Computer Vision and Image Understanding 136 (2015) 128–145

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Hierarchical temporal graphical model for head pose estimation and subsequent attribute classification in real-world videos

Meltem Demirkus *, Doina Precup, James J. Clark, Tal Arbel

Centre for Intelligent Machines, McGill University, Montreal, Quebec H3A 2A7, Canada

ARTICLE INFO

Article history:
Received 16 May 2014
Accepted 11 March 2015

Keywords:
Face
Attribute
Pose
Probabilistic
Real-world
Unconstrained
Video
Hierarchical
Graphical
Temporal

ABSTRACT

Recently, head pose estimation in real-world environments has been receiving attention in the computer vision community due to its applicability to a wide range of contexts. However, this task still remains as an open problem because of the challenges presented by real-world environments. The focus of most of the approaches to this problem has been on estimation from single images or video frames, without leveraging the temporal information available in the entire video sequence. Other approaches frame the problem in terms of classification into a set of very coarse pose bins. In this paper, we propose a hierarchical graphical model that probabilistically estimates continuous head pose angles from real-world videos, by leveraging the temporal pose information over frames. The proposed graphical model is a general framework, which is able to use any type of feature and can be adapted to any facial classification task. Furthermore, the framework outputs the entire pose distribution for a given video frame. This permits robust temporal probabilistic fusion of pose information over the video sequence, and also probabilistically embedding the head pose information into other inference tasks. Experiments on large, real-world video sequences reveal that our approach significantly outperforms alternative state-of-the-art pose estimation methods. The proposed framework is also evaluated on gender and facial hair estimation. By incorporating pose information into the proposed hierarchical temporal graphical mode, superior results are achieved for attribute classification tasks.

1. Introduction

The amount of real-world video sequences has increased substantially as the cost of the cameras has decreased in recent years, leading to a considerable increase in the range of possible real-world video applications, including video indexing, human tracking, face recognition and verification, social networking and human computer interaction. Despite the huge literature on optimizing and automating such applications, due to the challenges presented by real-world environments (see Fig. 1), it is still hard to develop these applications. Recently there has been a lot of effort in the computer vision community, to further improve these systems in the context of real-world scenarios [1–22]. Some of this effort is to make face recognition/verification, facial attribute classification and human computer interaction to benefit from head pose estimates as prior information in order to boost their performance [1,23–34]. For instance, for face verification/facial attribute classification tasks, facial fiducial points are used to estimate pose, so as to be able to map face images from the real-world environment to a common coordinate system [24,35,36,32].

Head pose estimation methods based on 2D images can be divided into several main groups, namely geometric, tracking, appearance template and manifold subspace embedding methods [38]. *Geometric methods* [39] use the facial landmark locations to estimate the head pose from their relative configuration. *Tracking methods* [40–42] use the relative movement between consecutive video frames to estimate the global head movement. *Appearance template methods* [43,44,28] use image-based comparison techniques to match a test image to a set of training images with corresponding pose labels. [28], for instance, estimates the 3D head pose from high quality depth images using random regression forests. It achieves this through regression between depth features and continuous head pose angles. However, this approach uses every image region for the head pose estimation, thus it is not robust to real-world facial occlusion, such as from coffee cups and facial hair, which can introduce additional spurious features to the face representation. *Manifold, subspace embedding methods*

* Corresponding author.
E-mail addresses: demirkus@cim.mcgill.ca (M. Demirkus), dprecup@cs.mcgill.ca (D. Precup), clark@cim.mcgill.ca (J.J. Clark), arbel@cim.mcgill.ca (T. Arbel).
URLs: http://www.cim.mcgill.ca/~demirkus/ (M. Demirkus), http://www.cs.mcgill.ca/~dprecup/ (D. Precup), http://www.cim.mcgill.ca/~clark/ (J.J. Clark), http://www.cim.mcgill.ca/~arbel/ (T. Arbel).

http://dx.doi.org/10.1016/j.cviu.2015.03.005
1077-3142/

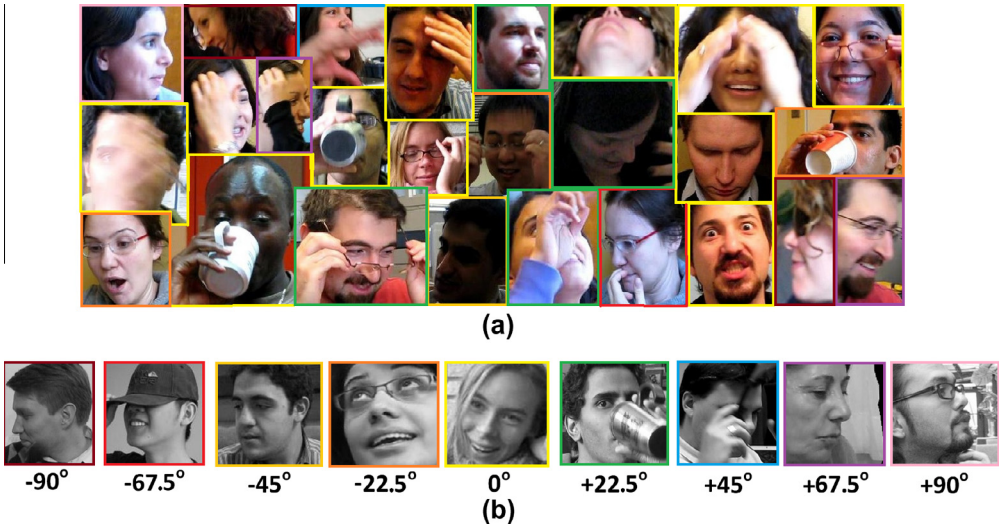

Fig. 1. Sample tracked face images from McGill Real-world Face Video Database [37]: (a) Examples of challenges exhibited in real-world environment, such as a wide variability in illumination conditions and background clutter, arbitrary head poses and scales, arbitrary partial occlusions etc. (b) Head pose (yaw angle) ground truth labels of sample face images provided by the probabilistic labeling strategy in [37]. Image courtesy of [32].

[45,44,46–51], on the other hand, use linear and nonlinear subspace techniques to project an image onto the head pose manifold, which is learned from a training set. When such techniques are used for video frames, they implicitly model a given video sequence temporally by mapping similar frames onto nearby locations in the manifold. The highest accuracies published in the head pose literature are provided by the manifold learning methods (see for example [51]). However, most of the aforementioned approaches are developed for constrained environments, and have several stages which are not compatible with real-world unconstrained environments (e.g. no extreme head pose or major occlusion is allowed). For instance, they often assume that the entire set of facial features typical for near frontal poses is always visible. Facial features are often manually labeled in the testing data, rather than automatically extracted. Most approaches are trained and tested on images which do not exhibit wide appearance variation. The testing databases mostly contain images with solid or constant background, limited range of facial expressions, no random illumination, and limited or no facial occlusion (e.g. CMU Multi-PIE [52], FERET [53] and CAS-PEAL [54]). Finally, the current tracking methods rely on video sequences with a known initial head pose, and typically, they must be reinitialized (at times, manually) whenever the tracking fails (due to a failure in the face detection or due to occlusion). However, many of these requirements and assumptions are not practical in the context of real-world videos.

Different methods have been proposed to solve the problem of head pose estimation in real-world environments, such as [27,30,33,23,55,29,32]. [27,30] treat the problem as a classification problem (assigning a face image to one of very coarsely defined pose bins), and address it in the context of single, low resolution video frames of crowded scenes under poor and/or limited (e.g. indoor) lighting. Approaches such as [23,29] use relatively high quality video frames/images and perform classification on finer pose bins. Some other approaches, on the other hand, define the pose estimation problem as a continuous pose angle estimation task [55,32,33]. Refs. [55,56] shows that when facial landmarks are located on the face, they can be successfully used to estimate head pose. However, such approaches have the vulnerability that it is difficult to extract such features when a major facial occlusion is present (see Fig. 2) or when the pose angle is more than 45°, leading to occlusion of facial landmark regions (e.g. eyes) in the image. Finally, most of the aforementioned approaches either focus on only a specific set of features (e.g. facial landmark points) to represent faces, or do not leverage the temporal pose information available between consecutive video frames.

The problem addressed in this paper is the automatic head pose (yaw angle) estimation in real-world videos that are affected by the joint occurrence of arbitrary face scales, extreme head poses, non-uniform illumination conditions, partial occlusions, motion blur, background clutter, wide variability in image quality, and subject variability (see Figs. 1 and 2). We propose a novel hierarchical, temporal graphical model, which uses a number of complementary robust local invariant facial features, and leverages the dependencies between consecutive video frames, in order to substantially improve head pose estimation in real world scenarios. Furthermore, at each frame, the system assesses the probability density function over the pose angles, ranging from −90° to +90° (Fig. 2). The proposed hierarchical and temporal graphical model (Figs. 4 and 3) uses spatial codebook representations inferred from different local features, which have a high degree of invariance to various transforms, such as changes in scale, viewpoint, rotation and translation. The local feature detectors and descriptors are chosen such that they extract complementary information from the tracked face image (see Fig. 2): (i) facial edge points inferred from eyebrows, mouth etc. (Geometric Blur features [57]) (ii) facial anatomical regions extracted from eyes, forehead, cheeks etc. (Boundary Preserving Local Region features [58]) (iii) densely sampled patch-based features over the whole face image (SIFT and CSIFT [59,60]), and (iv) facial landmarks [55,56]. Some of the reasons for employing multiple features include: (i) to ensure that if one type of feature is not detected from a face, the other(s) can compensate for it, in order to robustly estimate head pose, and (ii) to complement each other. To calculate the pose distribution for each feature type, the codebook statistics are employed. These statistics are used in a graphical model to estimate the single video frame pose probability distribution. Next, the framework temporally models these head pose probabilities over a video sequence. Finally, the method performs non-parametric density estimation to obtain continuous head pose probabilities.

The proposed hierarchical temporal graphical model is a general framework, which can use any type of feature and can be adapted to any face related inference task while providing the