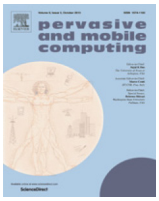
Contents lists available at ScienceDirect

Pervasive and Mobile Computing

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



Unsupervised understanding of location and illumination changes in egocentric videos

Alejandro Betancourt[a,b,*], Natalia Díaz-Rodríguez[c], Emilia Barakova[b], Lucio Marcenaro[a], Matthias Rauterberg[b], Carlo Regazzoni[a]

[a] Department of Engineering (DITEN), University of Genova, Genova, Italy
[b] Department of Industrial Design, Eindhoven University of Technology, Eindhoven, Netherlands
[c] Computer Science Department, University of California Santa Cruz, CA, USA

ARTICLE INFO

Article history:
Received 12 April 2016
Received in revised form 27 March 2017
Accepted 31 March 2017
Available online xxxx

Keywords:
Machine learning
Unsupervised learning
Egocentric videos
First person vision
Wearable camera

ABSTRACT

Wearable cameras stand out as one of the most promising devices for the upcoming years, and as a consequence, the demand of computer algorithms to automatically understand the videos recorded with them is increasing quickly. An automatic understanding of these videos is not an easy task, and its mobile nature implies important challenges to be faced, such as the changing light conditions and the unrestricted locations recorded. This paper proposes an unsupervised strategy based on global features and manifold learning to endow wearable cameras with contextual information regarding the light conditions and the location captured. Results show that non-linear manifold methods can capture contextual patterns from global features without compromising large computational resources. The proposed strategy is used, as an application case, as a switching mechanism to improve the hand-detection problem in egocentric videos.

1. Introduction

The emergence of wearable video devices such as action cameras, smart glasses and low-temporal life-logging cameras has detonated a recent trend in computer science known as First Person Vision (FPV) or Egovision. The 90's idea of a wearable device with autonomous processing capabilities is nowadays possible and is considered one of the most relevant technological trends of the recent years [1]. The ubiquitous and personal nature of these devices opens the door to critical applications such as Activity Recognition [2,3], User–Machine Interaction [4], Ambient Assisting Living [5–7] Augmented Memory [8,9] and Blind Navigation [10], among others.

One of the key features of wearable cameras is their capability to move across different locations and record exactly what the user is looking at. This is an unrestricted video perspective that requires existent methods to perform good in the unknown number of locations and the changing light conditions implied by this video perspective. A common way to deal with this problem is to predefine a particular application or location and bound the algorithms based on this. This is the case of gesture recognition for virtual museums proposed in [4] or the activity recognition methods based on the kitchen dataset [5,11]. Another way to alleviate the large number of recorded locations is by using exhaustive video labeling of the

[*] Corresponding author.
E-mail addresses: alejandro.betancourt@tpv-tech.com (A. Betancourt), ndiaz@decsai.ugr.es (N. Díaz-Rodríguez), e.i.barakova@tue.nl (E. Barakova), lucio.marcenaro@unige.it (L. Marcenaro), g.w.m.Rauterberg@tue.nl (M. Rauterberg), carlo@dibe.unige.it (C. Regazzoni).

http://dx.doi.org/10.1016/j.pmcj.2017.03.016
1574-1192/© 2017 Elsevier B.V. All rights reserved.

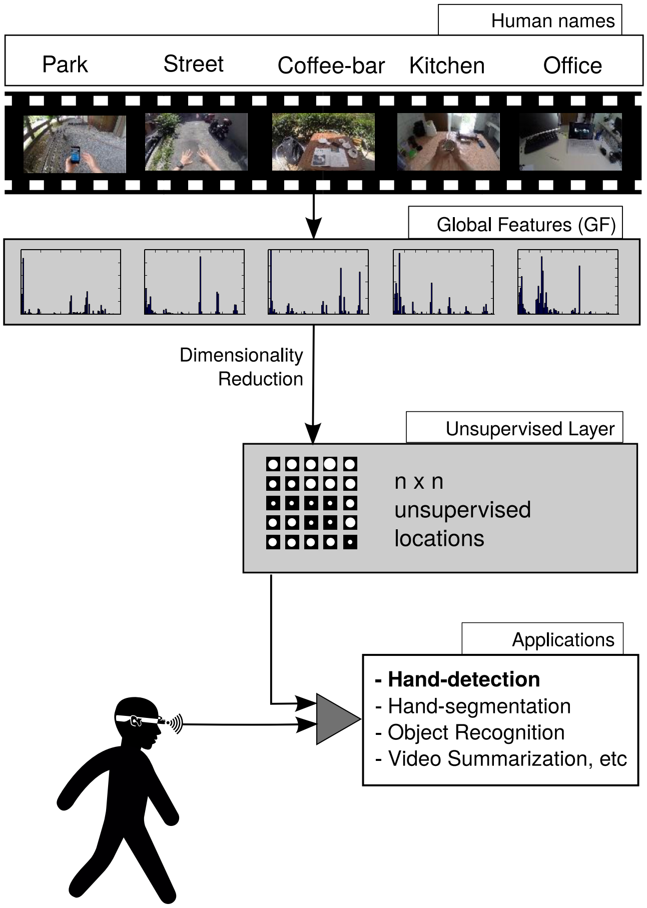


Fig. 1. Unsupervised strategy to extract contextual information about light and location using global features.

recorded locations and objects as is done in [6] to detect daily activities. The authors in [12] use global histograms of color to reduce the effect of light changes in a color-based hand-segmenter.

The approach of [12] shows that contextual information, such as light conditions, are valuable sources of information that can be used to improve the performance and applicability of current FPV methods. This idea is also applicable to other FPV related functionalities such as activity recognition, on which a device that can understand user's location can easily reduce the number of possible activities and take more accurate decisions. Pervasive computing refers to the devices that can modify their behavior based on contextual variables as context-aware devices [13], and its benefits are widely explored for example in assisted living [14] and anomaly detection [15].

This paper is motivated by the potential impact of contextual information, such as light conditions and location, on different FPV methods. The strategy presented, is a first step toward our envision of a device that can understand the environment of the user and modify its behavior accordingly. The proposed approach understands the contextual information on which the user is involved as a set of different characteristics that can point to previously recorded conditions, and not as a scene classification problem based on manual labels assigned to particular locations (e.g., kitchen, office, street). In this way, this study devises an unsupervised procedure for wearable cameras to switch between different models or search spaces according to the light conditions or location on which the user is involved. Fig. 1 summarizes our approach.

From Fig. 1 it is clear that the transition from the global features to the unsupervised layer can be seen as a dimensional reduction from the global feature space (high dimensional space) to a simplified low dimensional space (intrinsic dimension). The latter provides an unsupervised location map to be used later to switch between different behaviors at different hierarchical levels. These dimensional reductions are known as manifold methods, and their capabilities to capture complex patterns are defined by their algorithmic and/or theoretic formulation [16].

Regarding the global features to be used, relevant information can be obtained from recent advances in FPV [1] and scene recognition [17,18]. Given the restricted computational resources of wearable devices, we use computationally efficient features such as color histograms and GIST descriptors. However, the proposed approach can be extended with more complex data such as deep features [15]. In that case three important issues must be considered: (i) the computational cost will restrict the applicability in wearable devices; (ii) it will require large amounts of training videos and manual labels; (iii) the use of existent "pre-trained" neural architectures compromises the unsupervised nature of our approach.