Journal of Visual Communication and Image Representation 52 (2018) 1–12

Contents lists available at ScienceDirect

Journal of Visual Communication and Image Representation

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

Personal-location-based temporal segmentation of egocentric videos for lifelogging applications☆

Antonino Furnari*, Sebastiano Battiato, Giovanni Maria Farinella

Department of Mathematics and Computer Science, University of Catania, Italy

ARTICLE INFO

Keywords:
Egocentric Vision
Lifelogging
Personal locations
Temporal segmentation

ABSTRACT

Temporal video segmentation is useful to exploit and organize long egocentric videos. Previous work has focused on general purpose methods designed to deal with data acquired by different users. In contrast, egocentric video tends to be very personal and meaningful for the specific user who acquires it. We propose a method to segment egocentric video according to the personal locations visited by the user. The method aims at providing a personalized output and allows the user to specify which locations he wants to keep track of. To account for negative locations (i.e., locations not specified by the user), we propose a negative rejection method which does not require any negative sample at training time. For the experiments, we collected a dataset of egocentric videos in 10 different personal locations, plus various negative ones. Results show that the method is accurate and compares favorably with the state of the art.

1. Introduction

Wearable devices allow people to acquire a huge quantity of data about their behavior and activities in an automatic and continuous fashion [1]. The practice of acquiring data of one's own life for a variety of purposes is commonly referred to as lifelogging. While the technology to acquire and store lifelog data coming from different sources is already available, the real potential of such data depends on our ability to make sense of it. Wearable cameras, in particular, can be used to easily acquire hours of egocentric videos concerning the activities we perform, the people we meet, and the environments in which we spend our time. As observed in [2], egocentric video is generally difficult to exploit due to the lack of explicit structure, e.g., in the form of scene cuts or video chapters. Moreover, according to the considered goal, long egocentric videos tend to contain much uninformative content like, for instance, transiting through a corridor, walking outdoors or driving to the office. Consequently, automated tools to enable easy access to the information contained in such videos are necessary.

Toward this direction, researchers have already investigated methods to produce short informative video summaries from long egocentric videos [3–5], recognize activities performed by the camera wearer [6–11], temporally segment the video according to detected ego-motion patterns [2,12], and segment egocentric photo-streams [13–15]. Past literature aimed at investigating general-purpose methods, which are generally trained and tested on data acquired by many users in order to ensure the generality of the algorithms. This approach, however, risks to overlook the subjective nature of egocentric video, which can be leveraged to provide tailored and user-specific services.

1.1. Personal locations

Towards the exploitation of user-specific information, in [16], we introduced the concept of *personal location* as:

> *a fixed, distinguishable spatial environment in which the user can perform one or more activities which may or may not be specific to the considered location.*

Personal locations are defined at the instance level (e.g., my office, the lab), rather than at the category level (e.g., an office, a lab) and hence they should not be confused with the general concept of visual scene [17]. Indeed, a given set of personal locations could include different instances of the same scene category (e.g., office vs lab office). Moreover, personal locations are user-specific since different users will be naturally interested in monitoring different personal locations (e.g., each user will be interest in monitoring the activities performed in his own office). Personal locations are constrained spaces (i.e., they are not defined as a whole room but rather refer to a part of it, e.g., the "office desk"), and hence they are naturally related to a restricted set of activities which can be performed in the considered locations [18]. For

☆ This paper has been recommended for acceptance by Cathal Gurrin.

* Corresponding author.

E-mail addresses: furnari@dmi.unict.it (A. Furnari), battiato@dmi.unict.it (S. Battiato), gfarinella@dmi.unict.it (G.M. Farinella).

https://doi.org/10.1016/j.jvcir.2018.01.019
Received 15 February 2017; Received in revised form 22 September 2017; Accepted 28 January 2018
Available online 31 January 2018

instance, the “office” personal location is naturally associated with office-related activities such as “writing e-mails” and “surfing the Internet”, while the “piano” personal location is generally related just to “playing piano”. Hence, being able to recognize when the user is located at a given personal location directly reveals information on a broad spectrum of activities which the user may be performing. The advantage of recognizing personal locations, rather than activities directly, is that providing supervision to recognize complex activities requires many samples (which is not practical for user-specific applications), while providing supervision to recognize personal locations is much more feasible, especially in egocentric settings [16].

1.2. Temporal segmentation of egocentric video

In this paper, we propose to segment egocentric videos into coherent segments related to personal locations specified by the user. We assume that the user selects a number of personal locations he wants to monitor and provides labeled training samples for them. The process of acquiring training data should not burden the user and be as simple as possible. Therefore, we adopt the acquisition protocol specified in [16,19]. According to this protocol, the user acquires training data for a specific location by turning on his wearable device and looking around briefly to acquire a 30-s video of the environment.

At test time, the system analyzes the egocentric video acquired by the user and segments it into coherent shots related to the specified personal locations. Given the large variability of visual content generally acquired by wearable devices, the user cannot easily provide an exhaustive set of personal locations he will visit. Therefore, the system should be able to correctly identify and reject all frames not related to any of the personal locations specified by the user. We will refer to these frames as “negatives” in the rest of the paper. In our context, negatives arise from two main sources: (1) the user moving from a personal location to another (*transition negatives*), and (2) the user spending time in a location which is not of interest (*negative locations*). Examples of transition negatives can be a corridor or an urban street, while examples of negatives locations might be a conference room, an office other than the user’s office, another car, etc. Please note that, while negative samples need to be correctly detected by the system, in real-world applications no negative training data can be provided by the user. Therefore, we design our method to learn solely from positive training data.

Fig. 1 shows a scheme of the proposed temporal segmentation system and illustrates three possible applications for it, which are discussed in the following. The output of the algorithm is a temporal segmentation of the input video. Each segment is associated to a label which identifies the related personal location or whether it is a negative segment (i.e., it is not related to any user-specified personal location). Such output can be used for different purposes. The most straightforward objective consists in producing a video index to help the user browse the video. This way, the user can easily jump to the part of the video he is more interested in and discard negative segments which may not be relevant. A second possible use of the output temporal segmentation consists in producing coherent video shots related to the personal locations specified by the user (e.g., of egocentric videos acquired over different days). Given the segmented shots and related meta-data (e.g., time stamps), the system could answer questions such as “show me what I was doing this morning when I first entered my office” or “tell me how many coffees I had today” (e.g., how many times I was at the Coffee Vending Machine personal location). Moreover, video shots can be used as a basis for egocentric video summarization [4,20]. A third use of the segmented video consists in estimating the time spent by the user at each location. In this case, the system would be able to answer questions such as “how much time did I spend driving this week?” or “how much time did I spend in my office today?”. This kind of estimate does not require accurate temporal segmentation but only overall correct per-frame predictions.

1.2.1. Contributions

This paper extends our previous work [21]. In particular, we present the proposed method in greater details and analyzes the impact of each component and related parameters more thoroughly. We extend the experimental analysis by defining a novel performance measure designed to evaluate segmentation accuracy from a shot-retrieval point of view. New comparisons with many state of the art methods are also introduced. Finally, we publicly release the code implementing the proposed method and evaluation measures.

The main contributions of this paper can be summarized as follows: (1) It is proposed to segment egocentric videos to highlight personal locations using minimal user-specified training data. To study the problem we collect and release a dataset comprising more than 2 h of labeled egocentric video covering 10 different locations plus various negatives. (2) A method to segment egocentric videos and reject negative samples is proposed. The method can be trained using only the available positive samples. (3) A measure to evaluate the accuracy of temporal video segmentation methods is defined. The measure penalizes methods which produce over-segmented or under-segmented results.

Experiments show that the proposed system can produce accurate segmentations of the input video with little supervision, outperforming baselines and existing approaches. The code related to this study, as well as the proposed dataset and a video of our demo, can be downloaded at http://iplab.dmi.unict.it/PersonalLocationSegmentation/.

The remainder of the paper is organized as follows. Section 2 summarizes the related work. Section 3 presents the proposed method. Section 4 introduces the involved dataset, defines the considered evaluation measures and reports the experimental settings. Results are discussed in Section 5, whereas Section 6 concludes the paper.

2. Related works

Location awareness. Our work is related to previous studies on context and location awareness in wearable and mobile computing. According to Dey et al. [22], context aware systems should be able to *“use context to provide relevant information and/or services to the user, where relevancy depends on the user’s task”*. Visual location awareness, in particular, has been investigated by different authors over the years. Starner et al. [23] addressed the recognition of basic tasks and locations related to the Patrol game from egocentric videos in order to assist the user during the game. Aoki et al. [24] proposed to recognize personal locations from egocentric video using the approaching trajectories observed by the wearable camera. Torralba et al. [25] designed a context-based vision system for place and scene recognition. Farinella et al. [26,27] engineered efficient computational methods for scene recognition which can be easily deployed to embedded devices. Rhinehart et al. [18] explored the relationship between actions and locations to improve both localization and action prediction. Furnari et al. [16] performed a benchmark of different wearable devices and image representations for personal location recognition.

Temporal video segmentation. Temporal video segmentation methods aim at decomposing an input video into a set of meaningful segments which can be used as basic elements for indexing [28]. The topic has been widely investigated in the domain of movie and broadcast video [29–33]. In particular, Hanjalic et al. [29] proposed to consider a video as composed by scenes and shots. Shots are elementary video units acquired without interruption by a single camera. Scenes contain semantically coherent material and are generally composed by different temporally contiguous shots. Most state of the art algorithms achieve temporal segmentation by first detecting shots and then merging contiguous highly correlated shots to form scenes. Chasanis et al. [30] propose to cluster shots according to their visual content and apply a sequence alignment algorithm to obtain the final segmentation. Sidiropoulos et al. [31] jointly exploit low-level and high-level audiovisual features within the Scene Transition Graph to obtain temporal