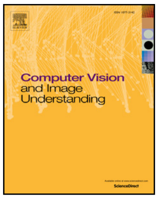
Computer Vision and Image Understanding 148 (2016) 181–192

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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



Characterizing everyday activities from visual lifelogs based on enhancing concept representation

Peng Wang[a,*], Lifeng Sun[a], Shiqiang Yang[a], Alan F. Smeaton[b], Cathal Gurrin[b]

[a] *National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China*
[b] *Insight Centre for Data Analytics, Dublin City University, Glasnevin, Dublin 9, Ireland*

ARTICLE INFO

Article history:
Received 14 April 2015
Accepted 26 September 2015

Keywords:
Lifelogging
Assistive living
SenseCam
Activity classification
Wearable camera

ABSTRACT

The proliferation of wearable visual recording devices such as SenseCam, Google Glass, etc. is creating opportunities for automatic analysis and usage of digitally-recorded everyday behavior, known as visual lifelogs. Such information can be recorded in order to identify human activities and build applications that support assistive living and enhance the human experience. Although the automatic detection of semantic concepts from images within a single, narrow, domain has now reached a usable performance level, in visual lifelogging a wide range of everyday concepts are captured by the imagery which vary enormously from one subject to another. This challenges the performance of automatic concept detection and the identification of human activities because visual lifelogs will have such variety of semantic concepts across individual subjects. In this paper, we characterize the everyday activities and behavior of subjects by applying a hidden conditional random field (HCRF) algorithm on an enhanced representation of semantic concepts appearing in visual lifelogs. This is carried out by first extracting latent features of concept occurrences based on weighted non-negative tensor factorization (WNTF) to exploit temporal patterns of concept occurrence. These results are then input to an HCRF-based model to provide an automatic annotation of activity sequences from a visual lifelog. Results for this are demonstrated in experiments to show the efficacy of our algorithm in improving the accuracy of characterizing everyday activities from individual lifelogs. The overall contribution is a demonstration that using images taken by wearable cameras we can capture and characterize everyday behavior with a level of accuracy that allows useful applications which measure, or change that behavior, to be developed.

1. Introduction

There is growing interest in creating large volumes of personal, first-person video or long duration image sequences, for lifelogging or quantified-self types of applications. These use wearable visual recording devices like Google Glass or Microsoft's SenseCam. *Visual lifelogging* is the term used to describe a class of personal sensing and digital recording of all of our everyday behavior which employs wearable cameras to capture image or video sequences of everyday activities. As the enabler for visual lifelogging, camera-enabled sensors are used in wearable devices to record still images [40] or video [3,13,32] taken from a first-person view, i.e. representing the subject's view of everyday activities. Visual lifelogging has already been widely applied in assistive living applications including aiding human memory recall, diet monitoring, chronic disease diagnosis, recording activities of daily living and so on. Example visual lifelogging projects include Steve Mann's WearCam [31,32], the DietSense project at UCLA [38], the WayMarkr project at New York University [4], the InSense system at MIT [3], and the IMMED system [33]. Microsoft Research catalysed research in this area with the development of the SenseCam [12,40] which was made available to other research groups in the late 2000s.

In terms of sensing devices, visual lifelogging can be categorized roughly into in-situ lifelogging and wearable lifelogging. In-situ lifelogging can be described simply as lifelogging in instrumented environments such as homes or workplaces. This means that human activities can be captured through video sensors installed in the local infrastructure [1]. Typical use of video sensors for in-situ lifelogging also includes works as reported in [15,17–19,48] and [16]. Jalal et al. [19] proposed a depth video-based activity recognition system for smart spaces based on feature transformation and HMM recognition. Similar technologies are applied in other work by the same authors in [15] and [18] which can recognize human activities from body depth silhouettes. In related work by Song et al. [48], depth data is utilized to represent the external surface of the human body. By proposing the body surface context features, human action recognition is robust to

[*] Corresponding author.
E-mail addresses: pwang@tsinghua.edu.cn (P. Wang), sunlf@tsinghua.edu.cn (L. Sun), yangshq@tsinghua.edu.cn (S. Yang), alan.smeaton@dcu.ie (A. F. Smeaton), cathal.gurrin@computing.dcu.ie (C. Gurrin).

http://dx.doi.org/10.1016/j.cviu.2015.09.014
1077-3142/© 2015 Elsevier Inc. All rights reserved.

Fig. 1. The Microsoft SenseCam as worn by subjects.

translations and rotations. As with Jalal's work in [16,17], Song's work [48] still depends on static scenes with an embedded sensing infrastructure. Current activity recognition in such settings usually assume there is only one actor in the scene and how these solutions can scale up to more realistic and challenging settings such as outdoors are difficult.

To alleviate such challenges, we focus on activity recognition within non-instrumented environments using wearable visual sensing. In wearable lifelogging, the sensing devices are portable and worn directly by the subjects and can include head-mounted cameras in works by Hori and Aizawa [13] and Mann et al. [32] or cameras mounted on the front of chests in works by Blum et al. [3] and by Sellen et al. [40]. In [14], the key issues and main challenges in generating wearable diaries and lifelogging systems are discussed. In [11], other sensors such as accelerometers, GPS, image and audio are recorded using a smartphone and applied in an application based on annotating daily activities. Though effective to a limited extent, a direct mapping from low-level features like colors and textures to semantic labels lacks flexibility in characterizing the semantics of activities such as understanding occurrences of scenes, objects, etc. in images. Recent work in [47] has also highlighted the same problem.

As a new form of multimedia, the effective management of large visual lifelogs requires semantic indexing and retrieval, for which we can use the preliminary work already done in other domains. State-of-the-art techniques for image/video analysis use statistical approaches to map low-level image features like shapes and colors to high-level semantic concepts like "indoor", "dog" or "walk". According to the TRECVid benchmark [43], acceptable detection results have been achieved, particularly for concepts for which there exists enough annotated training data. Introducing automatic detection of semantic concepts from visual lifelogs enables searching through those lifelogs based on their content and this is particularly useful in characterizing everyday life patterns [5,8]. However, because of the wide variety of activities that people usually engage in and the differences in those activities from person to person, a very wide range of semantic concepts can appear in visual lifelogs, which increases the challenges in developing automatic concept detectors from which we can detect everyday activities. Moreover, due to subjects' movements as lifelog images are captured, even images captured passively within the same lifelogged event may have significant visual differences. This poses burdens on the characterization of activities based on the detected concepts, especially in applications where the detection of everyday behavior is to be done in near real-time.

The SenseCam, shown in Fig. 1, is a sensor-augmented wearable camera designed to capture a digital record of the wearer's day by recording a series of images and a log of sensor data. It captures the view of the wearer from a fisheye lens and pictures are taken at the rate of about one every 50 s without the trigger of other sensors. The on-board sensors for measuring ambient light levels, movement, and the presence of other people through a passive infra-red sensor, are also used to trigger additional capture of pictures when sudden changes are detected in the environment of the wearer as well as to prevent images being captured when the wearer, and the SenseCam, are being moved which would result in blurring of images. SenseCam has been shown to be effective in supporting recall of memory from the past for individuals [40,42], as well as having applications in diet monitoring [35], activity detection [54], sports training [35], etc. Due to its advantages of multiple sensing capabilities, light weight and unobtrusive logging with a long battery life, we employ SenseCam as a wearable device to log details of subjects' everyday lives.

Temporal patterns of concept occurrence can characterize image sequences, but at a higher level. Consider the "cooking" activity, where visual concepts like "fridge", "microwave", "oven" often occur in sequence and frequently interact with the concept of "hands". For example, "opening fridge" is typically observed before "starting microwave". Such patterns can also be regarded as temporal semantics of concepts. To deal with such concept temporal semantics, the major contributions of this paper can be highlighted as: first, we proposed a time-aware concept detection enhancement algorithm based on weighted non-negative tensor factorization (WNTF) for which a multiplicative solution is derived. The effectiveness of this factorization method is also proven. The second contribution is an everyday activity characterization based on hidden conditional random fields (HCRF), proposed by merging time-varying dynamics of concept attributes.

The rest of the paper is organized as follows: in Section 2 we present related work on concept detection and event processing used to drive a characterization of everyday activities. An overview of our proposed solution is presented in Section 3. In Section 4, we describe tensor factorization approaches to tackle the concept enhancement problem at a frame-level of concept indexing. This is followed by an HCRF-based algorithm to combine concept semantics from a frame level for higher-level activity characterization in Section 5. The experimental implementation and analysis of our results are presented in Section 6. Finally, we close the paper with conclusions and pointers to future work.

2. Related work

2.1. Automatic cncept detection and enhancement

Compared to low-level features like color, texture, shape, etc. which in their raw form do not convey much meaning, semantic