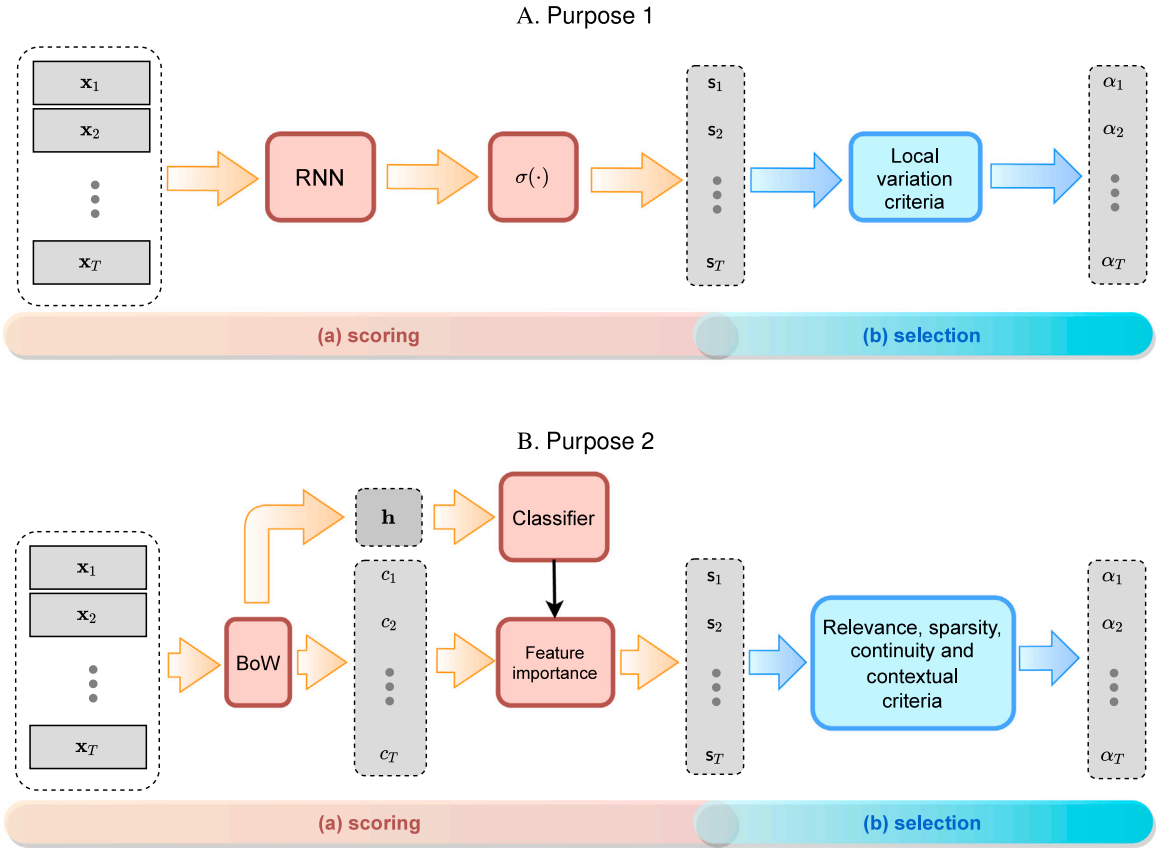

Fig. 3. Schematic diagrams of the scoring and selecting procedures for Purpose 1 (above) and Purpose 2 (below). See text for details.

ongoing. Both requirements imply that it is the temporally *local variation* of the scores that is relevant here for frame selection (Fig. 3-A-b). Therefore, we propose to use the *derivative* of the scores, and use a threshold θ (empirically set to 0.2) for frame selection, i.e.

$$\alpha_t = 1, \text{ if } \frac{\partial \mathsf{s}_t}{\partial t} > \theta. \tag{1}$$

An alternative idea would be thresholding the scores themselves, but this is not a good strategy for this summarisation purpose, since this would easily lead to two unfavourable effects. First, long video segments with scores above a threshold would end up being selected. As a result, some steps may unnecessarily be over-represented in the summary. Second, segments with scores relatively low would go undetected even though they can potentially be relevant. This would result in some task steps being omitted in the summary.

2.2.2.2. Purpose 2 (Fig. 3-b-b). Given the scores s_t for a test sequence, the selection follows the following guidelines: relevance and sparsity (for a summary, keeping only the most relevant frames is required); diversity (temporally close frames, even if relevant, are likely to be similar and therefore, redundant); and temporal continuity and context (humans viewers are unlikely to correctly perceive/recognise the contents of isolated frames within a video). The implementation of these principles (Section 3.3.2) defines the summary α_t.

For video classification purposes, the sequence of frame descriptors corresponding to summary α is formed as $\mathbf{s} = [\mathbf{x}_t : \forall\, t \;\text{ such that }\; \alpha_t = 1]$. Thus, the original sequence of descriptors of length T is reduced to length $T_s = \sum_t \alpha_t \ll T$. Let $\mathbf{X}_k = \{\mathbf{x}_{1:k}\}$ be the set of all frame-level descriptors from frame 1 to k, and $\mathbf{h}^{(k)} = \text{bow}(\mathbf{X}_k)$ the corresponding BoW histogram. When a summary $\mathbf{s}$ is considered, the set $\mathbf{X}_k = \{\mathbf{s}_{1:k}\}$ is alternatively used. For a given sequence of length T_s, classification is repeatedly performed on $\mathbf{h}^{(t)}$ for $t \in \{1, \dots, T_s\}$, so that a class-confidence $c(\ell; t)$ is obtained for each video class $\ell \in \{1, \dots, C\}$ at each frame t. This confidence is not part of the selection mechanism itself, but it will be used for evaluating the summaries.

2.3. Evaluation

One useful property of video summaries is their length, or their ratio with respect to the original video, since shorter summaries are generally preferred over longer ones, as long as they are equally informative. Additionally, for purpose-oriented summaries, purpose-specific quality measures are called for to evaluate how much the summary fits its purpose. We now discuss the proposed evaluation metrics for the two purposes considered in this work. Although generally speaking different purposes will require different evaluation metrics, certainly, this does not prevent *some* metrics to be reused or adapted for *some* (similar) purposes. Regarding proposed metrics such as rank order statistics (Otani et al., 2019), they require ground-truth importance scores, which are not available in our case, and hence not applicable. Additionally, these metrics are oriented to generic summarisation methods, and hence not meaningful to the purposes of our work. Similarly, the CLUSA metric (Abdalla et al., 2019) works on video segment-level

user annotations, which are neither available in the used datasets nor required in our approach.

In the following, we use T and T_s as the lengths, expressed as the number of frames, of the video and its summary, respectively.

Purpose 1. We denote as $a_t = 1$ if frame t is annotated, and $T_a = \sum_t a_t$ as the number of annotated frames in the original video. We first introduce some basic generic metrics that are not actually purpose-specific but provide context to help judge the summarisation performance. One of these metrics is also used to derive one of the purpose-specific metrics.

Basic generic metrics. The compactness of the annotation ($\mathsf{CA} = \frac{T_a}{T} \in [0,1]$) is the percentage of the video frames which are part of the annotation, and depends on the original video, not on its summary. It is therefore a reference metric to help better understand the summary quality in terms of the other metrics. The compactness of the summary with respect to the full video ($\mathsf{C_f}=\frac{T_s}{T}$), and with respect to annotation ($\mathsf{C}_a=\frac{T_s}{T_a}$) provide two complementary ideas of how much the video is summarised. The annotation detection ($\mathsf{AD}=\frac{\sum_t \alpha_t \cdot a_t}{\sum_t \max(\alpha_t, a_t)}$) provides a measure of how much the summary overlap the annotation, normalised by the union of both. However, given the purpose and summarisation method chosen, a higher AD does not necessarily mean a better summary, as discussed above in the selection strategy (Section 2.2.2).

Purpose-specific metrics. A meaningful metric which actually measures the summary quality for this summarisation purpose is the *steps coverage* ($\mathsf{SC} \in [0,1]$), which uses a detailed annotation so that a group of annotated frames is considered detected if at least 5 of their frames are part of the summary. Since SC is the ratio of actions covered, the higher it is the better the summary fits its purpose. We will later see that the ratio $\mathsf{E}=\mathsf{SC}/\mathsf{C_f}$can be useful in some particular cases, as it captures the effectiveness of the summary in covering the steps with respect to its the length. This ratio is unbounded, and for two summaries with either the same $\mathsf{C_f}$or the same SC, the larger E, the better.

Purpose 2. A good summary for this purpose should provide *earlier* high confidence on the true class of the corresponding video. The following measures try to capture this notion.

The *true-to-max confidence ratio* (T2MR) relates the confidence of the true video class label ℓ^* to the maximum confidence at each frame t, $c_{\max}(t) = \max_\ell c(\ell; t)$, and averages this for every frame for the duration of the summary length, T_s. In the case of the full video, its first T_s frames are considered. Thus,

$$\mathsf{T2MR} = \frac{1}{T_s} \sum_{t=1}^{T_s} \frac{c(\ell^*; t)}{c_{\max}(t)} \in [0,1]\,. \tag{2}$$

The higher T2MR, the better, since the maximum confidence is closer to or matches that of the true class, and a correct prediction is therefore more likely. However, although T2MR reflects the confidence ratio, the potential classification ability is not properly considered. Therefore, to complement T2MR, the *true-class rank* (TCR) considers how well the true class is ranked. For C classes, the true class can be ranked ranging from the first position (correct classification) to C-th position (the worst-case misclassification). The rank at each frame t is therefore $r(t) = k$ such that $\ell_k = \ell^*$, where $c(\ell_i; t) > c(\ell_j; t)$ for $i < j$. Then,

$$\mathsf{TCR} = \frac{1}{T_s} \sum_{t=1}^{T_s} \frac{C - r(t)}{C - 1} \in [0,1]\,, \tag{3}$$

where, for easier interpretation, we have linearly mapped $[1, C]$ to $[1, 0]$, so that the best rank ($r = 1$), contributes the most (1), and the worst rank ($r = C$) contributes the least (0). Thus, the higher TCR, the better the average ranking is.

3. Experiments

The proposed methodology is evaluated on two datasets (Section 3.1), by comparing the summarisation performance with sensible baselines (Section 3.4) in terms of their adequacy to the intended purposes (Section 3.5). Validation protocols (Section 3.2) and implementations details (Section 3.3) are also given.

3.1. Datasets

Two egocentric video datasets are used for the experiments. The Bristol Egocentric Object Interactions Dataset (BEOID) (Damen, Leelasawassuk, Haines, Calway, & Mayol-Cuevas, 2014a; Damen et al., 2014b) consists of 58 videos of six locations, which we will compactly refer to as desk, door, printer, sink, row, and treadmill, where operators carry out the given verbal instructions.

The Extended Georgia Tech Egocentric Activity Gaze+ (EGTEA-Gaze+) (Li, Liu, & Rehg, 2018, 2019) consists of 86 sequences. The 32 participating subjects perform seven different recipes, which we will refer to as greekSalad, pastaSalad, pizza, bacon&eggs, continentalBreakfast, cheeseBurger and turkeySandwich. Additional annotations of EGTEA-Gaze+ provided in (Hahn, 2019) for (Hahn, Ruiz, Alayrac, Laptev, & Rehg, 2018), such as actual start-end frame numbers of the videos, were used.

We will use the term *scenario* to mean either the locations in BEOID or the recipes in EGTEA-Gaze+. All frames were considered in BEOID videos, but the videos in EGTEA-Gaze+, which are longer, were temporally subsampled and one frame every twelve were taken. Both datasets are considered when testing Purpose 1, but the nature of BEOID dataset makes it unsuitable for Purpose 2 since the different locations are very different one to each other, and distinguishing between them can thus be a relatively simple task, whereas all recipes in EGTEA-Gaze+ share a common location (a kitchen) with many similar objects (pans, kettles, fridge, etc.) and actions shared across recipes, thus resulting in a more challenging task.

3.2. Validation protocols

Regarding the validation protocols, for Purpose 1, a leave-one-recipe-out is employed. The RNN models are therefore trained with snippets of the videos in the training recipes. Testing is performed on the full videos of the test scenarios. For Purpose 2, a leaving-one-video-out procedure is used. Therefore, after excluding instances whose files were problematic, 81 videos were considered for a total of 7 categories. Five repetitions were performed to account for the non-deterministic aspects in the clustering procedure. Results are reported for these $81{\cdot}5 = 405$ instances.

3.3. Implementation details

For the frame-level feature vector, the activations of layer avg-pool of the Keras implementation (Chollet et al.) of the Inception V3 model (Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016) pretrained on ImageNet, is extracted for each frame of the video sequences and represented by $\mathbf{x}_t \in \mathbb{R}^{2048}$ for the tth frame.

3.3.1. Purpose 1

For the scoring, a bidirectional long-term short memory (LSTM) network (Sak, Senior, & Beaufays, 2014) is used as the RNN, and trained supervisedly using action annotations. The PyTorch framework (Paszke et al., 2019) was used for this LSTM-based network. The mean squared error was used as the loss function. Since a pretrained CNN is used as the input to the RNN network, this is typically seen as a CNN-RNN network. The RNN part of this network is trained for 20 epochs using snippets of the training videos. The batch size is a single sequence, which allows for different-length sequences to be used. Regarding the

appropriate length of the training snippets, it can be considered that the shorter they are, the more training instances will be available, but less temporal context will be used. In some preliminary experiments, snippets of average lengths $L \in \{20, 40, 80, 160, 320\}$ frames were tested, with not much difference among them, but with $L = 80$ providing a better tradeoff between training set size and temporal-context modelling, and used for all the reported experiments. The training snippets were sampled from the training videos with a 20% overlap. At inference time, the full-length video is used as input.

3.3.2. Purpose 2

For clustering, a multibatch k-means is performed for the frames in the training videos, and a classifier is trained on the resulting BoW of those videos. The number of clusters tested were in the order of a few tens or a few hundreds, with similar results, and $k = 50$ was finally selected. The BoW pipeline is popular and detailed references can be found, as in Traver, Latorre-Carmona, Salvador-Balaguer, Pla, and Javidi (2014). Preliminary results with and without class imbalance correction gave similar results, and no correction was finally applied. A total of 10,000 data points were sampled from the available training frames. A random forest was used as the classifier, whose parameters were selected to avoid overfitting, with reasonable results obtained with 50 trees of maximum depth of 5 levels, and up to 5 leaf nodes and 3 features to consider when defining the best split. The Python's scikit-learn machine learning library (Pedregosa et al., 2011) was used for the clustering and classification procedures.

For implementing the selecting principles (Section 2.2.2.2), the score signal s_t is first temporally smoothed with a Gaussian kernel of standard deviation $\sigma = 150$ (frame units) to regularise it and deal with estimation noise. Then, the peaks with at least a prominence of $\theta_p = 0.025$ (relevance and sparsity) and with a minimum separation of $\Delta_t = 30$ frames (diversity and sparsity) are selected, and a temporal window corresponding to $t = 3$ s centred at the peak locations is taken (for temporal continuity and context). Algorithm 1 formalises the procedure. For peak processing, the module signal from the scipy package was used.

3.4. Baselines

Since we are interested in evaluating how the purpose specificity helps in generating purpose-specific summaries, we propose baselines which are partly similar to the purpose-specific summarisers, but lack the purpose-related information at some point, thus effectively turning them into more *purpose agnostic*. Note that these can be considered *strong* baselines in the sense that they still rely on some information (data or algorithmic) that a general-purpose algorithm would lack. In other words, the baselines can be seen as "privileged" summarisers since they exploit either the available scores (the baseline for Purpose 1) or ideas of the frame selection (the baselines for Purpose 2), as detailed below.

For Purpose 1, the baseline shares the scoring strategy of the proposed purpose-specific summariser, but differs in the selection strategy. Recent summarisation works (Mahasseni et al., 2017; Zhou et al., 2018a) use the combination of Kernel Temporal Segmentation, KTS (Potapov, Douze, Harchaoui, & Schmid, 2014) with 0/1 knapsack (Kellerer, Pferschy, & Pisinger, 2004) for selecting frames given the frame-level scores. Thus, KTS is used here as a baseline. For this particular case, to better quantify the performance of the purpose-oriented with respect to the oracle baseline, since they will have a common C_f, the ratio $E{=}SC/C_f$ is useful. To gain further insight into the performance of a fully purpose-agnostic algorithm, the Deep Semantic Features (DSF) algorithm (Otani et al., 2016)[1] is used in BEOID as an additional baseline. Since DSF aims at extracting representative and diverse video segments, it represents well general-purpose summarisation methods. As with KTS, DSF is evaluated with different summarisation ratios p as well.

It is worth noting that, interestingly, one significant advantage of the proposed approach over global techniques such as KTS is that it lends itself more easily to on-line summarisation, since it does not require any temporal segmentation nor first inspecting the full video for neither frame scoring nor selecting.

As for the baseline for Purpose 2, a uniform sampling was considered by selecting as many equally-spaced frames as the number of peaks selected by the summarisation method. Thus, the baseline does not use any frame scoring but exploits the algorithmic idea behind the selection strategy of how much to sample from the original input video.

[1] https://github.com/adityashukla17/Video-Summarization.

Table 2
Global results of Purpose 1 in BEOID.

Location: All, $n = 58$, CA: 0.38 (0.17)					
Method	C_f	C_a	AD	SC	$E = SC/C_f$
KTS(0.15)	0.15 (0.01)	0.48 (0.23)	0.11 (0.02)	0.70 (0.23)	4.82 (1.58)
KTS(0.35)	0.35 (0.01)	1.13 (0.54)	0.21 (0.06)	0.88 (0.14)	2.56 (0.43)
KTS(0.55)	0.55 (0.01)	1.79 (0.84)	0.28 (0.09)	0.96 (0.10)	1.75 (0.18)
KTS(p^*)	0.20 (0.12)	0.58 (0.29)	0.15 (0.08)	0.76 (0.21)	4.87 (3.16)
DSF(0.15)	0.21 (0.16)	0.70 (0.59)	0.15 (0.16)	0.48 (0.33)	2.66 (1.99)
DSF(0.35)	0.33 (0.11)	1.07 (0.59)	0.22 (0.14)	0.68 (0.27)	2.11 (0.77)
DSF(0.55)	0.47 (0.08)	1.53 (0.76)	0.28 (0.13)	0.84 (0.20)	1.78 (0.37)
GiPo	0.21 (0.12)	0.59 (0.30)	0.15 (0.08)	0.82 (0.21)	**5.01** (2.94)

Table 3
Global results of Purpose 1 in EGTEA-Gaze+.

Recipe: All, $n = 79$, CA: 0.27 (0.11)					
Method	C_f	C_a	AD	SC	$E = SC/C_f$
KTS(0.15)	0.15 (0.00)	0.70 (0.44)	0.24 (0.09)	0.54 (0.20)	3.63 (1.37)
KTS(0.35)	0.35 (0.00)	1.66 (1.03)	0.38 (0.12)	0.77 (0.16)	2.20 (0.45)
KTS(0.55)	0.55 (0.00)	2.61 (1.62)	0.40 (0.15)	0.91 (0.11)	1.66 (0.20)
KTS(p^*)	0.21 (0.20)	0.98 (1.29)	0.24 (0.14)	0.53 (0.28)	3.63 (2.65)
GiPo	0.22 (0.20)	0.99 (1.30)	0.17 (0.08)	0.81 (0.21)	**6.56** (4.79)

3.5. Results

3.5.1. Purpose 1

In the following, performance for each metric is given as mean (std. dev). In the BEOID case, the overall results (Table 2) indicate that, in comparison to the KTS and DSF baselines, the GiPo approach produces summaries with the best tradeoff between steps coverage SC and summary length C_f. For the same p, DSF(p) gets lower steps coverage (SC) and poorer overall effectiveness (E) than KTS(p). Additionally, to get similar values of SC, longer summaries (C_f) are required in DSF than in KTS. It is interesting to note that for the same summary ratio (C_f=0.21), DSF gets SC = 0.48, i.e. only about half GiPo's steps coverage (SC = 0.82). And, for a similar ratio (C_f=0.20), however, KTS(p^*) gets SC = 0.76, significantly higher than SC = 0.48 with DSF. These relative performances are consistent with how purpose-aware each of these baselines are: KTS is partially purpose-aware since it shares the scores produced by GiPo but uses a purpose-agnostic selecting algorithm, whereas DSF is totally purpose-agnostic. The results per location (Tables 4 and 5) are similar, and support this general trend. For some summarisation percentages p, SC is higher with KTS(p) than with GiPo, but with longer summaries (higher C_f). Even the privilegedly informed KTS(p^*) has lower performance than GiPo in all but one scenario (treadmill).

Since the row and treadmill scenarios include many repetitive actions, we explored the effect on summarisation of an unsophisticated cycle detector and removal (CDR) algorithm (Appendix). It can be observed (Table 5) that the inclusion of CDR (denoted as GiPo w/

Algorithm 1. *Frame selection for Purpose 2*

Input: *Frame scores* $\mathsf{s} = (s_1, \ldots, s_T)$

Output: *Summary* $\alpha = (\alpha_1, \ldots, \alpha_T)$

$\mathsf{s} \leftarrow \mathsf{s} * G_\sigma$ ▷ *Smooth scores by convolution with Gaussian kernel* G_σ

$\mathcal{P} \leftarrow \mathit{findPeaks}(\mathsf{s}, \Delta t)$ ▷ *Locate peaks of minimum separation* Δt

$\{p\} \leftarrow \mathit{getPeakProminences}(\mathsf{s}, \mathcal{P})$ ▷ *Compute how prominent each peak is*

$p_i = \dfrac{p_i}{\max_i p_i},\ i \in \{1, \ldots, |\mathcal{P}|\}$ ▷ *Normalise peak prominences*

$\mathcal{P} \leftarrow \left\{z_i \in \mathcal{P} : p_i \geq \theta_p\right\}$ ▷ *Select peaks with minimum prominence* θ_p

$\alpha_i \leftarrow \delta(i \in \mathcal{P}),\ i \in \{1, \ldots, T\}$ ▷ *Select locations corresponding to peaks;* $\delta(\cdot)$ *represents the Kronecker delta function*

$\alpha \leftarrow \mathit{binaryDilation}(\alpha, t)$ ▷ *Expand selection around* α_i *to have segments of minimum length t (seconds)*

end

Table 4
Per-location results of Purpose 1 in BEOID.

Location: door, $n = 10$, CA: 0.3 (0.09)

Method	C_f	C_a	AD	SC	$E = SC/C_f$
KTS(0.15)	0.14 (0.01)	0.50 (0.15)	0.10 (0.01)	0.80 (0.22)	**5.79** (1.58)
KTS(0.35)	0.33 (0.02)	1.24 (0.40)	0.18 (0.03)	0.97 (0.07)	2.91 (0.19)
KTS(0.55)	0.54 (0.01)	1.97 (0.61)	0.23 (0.05)	0.97 (0.07)	1.81 (0.14)
DSF(0.15)	0.52 (0.11)	1.80 (0.35)	0.39 (0.22)	0.90 (0.12)	1.79 (0.33)
DSF(0.35)	0.52 (0.11)	1.80 (0.35)	0.39 (0.22)	0.90 (0.12)	1.79 (0.33)
DSF(0.55)	0.52 (0.11)	1.80 (0.35)	0.39 (0.22)	0.90 (0.12)	1.79 (0.33)
KTS(p^*)	0.22 (0.07)	0.80 (0.31)	0.14 (0.04)	0.85 (0.23)	4.19 (1.49)
GiPo	0.23 (0.06)	0.85 (0.31)	0.15 (0.04)	0.95 (0.10)	4.29 (0.77)
KTS(p^*)	0.22 (0.07)	0.80 (0.31)	0.14 (0.04)	0.85 (0.23)	4.19 (1.49)
GiPo w/ CDR	0.23 (0.06)	0.85 (0.31)	0.15 (0.04)	0.95 (0.10)	**4.29** (0.77)

Location: sink, $n = 10$, CA: 0.4 (0.06)

Method	C_f	C_a	AD	SC	$E = SC/C_f$
KTS(0.15)	0.15 (0.00)	0.38 (0.07)	0.12 (0.01)	0.53 (0.12)	3.59 (0.79)
KTS(0.35)	0.35 (0.00)	0.90 (0.15)	0.23 (0.02)	0.80 (0.14)	2.30 (0.40)
KTS(0.55)	0.55 (0.00)	1.41 (0.24)	0.30 (0.03)	0.87 (0.14)	1.58 (0.25)
DSF(0.15)	0.10 (0.02)	0.26 (0.08)	0.05 (0.02)	0.13 (0.14)	1.35 (1.20)
DSF(0.35)	0.29 (0.02)	0.76 (0.16)	0.22 (0.04)	0.57 (0.14)	1.94 (0.44)
DSF(0.55)	0.47 (0.02)	1.22 (0.22)	0.25 (0.05)	0.83 (0.14)	1.75 (0.30)
KTS(p^*)	0.19 (0.21)	0.46 (0.47)	0.13 (0.10)	0.60 (0.17)	4.64 (2.00)
GiPo	0.19 (0.21)	0.47 (0.47)	0.13 (0.10)	0.63 (0.19)	**4.84** (2.12)
KTS(p^*)	0.16 (0.16)	0.39 (0.36)	0.12 (0.08)	0.57 (0.18)	4.91 (2.03)
GiPo w/ CDR	0.16 (0.16)	0.39 (0.35)	0.12 (0.08)	0.60 (0.21)	**4.98** (1.94)

Location: printer, $n = 10$, CA: 0.36 (0.12)

Method	C_f	C_a	AD	SC	$E = SC/C_f$
KTS(0.15)	0.14 (0.00)	0.46 (0.19)	0.11 (0.01)	0.60 (0.20)	**4.18** (1.35)
KTS(0.35)	0.34 (0.01)	1.09 (0.45)	0.21 (0.04)	0.90 (0.15)	2.63 (0.47)
KTS(0.55)	0.55 (0.00)	1.74 (0.69)	0.27 (0.07)	0.97 (0.10)	1.76 (0.18)
DSF(0.15)	0.25 (0.03)	0.77 (0.24)	0.14 (0.12)	0.37 (0.18)	1.43 (0.63)
DSF(0.35)	0.25 (0.03)	0.77 (0.24)	0.14 (0.12)	0.37 (0.18)	1.43 (0.63)
DSF(0.55)	0.39 (0.09)	1.30 (0.69)	0.21 (0.13)	0.57 (0.21)	1.44 (0.35)
KTS(p^*)	0.20 (0.06)	0.61 (0.24)	0.14 (0.04)	0.73 (0.20)	4.02 (1.46)
GiPo	0.21 (0.06)	0.63 (0.24)	0.15 (0.04)	0.80 (0.22)	4.17 (1.34)
KTS(p^*)	0.18 (0.07)	0.54 (0.25)	0.13 (0.04)	0.60 (0.25)	**3.69** (1.43)
GiPo w/ CDR	0.18 (0.07)	0.56 (0.25)	0.13 (0.04)	0.63 (0.28)	3.66 (1.30)

Location: desk, $n = 10$, CA: 0.33 (0.10)

Method	C_f	C_a	AD	SC	$E = SC/C_f$
KTS(0.15)	0.15 (0.00)	0.49 (0.16)	0.11 (0.01)	0.63 (0.19)	4.28 (1.29)
KTS(0.35)	0.35 (0.00)	1.15 (0.36)	0.20 (0.04)	0.85 (0.14)	2.46 (0.41)
KTS(0.55)	0.55 (0.00)	1.81 (0.57)	0.26 (0.06)	1.00 (0.00)	1.82 (0.00)
DSF(0.15)	0.13 (0.02)	0.40 (0.09)	0.11 (0.05)	0.27 (0.12)	2.11 (0.89)
DSF(0.35)	0.28 (0.03)	0.94 (0.33)	0.21 (0.09)	0.55 (0.17)	1.96 (0.64)
DSF(0.55)	0.48 (0.03)	1.58 (0.53)	0.28 (0.06)	0.76 (0.16)	1.61 (0.36)
KTS(p^*)	0.18 (0.05)	0.58 (0.12)	0.13 (0.04)	0.64 (0.12)	3.73 (1.32)
GiPo	0.19 (0.05)	0.58 (0.13)	0.14 (0.04)	0.80 (0.15)	**4.43** (0.88)
KTS(p^*)	0.13 (0.05)	0.44 (0.25)	0.10 (0.04)	0.63 (0.23)	**4.92** (1.20)
GiPo w/ CDR	0.13 (0.05)	0.45 (0.25)	0.10 (0.04)	0.60 (0.21)	4.66 (1.06)

Table 5
Per-location results of Purpose 1 in BEOID (contd. from Table 4).

Location: treadmill, $n = 9$, CA: 0.19 (0.05)

Method	C_f	C_a	AD	SC	$E = SC/C_f$
KTS(0.15)	0.15 (0.00)	0.82 (0.20)	0.09 (0.01)	0.94 (0.16)	6.30 (1.05)
KTS(0.35)	0.35 (0.00)	1.92 (0.47)	0.14 (0.03)	1.00 (0.00)	2.86 (0.00)
KTS(0.55)	0.55 (0.00)	3.01 (0.74)	0.17 (0.04)	1.00 (0.00)	1.82 (0.00)
DSF(0.15)	0.14 (0.01)	0.79 (0.18)	0.09 (0.04)	0.88 (0.22)	6.35 (1.55)
DSF(0.35)	0.31 (0.02)	1.80 (0.41)	0.15 (0.05)	0.94 (0.17)	2.98 (0.54)
DSF(0.55)	0.48 (0.02)	2.76 (0.61)	0.19 (0.06)	1.00 (0.00)	2.08 (0.08)
KTS(p^*)	0.10 (0.02)	0.54 (0.20)	0.07 (0.01)	0.94 (0.16)	**10.39** (3.78)
GiPo	0.10 (0.02)	0.54 (0.20)	0.07 (0.01)	0.89 (0.21)	9.81 (4.12)
KTS(p^*)	0.03 (0.01)	0.13 (0.04)	0.02 (0.01)	0.83 (0.24)	**38.77** (18.17)
GiPo w/ CDR	0.03 (0.01)	0.13 (0.04)	0.02 (0.01)	0.78 (0.34)	31.05 (18.54)

Location: row, $n = 9$, CA: 0.68 (0.05)

Method	C_f	C_a	AD	SC	$E = SC/C_f$
KTS(0.15)	0.15 (0.00)	0.22 (0.02)	0.14 (0.00)	0.71 (0.25)	**4.77** (1.67)
KTS(0.35)	0.35 (0.00)	0.51 (0.03)	0.30 (0.01)	0.85 (0.13)	2.44 (0.39)
KTS(0.55)	0.55 (0.00)	0.81 (0.05)	0.44 (0.02)	0.97 (0.08)	1.77 (0.14)
DSF(0.15)	0.11 (0.02)	0.16 (0.04)	0.11 (0.02)	0.43 (0.21)	3.77 (1.35)
DSF(0.35)	0.30 (0.02)	0.45 (0.04)	0.20 (0.04)	0.84 (0.19)	2.79 (0.67)
DSF(0.55)	0.48 (0.03)	0.70 (0.07)	0.33 (0.05)	1.00 (0.00)	2.10 (0.16)
KTS(p^*)	0.32 (0.02)	0.47 (0.05)	0.28 (0.01)	0.82 (0.13)	2.58 (0.40)
GiPo	0.32 (0.02)	0.47 (0.05)	0.28 (0.01)	0.89 (0.17)	2.78 (0.55)
KTS(p^*)	0.12 (0.04)	0.17 (0.06)	0.11 (0.03)	0.66 (0.26)	5.64 (1.46)
GiPo w/ CDR	0.12 (0.04)	0.17 (0.06)	0.11 (0.03)	0.86 (0.17)	**8.16** (2.88)

CDR in the tables) induces a moderate reduction in SC, and a significant improvement in the compactness of the summary (about three times smaller C_f, and about three-four times smaller C_a). It can be an application-dependent choice whether to prioritise on SC or C_f.It is also interesting to observe that endowed with the ideal reduction, KTS(p^*) can provide competent results. This observation suggests that a proper combination of the proposed GiPo approach and existing solutions can prove beneficial. It is important to remember, however, the advantage of the proposed approach for scenarios of on-line summarisation, as mentioned above (Section 3.4).

When the CDR is applied on the sequences without any (apparent) cycle, the results (Tables 4) are exactly the same for the door scenario, but may have a negative impact in the summaries of videos of some other sequences (e.g. in desk, SC drops from 0.80 without CDR to 0.63 with CDR) due to the yet imperfect nature of the CDR. Although not the main topic in this work, it is interesting to note that detecting and removing cycles is a very important yet largely unexplored area in the summarisation literature.

Similar observations emerge in the EGTEA-Gaze+ case. Both the global results (Table 3) and per-recipe ones (Table 6) indicate that whenever KTS has a good SC it is at the expense of longer summaries (higher C_f), while the purpose-oriented summarisation provides the best results, i.e. the highest steps coverage with the shortest summaries. However, when provided with the "ideal" summary length, the steps coverage produced by KTS(p^*) is significantly lower (worse). This clearly illustrates that the proposed purpose-oriented summarisation is effective for the purpose it was designed for, and that general-purpose approaches are suboptimal at best. Even though the baseline uses scores learned with the purpose-oriented method, its selection mechanism is meant for a general-purpose summarisation. This observation indicates that it is important that both the scoring and the selecting procedures

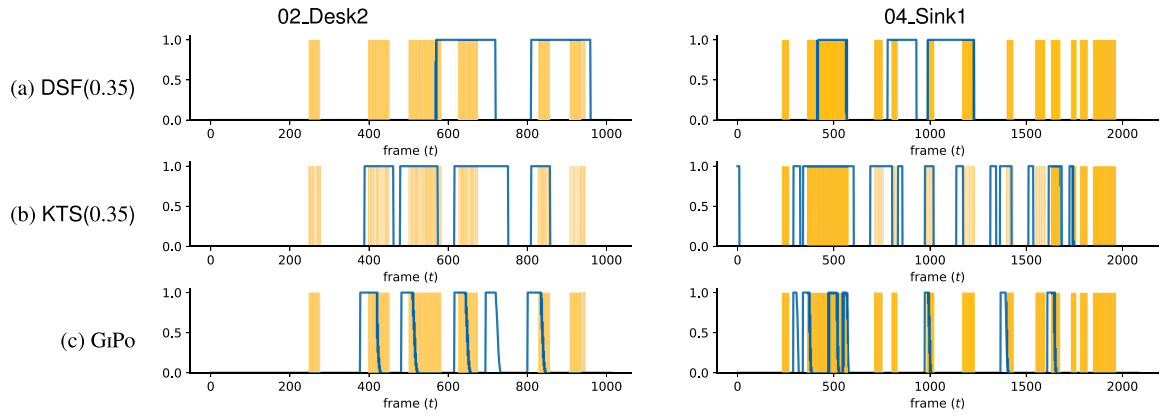

Fig. 4. Example of summaries for Purpose 1 with DSF (top), KTS (middle) and GIPo (bottom) for two sequences in desk (left) and sink (right) scenarios. Blue lines represent the summary α (i.e. the selected frames), and shaded orange regions correspond to the frames with ground-truth annotated actions.

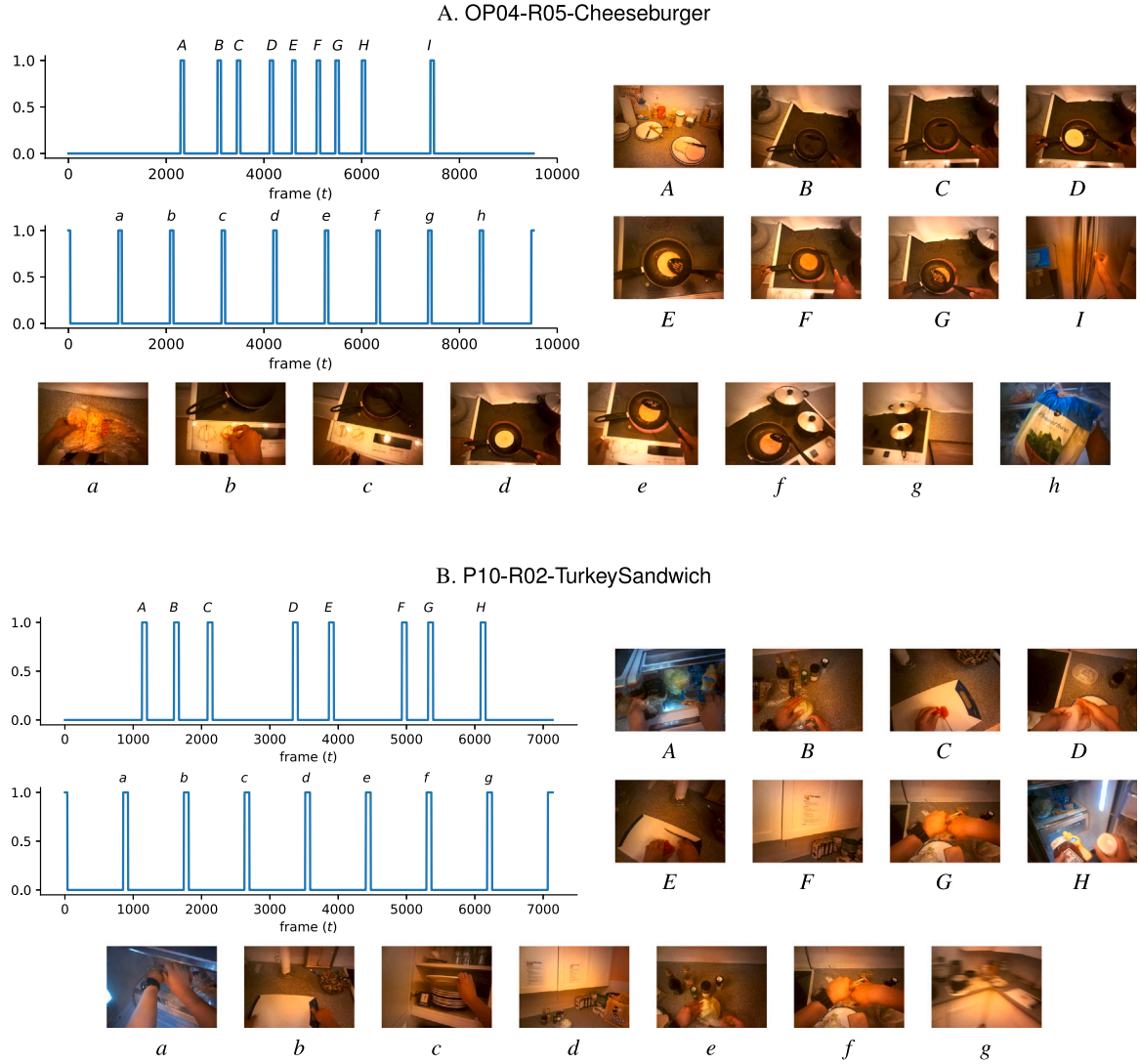

Fig. 5. Example of summaries for Purpose 2 for sequences of cheeseBurger (above) and turkeySandwich (below). Frame segments selected by GIPo are marked in upper-case letters ($A, B, \ldots$) and those selected by the uniform-sampling baseline are marked in lower-case letters ($a, b, \ldots$). The displayed images correspond to the centre frame of these segments.

Table 6
Per-recipe results of Purpose 1 in EGTEA-Gaze+.

Recipe: pizza, $n = 5$, CA: 0.38 (0.05)					
Method	C_f	C_a	AD	SC	$E = SC/C_f$
KTS(0.15)	0.15 (0.00)	0.39 (0.05)	0.16 (0.03)	0.33 (0.17)	2.24 (1.11)
KTS(0.35)	0.35 (0.00)	0.92 (0.12)	0.43 (0.05)	0.61 (0.19)	1.75 (0.55)
KTS(0.55)	0.55 (0.00)	1.45 (0.20)	0.50 (0.09)	0.87 (0.04)	1.58 (0.07)
KTS(p^*)	0.24 (0.15)	0.59 (0.35)	0.26 (0.12)	0.55 (0.29)	**3.89** (3.33)
GiPo	0.24 (0.15)	0.59 (0.35)	0.23 (0.13)	0.70 (0.38)	2.59 (1.52)
Recipe: greekSalad, $n = 9$, CA: 0.31 (0.10)					
Method	C_f	C_a	AD	SC	$E = SC/C_f$
KTS(0.15)	0.15 (0.00)	0.53 (0.18)	0.29 (0.10)	0.48 (0.15)	3.24 (0.94)
KTS(0.35)	0.35 (0.01)	1.25 (0.41)	0.47 (0.11)	0.71 (0.15)	2.06 (0.42)
KTS(0.55)	0.55 (0.00)	1.97 (0.63)	0.51 (0.15)	0.92 (0.07)	1.68 (0.13)
KTS(p^*)	0.20 (0.21)	0.72 (0.77)	0.27 (0.17)	0.55 (0.24)	3.97 (1.77)
GiPo	0.20 (0.21)	0.73 (0.77)	0.20 (0.07)	0.78 (0.22)	**6.25** (2.83)
Recipe: pastaSalad, $n = 19$, CA: 0.33 (0.09)					
Method	C_f	C_a	AD	SC	$E = SC/C_f$
KTS(0.15)	0.15 (0.00)	0.52 (0.29)	0.22 (0.05)	0.57 (0.12)	3.78 (0.79)
KTS(0.35)	0.35 (0.00)	1.21 (0.69)	0.41 (0.09)	0.78 (0.11)	2.23 (0.31)
KTS(0.55)	0.55 (0.00)	1.90 (1.09)	0.47 (0.12)	0.86 (0.12)	1.56 (0.22)
KTS(p^*)	0.18 (0.17)	0.60 (0.60)	0.22 (0.10)	0.52 (0.19)	3.87 (1.52)
GiPo	0.18 (0.17)	0.60 (0.60)	0.15 (0.06)	0.76 (0.15)	**6.56** (3.76)
Recipe: bacon&eggs, $n = 13$, CA: 0.33 (0.09)					
Method	C_f	C_a	AD	SC	$E = SC/C_f$
KTS(0.15)	0.15 (0.00)	0.47 (0.11)	0.23 (0.06)	0.59 (0.16)	3.96 (1.07)
KTS(0.35)	0.35 (0.00)	1.11 (0.26)	0.42 (0.11)	0.76 (0.14)	2.19 (0.40)
KTS(0.55)	0.55 (0.00)	1.74 (0.40)	0.46 (0.09)	0.89 (0.12)	1.63 (0.23)
KTS(p^*)	0.26 (0.24)	0.83 (0.85)	0.26 (0.15)	0.59 (0.27)	3.91 (3.61)
GiPo	0.26 (0.24)	0.83 (0.85)	0.20 (0.09)	0.85 (0.17)	**6.26** (4.27)
Recipe: continentalBreakfast, $n = 11$, CA: 0.23 (0.07)					
Method	C_f	C_a	AD	SC	$E = SC/C_f$
KTS(0.15)	0.15 (0.00)	0.73 (0.29)	0.30 (0.09)	0.56 (0.13)	3.80 (0.91)
KTS(0.35)	0.35 (0.00)	1.72 (0.68)	0.40 (0.09)	0.82 (0.13)	2.36 (0.37)
KTS(0.55)	0.55 (0.00)	2.71 (1.07)	0.37 (0.11)	0.93 (0.07)	1.71 (0.13)
KTS(p^*)	0.30 (0.26)	1.36 (1.42)	0.31 (0.14)	0.63 (0.25)	3.82 (2.92)
GiPo	0.30 (0.26)	1.36 (1.42)	0.17 (0.06)	0.88 (0.19)	**6.53** (6.32)
Recipe: turkeySandwich, $n = 12$, CA: 0.11 (0.04)					
Method	C_f	C_a	AD	SC	$E = SC/C_f$
KTS(0.15)	0.15 (0.00)	1.48 (0.43)	0.23 (0.11)	0.56 (0.27)	3.80 (1.84)
KTS(0.35)	0.35 (0.00)	3.51 (1.01)	0.23 (0.10)	0.79 (0.19)	2.28 (0.56)
KTS(0.55)	0.55 (0.00)	5.52 (1.55)	0.19 (0.06)	0.96 (0.09)	1.75 (0.17)
KTS(p^*)	0.20 (0.19)	2.11 (2.32)	0.15 (0.14)	0.44 (0.40)	2.24 (2.36)
GiPo	0.21 (0.19)	2.15 (2.32)	0.12 (0.06)	0.82 (0.25)	**8.34** (6.69)
Recipe: cheeseBurger, $n = 10$, CA: 0.22 (0.05)					
Method	C_f	C_a	AD	SC	$E = SC/C_f$
KTS(0.15)	0.15 (0.00)	0.72 (0.22)	0.20 (0.10)	0.53 (0.31)	3.53 (2.08)
KTS(0.35)	0.35 (0.00)	1.69 (0.53)	0.34 (0.09)	0.78 (0.17)	2.23 (0.47)
KTS(0.55)	0.55 (0.00)	2.65 (0.83)	0.34 (0.09)	0.95 (0.10)	1.73 (0.18)
KTS(p^*)	0.13 (0.05)	0.59 (0.12)	0.21 (0.09)	0.48 (0.27)	3.86 (2.63)
GiPo	0.14 (0.05)	0.60 (0.11)	0.15 (0.05)	0.82 (0.17)	**7.10** (3.49)

to be purpose-aware for producing truly purpose-oriented summaries.

The behaviour of the algorithms in some random samples from two BEOID scenarios (Fig. 4) are in agreement with, and illustrate, the general trend discussed above. Being clueless about the purpose, DSF tends to select a few large video segments that just may overlap some action segments almost by chance (Fig. 4a). Guided by the purpose-specific learned scores, KTS is more focused towards where the actions happen, but still selects long video segments (Fig. 4b). Under the purpose-specific selection criteria, GiPo tends to select more and shorter video segments (Fig. 4c), usually covering the beginning of action parts, as desirable. Certainly, not being perfect, the GiPo approach misses some action steps, and occasionally mis-selects some non-action segment.

Table 7
Video classification (number of cases, out of 405, [and %]) for Purpose 2 on EGTEA-Gaze+. Random guess rate is 14.3% for $C = 7$ classes. − = misclassification, + = correct classification; f = full-video, s = summary.

Baseline			
f	s		
	−	+	
−	198 [48.9]	22 [5.4]	220 [54.3]
+	72 [17.8]	113 [27.9]	185 [45.7]
Total	270 [66.7]	135 [33.3]	405 [100]
GiPo			
f	s		
	−	+	
−	196 [48.4]	24 [5.9]	220 [54.3]
+	64 [15.8]	121 [29.9]	185 [45.7]
Total	260 [64.2]	145 [35.8]	405 [100]

3.5.2. Purpose 2

The summaries generated for this purpose on EGTEA-Gaze+ resulted to be about 70-frame long, which are on average about 4% of the original full videos.

Classification rates of the full videos (f), and the summaries (s), both with the baseline and with GiPo (Table 7), are (significantly) bigger than random guess, yet not great. These rates are about 10 percent points lower with the summaries than with the full videos. In addition, the recognition rates for the baseline and the GiPo summaries are similar. A likely interpretation of these results (poor full-video classification and strong baseline summary) can be that the BoW-based representation tend to favour global characterisations of the full sequences as opposed to sparse reduced ones, an aspect left for improvement as further work.

Regarding the performance metrics, it is interesting to observe (Table 8) that in the cases of correct classification for the summaries, i.e. (f:−, s:+) and (f:+, s:+), both metrics (TCR and T2MR) are better for the GiPo summaries than for the baselines summaries, as noticeable from the bold-faced numbers for the maximum values row-wise. This observation is confirmed by the averaged s/f ratios of the metrics ($T2MR_s/T2MR_f$ and TCR_s / TCR_f) computed instance-wise (right-hand side tables). These ratios are larger for GiPo than for the baseline, which suggests that despite the strong baseline summaries, the GiPo summaries provide earlier and higher confidence for discriminative purposes. Thus, in spite of the simplicity of the approach, it serves as an illustration of this second summarisation purpose.

Finally, the frames selected by GiPo and those selected by the baseline for two random sequences of two different scenarios, cheeseBurger and turkeySandwich, are compared. In the cheeseBurger case (Fig. 5-A), most selected frames by the GiPo method (A–I) relate to the burger preparation, and only one segment (I) out of the nine segments relate to some action/object ("open the fridge") that is not specific to this recipe, and might certainly be shared with some other recipes. However, the baseline method, by selecting uniformly-sampled segments (a–h) might select good segments by chance (a, d, e, f, h), but is more prone to select more non-discriminative parts (b, c, g). Regarding turkeySandwich (Fig. 5-B), many frames involving interactions with relevant objects of this recipe (lettuce, tomato, turkey, mustard/mayo) are selected (B, C, D, E, G, H), although some of these objects (lettuce, tomato) are shared among other recipes as well. Only a few segments selected (A and F) correspond to less clear or more generic objects. In contrast, most of the segments selected by the purpose-agnostic baseline (a, b, c, d, g) relate to generic, non-discriminative parts (fridge,

Table 8
Performance (mean and std. dev. for each metric) for Purpose 2 on EGTEA-Gaze+. − = misclassification, + = correct classification; f = full-video, s = summary.

TCR

Classif.		baseline		GIPO	
f	s	f	s	f	s
−	−	0.44 (0.29)	0.49 (0.28)	0.44 (0.30)	0.48 (0.27)
−	+	0.68 (0.28)	0.77 (0.17)	0.62 (0.22)	**0.87** (0.11)
+	−	0.48 (0.30)	0.72 (0.23)	0.49 (0.30)	0.66 (0.21)
+	+	0.72 (0.24)	0.88 (0.13)	0.70 (0.26)	**0.90** (0.10)

TCR_s/TCR_f

Classif.			
f	s	baseline	GIPO
−	−	1.75 (4.57)	1.72 (4.60)
−	+	1.43 (0.82)	**1.66** (0.82)
+	−	2.02 (1.78)	1.60 (1.15)
+	+	1.38 (0.63)	**1.63** (1.28)

T2MR

Classif.		baseline		GIPO	
f	s	f	s	f	s
−	−	0.55 (0.23)	0.57 (0.22)	0.56 (0.24)	0.55 (0.20)
−	+	0.74 (0.25)	0.80 (0.14)	0.67 (0.20)	**0.91** (0.10)
+	−	0.60 (0.23)	0.76 (0.19)	0.61 (0.22)	0.70 (0.18)
+	+	0.78 (0.21)	0.91 (0.12)	0.76 (0.22)	**0.93** (0.09)

$T2MR_s/T2MR_f$

Classif.			
f	s	baseline	GIPO
−	−	1.12 (0.43)	1.09 (0.42)
−	+	1.26 (0.60)	**1.48** (0.52)
+	−	1.40 (0.55)	1.27 (0.51)
+	+	1.27 (0.52)	**1.35** (0.55)

empty dishes, cupboard, kitchen bench), and some segments are recipe-related, by chance (e, f). Although there is certainly some room for improvement, these examples provides interesting insight into the effectiveness of the proposed approach and the potential utility of Purpose 2, and GIPO at large.

4. Discussion

The scoring-and-selecting paradigm has been shown to be suitable for purpose-specific summarisation. Thus, for a new summarisation purpose, frames are first scored according to how they roughly support this purpose; then, the selection strategy applies purpose-specific constraints and trade-offs. The evaluation metrics should then be aligned to capture how well the generated summaries fit the intended purpose. The two considered examples provide illustrative guidance and inspiration for applying this framework to other purposes. For instance, for the first purpose the frame relevance can be judged within a video independently to other videos, whereas the second purpose relates to (video-level) discriminative scenarios.

Further work can be directed at improving the performance of the considered purpose-oriented summarisation approaches, and at comparing and validating the results with user studies. GIPO-based summarisation challenges can be proposed as part of other egocentric datasets, mainly the largest ones such as EPIC-KITCHENS (Damen et al., 2020) and Ego4D.[2] Producing GIPO-based summaries from a set of videos where people perform similar but not exactly the same activities represent an interesting and challenging research possibility. Other efforts may address the exploration of other summarisation purposes and the feasibility of a more unified methodology that facilitates both defining user purposes, and automatically producing the corresponding summaries.

5. Conclusion

This work has addressed a largely unexplored area in video summarisation: summaries with a purpose. The problem and possible solutions have been illustrated on egocentric videos on two purposes grounded on meaningful end-user goals: reviewing performed steps, and supporting video-category browsing. Both purposes are formulated in terms of a scoring-and-selecting formalism, while the respective procedures are purpose-specific. Results provide evidence not only on that the proposed approaches are effective but also that this kind of ad-hoc purpose-specific solutions are clearly called for.

[2] To be introduced in the forthcoming *Ninth International Workshop on Egocentric Perception, Interaction and Computing: Introducing Ego4D — a Massive First-Person Dataset and Challenge*, https://eyewear-computing.org/EPIC_ICCV21.

CRediT authorship contribution statement

V. Javier Traver: Conceptualisation, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualisation, Project administration, Funding acquisition. **Dima Damen:** Methodology, Resources, Data curation, Writing – review & editing, Visualisation, Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Work supported by project UJI-B2018-44 from *Pla de promoció de la investigació de la Universitat Jaume I, Castelló, Spain*. The financial support for the research stay with code PRX18/00283, and for the research network with code RED2018-102511-T, both from *Ministerio de Ciencia, Innovación y Universidades*, are acknowledged. Part of the work was carried out as a research visitor of the first author at University of Bristol, during a short sabbatical leave from his institution, which is also acknowledged.

Appendix. Cycle detection and removal

Frame-level features $\mathbf{x}_t$ are clustered with k-means, and assigned to cluster c_t. Then a $k \times k$ transition matrix M is initialised to 0 and entries $M[\mathsf{c}_t, \mathsf{c}_{t+1}]$ incremented by one whenever the clusters of two consecutive frames differ, $\mathsf{c}_t \neq \mathsf{c}_{t+1}$. Positive entries (i, j) in the transition matrix, $M[i, j] > 0$, are used to start "walking" through potential cycles $p = (i, j)$, so that a third element k is added to get $p = (i, j, k)$ if $M[j, k] > 0$, then $p = (i, j, k, l)$ for some l such as $M[k, l] > 0$, and so on. Whenever the last item in the cycle p, say $l = p_{|p|}$, is found to be already in the current cycle ($l \in p_{1:|p|-1}$), this cycle p is saved into a set of found cycles, and its corresponding entries in the transition matrix are reset. If a cycle can no longer be "walked", it is also reset, as the entries in M are, and then another potential cycle is explored. This process is repeated until no more potential cycles are found.

The number of clusters k in k-means in this work is given by a manually-provided estimation of number of different situations for each scenario in BEOID. Further work would be aimed at making this simple cycle detector more robust and general.

To remove the repetitions in the generated summary α, the temporal segments $[t_1, t_2]$ corresponding to the found cycles are first identified, and then removed from the summary, i.e. $\alpha_{t_1:t_2} \leftarrow 0$. We keep the first (two) repetitions of a given cycle, and the number of remaining repetitions could be part of the actual video summary, e.g as superimposed text or graphical form.

References

Abdalla, K., Menezes, I., & Oliveira, L. (2019). Modelling perceptions on the evaluation of video summarization. *Expert Systems with Applications*, *131*, 254–265.

Agustí, P., Traver, V. J., & Pla, F. (2014). Bag-of-words with aggregated temporal pairwise word co-occurrence for human action recognition. *Pattern Recognition Letters*, *49*, 224–230.

Apostolidis, E. E., Adamantidou, E., Metsai, A. I., Mezaris, V., & Patras, I. (2021). Video summarization using deep neural networks: A survey. CoRR http://arxiv.org/abs/2101.06072.

Cai, S., Zuo, W., Davis, L. S., & Zhang, L. (2018). Weakly-supervised video summarization using variational encoder-decoder and web prior. In *ECCV* (pp. 193–210).

Chollet, F., et al. (2015). Keras implementation of Inception V3 model, with weights pre-trained on ImageNet https://keras.io/applications/#inceptionv3.

Chu, W., Song, Y., & Jaimes, A. (2015). Video co-summarization: Video summarization by visual co-occurrence. In *CVPR* (pp. 3584–3592).

Damen, D., Doughty, H., Farinella, G., Fidler, S., Furnari, A., Kazakos, E., et al. (2020). The EPIC-KITCHENS dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Damen, D., Leelasawassuk, T., Haines, O., Calway, A., & Mayol-Cuevas, W. W. (2014a). You-Do, I-Learn: Discovering task relevant objects and their modes of interaction from multi-user egocentric video. In *BMVC*.

Damen, D., Leelasawassuk, T., Haines, O., Wray, M., Moltisanti, D., Calway, A., et al. (2014b). Bristol egocentric object interactions dataset. http://people.cs.bris.ac.uk/~damen/BEOID/index.htm Last visit: 23rd. December 2019.

del Molino, A. G., Tan, C., Lim, J., & Tan, A. (2017). Summarization of egocentric videos: A comprehensive survey. *IEEE Transactions on Human-Machine Systems*, *47*(1), 65–76.

Doughty, H., Mayol-Cuevas, W. W., & Damen, D. (2019). The pros and cons: Rank-aware temporal attention for skill determination in long videos. In *CVPR* (pp. 7862–7871).

Gao, J., Yang, X., Zhang, Y., & Xu, C. (2020). Unsupervised video summarization via relation-aware assignment learning. *IEEE Transactions on Multimedia*.

Hahn, M. (2019). Tools for FPV. https://github.com/meera1hahn/ego_centric_vision Last access: 23rd December, 2019.

Hahn, M., Ruiz, N., Alayrac, J., Laptev, I., & Rehg, J. M. (2018). Learning to localize and align fine-grained actions to sparse instructions. CoRR http://arxiv.org/abs/1809.08381.

Huang, J.-H., Murn, L., Mrak, M., & Worring, M. (2021). GPT2MVS: Generative pre-trained transformer-2 for multi-modal video summarization. http://arxiv.org/abs/2104.12465.

Huang, J.-H., & Worring, M. (2020). Query-controllable video summarization. In *Proceedings of the international conference on multimedia retrieval* (pp. 242–250). ACM.

Hussain, T., Muhammad, K., Ding, W., Lloret, J., Baik, S. W., & de Albuquerque, V. H. C. (2021). A comprehensive survey of multi-view video summarization. *Pattern Recognition*, *109*.

Jang, Y., Sullivan, B., Ludwig, C., Gilchrist, I., Damen, D., & Mayol-Cuevas, W. (2019). EPIC-Tent: An egocentric video dataset for camping tent assembly. In *ICCV Workshops*.

Ji, Z., Zhao, Y., Pang, Y., Li, X., & Han, J. (2021). Deep attentive video summarization with distribution consistency learning. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(4), 1765–1775.

Jin, H., Song, Y., & Yatani, K. (2017). ElasticPlay: Interactive video summarization with dynamic time budgets. In *Proceedings of the ACM international conference on multimedia* (pp. 1164–1172).

Kanehira, A., Van Gool, L., Ushiku, Y., & Harada, T. (2018). Viewpoint-aware video summarization. In *CVPR*.

Kaushal, V., Iyer, R., Doctor, K., Sahoo, A., Dubal, P., Kothawade, S., et al. (2019). Demystifying multi-faceted video summarization: tradeoff between diversity, representation, coverage and importance. In *IEEE winter conference on applications of computer vision (WACV)* (pp. 452–461).

Kaushal, V., Kothawade, S., Tomar, A., Iyer, R. K., & Ramakrishnan, G. (2021). How good is a video summary? A new benchmarking dataset and evaluation framework towards realistic video summarization. CoRR http://arxiv.org/abs/2101.10514.

Kellerer, H., Pferschy, U., & Pisinger, D. (2004). *Knapsack Problems*. Springer.

Lee, Y. J., & Grauman, K. (2015). Predicting important objects for egocentric video summarization. *International Journal of Computer Vision*, *114*(1), 38–55.

Lee, S., Sung, J., Yu, Y., & Kim, G. (2018). A memory network approach for story-based temporal summarization of 360° videos. In *CVPR*.

Leiva, L. A., Traver, V. J., & Castelló, V. (2013). LiveThumbs: a visual aid for web page revisitation. In *ACM SIGCHI conference on human factors in computing systems (CHI)* (pp. 1797–1802).

Li, Y., Liu, M., & Rehg, J. M. (2018). In the eye of beholder: Joint learning of gaze and actions in first person video. In *ECCV* (pp. 639–655).

Li, Y., Liu, M., & Rehg, J. M. (2019). Georgia tech egocentric activity datasets. Last visit: 23rd. December 2019.

Liu, T., Meng, Q., Huang, J.-J., Vlontzos, A., Rueckert, D., & Kainz, B. (2021). Video summarization through reinforcement learning with a 3D spatio-temporal u-net. http://arxiv.org/abs/2106.10528.

Lu, Y., & Mayol-Cuevas, W. (2019). HIGS: hand interaction guidance system. In *2019 IEEE international symposium on mixed and augmented reality adjunct (ISMAR-Adjunct)* (pp. 376–381).

Ma, M., Mei, S., Wan, S., Hou, J., Wang, Z., & Feng, D. D. (2020). Video summarization via block sparse dictionary selection. *Neurocomputing*, *378*, 197–209.

Mahasseni, B., Lam, M., & Todorovic, S. (2017). Unsupervised video summarization with adversarial LSTM networks. In *CVPR*.

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.

Niebles, J. C., Wang, H., Wang, H., & Fei-Fei, L. (2006). Unsupervised learning of human action categories using spatial-temporal words. In *BMVC* (pp. 127.1–127.10).

Otani, M., Nakashima, Y., Rahtu, E., & Heikkila, J. (2019). Rethinking the evaluation of video summaries. In *CVPR*.

Otani, M., Nakashima, Y., Rahtu, E., Heikkilä, J., & Yokoya, N. (2016). Video summarization using deep semantic features. In *Asian conference on computer vision (ACCV)* (pp. 361–377).

Panda, R., Das, A., Wu, Z., Ernst, J., & Roy-Chowdhury, A. K. (2017). Weakly supervised summarization of web videos. In *ICCV* (pp. 3677–3686).

Panda, R., & Roy-Chowdhury, A. K. (2017). Collaborative summarization of topic-related videos. In *CVPR* (pp. 4274–4283).

Park, J., Lee, J., Kim, I., & Sohn, K. (2020). SumGraph: Video summarization via recursive graph modeling. In *ECCV* (pp. 647–663).

Paszke, A., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, R. Garnett (Eds.), *Advances in neural information processing systems (NeurIPS)* (pp. 8024–8035).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Potapov, D., Douze, M., Harchaoui, Z., & Schmid, C. (2014). Category-specific video summarization. In *ECCV* (pp. 540–555).

Ragusa, F., Furnari, A., Battiato, S., Signorello, G., & Farinella, G. M. (2019). Egocentric visitors localization in cultural sites. *Journal on Computing and Cultural Heritage*, *12*(2).

Rochan, M., & Wang, Y. (2019). Video summarization by learning from unpaired data. In *CVPR*.

Sak, H., Senior, A., & Beaufays, F. (2014). Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. http://arxiv.org/abs/1402.1128.

Sharghi, A., Gong, B., & Shah, M. (2016). Query-focused extractive video summarization. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *ECCV* (pp. 3–19).

Sharghi, A., Laurel, J. S., & Gong, B. (2017). Query-focused video summarization: Dataset, evaluation, and a memory network based approach. In *CVPR*.

Silva, M., Ramos, W., Ferreira, J. a., Chamone, F., Campos, M., & Nascimento, E. R. (2018). A weighted sparse sampling and smoothing frame transition approach for semantic fast-forward first-person videos. In *CVPR*.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *CVPR* (pp. 2818–2826).

Traver, V. J., Latorre-Carmona, P., Salvador-Balaguer, E., Pla, F., & Javidi, B. (2014). Human gesture recognition using three-dimensional integral imaging. *Journal of the Optical Society of America A*, *31*(10), 2312–2320.

Traver, V. J., Zorío, J., & Leiva, L. A. (2021). Glimpse: A gaze-based measure of temporal salience. *Sensors*, *21*(9).

Varini, P., Serra, G., & Cucchiara, R. (2017). Personalized egocentric video summarization of cultural tour on user preferences input. *IEEE Transactions on Multimedia*, *19*(12), 2832–2845.

Wei, H., Ni, B., Yan, Y., Yu, H., Yang, X., & Yao, C. (2018). Video summarization via semantic attended networks. In *AAAI* (pp. 216–223).

Wu, J., hua Zhong, S., & Liu, Y. (2020). Dynamic graph convolutional network for multi-video summarization. *Pattern Recognition*, *107*.

Xiao, S., Zhao, Z., Zhang, Z., Yan, X., & Yang, M. (2020). Convolutional hierarchical attention network for query-focused video summarization. In *AAAI conference on artificial intelligence* (pp. 12426–12433). AAAI Press.

Xiong, B., Kalantidis, Y., Ghadiyaram, D., & Grauman, K. (2019). Less is more: Learning highlight detection from video duration. In *CVPR* (pp. 1258–1267).

Xu, C., Gao, Z., Zhang, H., Li, S., & de Albuquerque, V. H. C. (2021). Video salient object detection using dual-stream spatiotemporal attention. *Applied Soft Computing*, *108*.

Yuan, L., Tay, F. E. H., Li, P., Zhou, L., & Feng, J. (2019). Cycle-SUM: Cycle-consistent adversarial LSTM networks for unsupervised video summarization. In *AAAI* (pp. 9143–9150).

Zhang, K., Grauman, K., & Sha, F. (2018). Retrospective encoders for video summarization. In *ECCV*.

Zhang, J., & Peng, Y. (2019). Object-aware aggregation with bidirectional temporal graph for video captioning. In *IEEE/CVF conference on computer vision and pattern recognition (CVPR)*.

Zhao, B., Li, X., & Lu, X. (2018). HSA-RNN: Hierarchical structure-adaptive RNN for video summarization. In *CVPR*.

Zhao, B., Li, X., & Lu, X. (2020). Property-constrained dual learning for video summarization. *IEEE Transactions on Neural Networks and Learning Systems*, *31*(10), 3989–4000.

Zhao, B., Li, X., & Lu, X. (2021a). TTH-RNN: Tensor-train hierarchical recurrent neural network for video summarization. *IEEE Transactions on Industrial Electronics*, *68*(4), 3629–3637.

Zhao, B., Li, H., Lu, X., & Li, X. (2021b). Reconstructive sequence-graph network for video summarization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhou, K., Qiao, Y., & Xiang, T. (2018a). Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In *Conf. on artificial intelligence (AAAI)* (pp. 7582–7589).

Zhou, K., Xiang, T., & Cavallaro, A. (2018b). Video summarisation by classification with deep reinforcement learning. In *BMVC*.

Zhu, W., Lu, J., Li, J., & Zhou, J. (2021). Dsnet: A flexible detect-to-summarize network for video summarization. *IEEE Transactions on Image Processing*, *30*, 948–962.