Image and Vision Computing 32 (2014) 707–727

Contents lists available at ScienceDirect

Image and Vision Computing

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

# Online learning and fusion of orientation appearance models for robust rigid object tracking ☆

Ioannis Marras [a,*], Georgios Tzimiropoulos [a,b], Stefanos Zafeiriou [a], Maja Pantic [a,c]

[a] Imperial College London, Department of Computing, SW7 2AZ, UK
[b] University of Lincoln, School of Computer Science, Lincoln LN6 7TS, UK
[c] University of Twente, Department of Computer Science, Enschede 7522 NB, The Netherlands

ARTICLE INFO

*Article history:*
Received 5 June 2013
Received in revised form 14 February 2014
Accepted 5 April 2014
Available online 20 May 2014

*Keywords:*
Rigid object tracking
Fusion of orientation appearance models
Subspace learning
Online learning
Face analysis
RGB-D

ABSTRACT

We introduce a robust framework for learning and fusing of orientation appearance models based on both texture and depth information for rigid object tracking. Our framework fuses data obtained from a standard visual camera and dense depth maps obtained by low-cost consumer depth cameras such as the Kinect. To combine these two completely different modalities, we propose to use features that do not depend on the data representation: angles. More specifically, our framework combines image gradient orientations as extracted from intensity images with the directions of surface normals computed from dense depth fields. We propose to capture the correlations between the obtained orientation appearance models using a fusion approach motivated by the original Active Appearance Models (AAMs). To incorporate these features in a learning framework, we use a robust kernel based on the Euler representation of angles which does not require off-line training, and can be efficiently implemented online. The robustness of learning from orientation appearance models is presented both theoretically and experimentally in this work. This kernel enables us to cope with gross measurement errors, missing data as well as other typical problems such as illumination changes and occlusions. By combining the proposed models with a particle filter, the proposed framework was used for performing 2D plus 3D rigid object tracking, achieving robust performance in very difficult tracking scenarios including extreme pose variations.



## 1. Introduction

Depth or range cameras have been developed for several years and are available to researchers for certain applications for about a decade. With the development of 3D capturing equipment, it has become faster and easier to obtain 3D shape and 2D texture information to represent a real 3D object in a scene. Subspace learning techniques have been widely used for fusing the two modalities. These techniques have provided valuable tools for understanding and capturing the intrinsic non-linear structure of visual data encountered in many important machine vision problems. At the same time, there has been a substantially increasing interest in related applications such as appearance-based object recognition and rigid object tracking. A fundamental problem of the majority of subspace learning techniques (both linear and non-linear) for appearance-based object representation is that they are not robust. In this paper, we are extending existed subspace techniques in a robust framework for learning and fusing of orientation appearance models based on both texture and depth information.

Outliers are common not only because of illumination changes, occlusions or cast shadows but also because the depth measurements provided by an depth camera could be very noisy and the obtained depth maps usually contain "holes" or missing parts. It should be mentioned here that there are several cases that should be considered as "partial occlusions", like extreme facial expressions, or when a hand/object covers partially the tracked object. Fig. 1 depicts a few examples for all these common cases by using Kinect for obtaining both texture and depth information. In contrary to the texture information, there are cases of partially object occlusions where there are no significant differences in the depth information when a hand/object touches an object, as the raw Kinect depth data is of low resolution with high noise levels. Furthermore, as it is shown in Fig. 1, there are many cases when the estimation of the nose tip in the 3D space is very possible to fail, such as cases of extreme facial expressions or when a hand covers partially the object by touching it. This is a problem for all methods for both

☆ This paper has been recommended for acceptance by Lijun Yin.

* Corresponding author at: Imperial College, Department of Computing, London, U.K.
*E-mail addresses:* i.marras@imperial.ac.uk (I. Marras), gt204@imperial.ac.uk (G. Tzimiropoulos), s.zafeiriou@imperial.ac.uk (S. Zafeiriou), m.pantic@imperial.ac.uk (M. Pantic).

http://dx.doi.org/10.1016/j.imavis.2014.04.017
0262-8856/

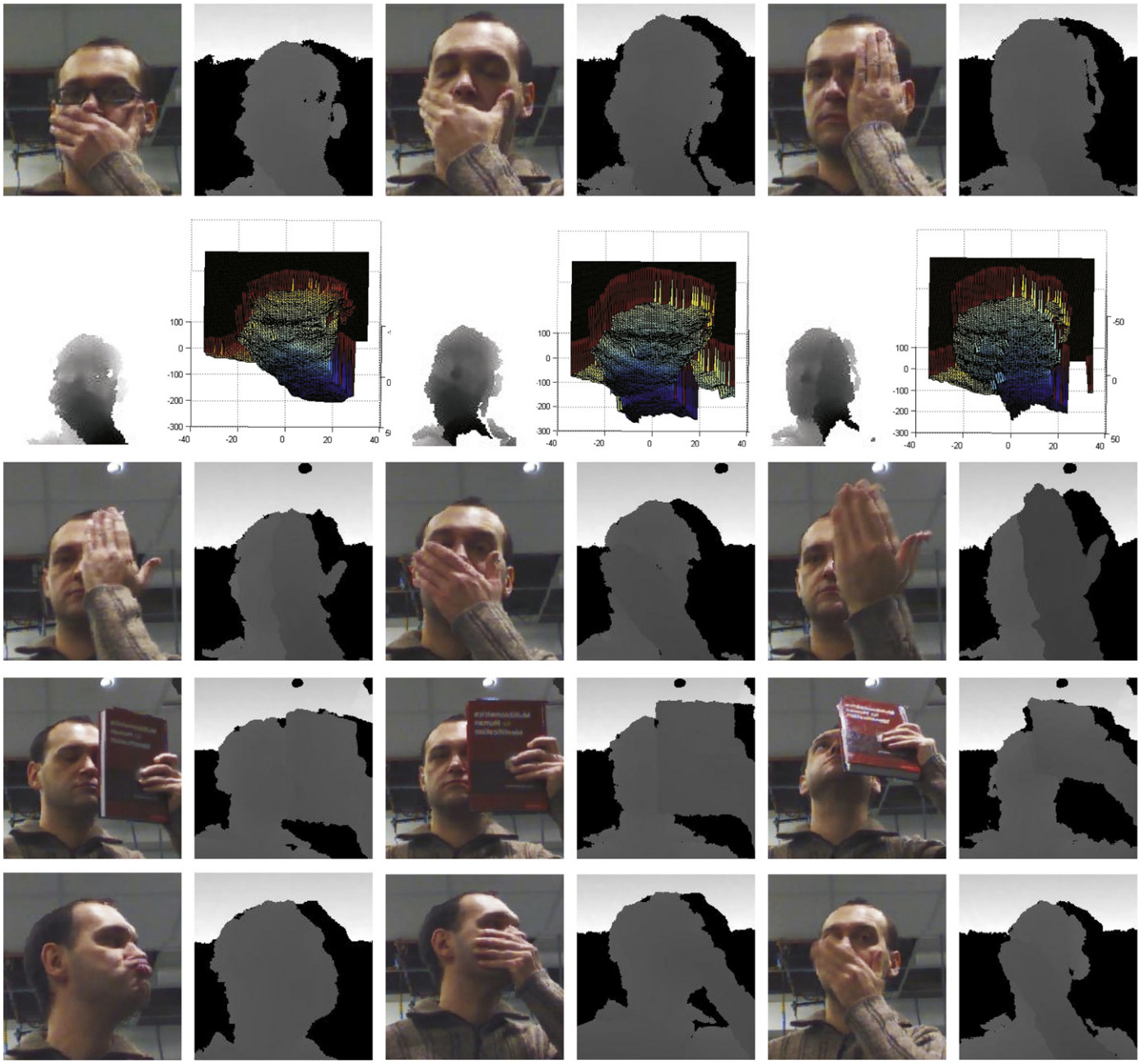

**Fig. 1.** Common examples of partial occluded faces, in both cropped texture and depth information by using Kinect. First row: a hand is touching the face. In contrary to the texture information, there are cases of partially object occlusions where there are no significant differences in the depth information when a hand/object touches the tracked object, as the raw Kinect depth data is of low resolution with high noise levels, Second row: depth information only for the face regions that are depicted in the first row, as well as their triangulated meshes without any mesh filtering, Third row: a hand, which is far from the face, is covering a face part. In this case, there are missing face parts inside the parallelepiped containing the face, Fourth row: in general an object is covering a face part, and Fifth row: cases where the nose tip estimation in the 3D space is not performing well, thus making a 3D tracking/pose estimation procedure, which is based on this estimation, to fail.

3D tracking/pose estimation and 3D face recognition such as [1,2] that are based on accurate nose tip estimation. In [2], the use of random regression forests for real time head pose estimation from high quality range scans, is introduced. Note that subspace learning for visual tracking requires robustness, efficiency and online adaptation. This combined problem has been rarely studied in literature. For example, in [3], the subspace is efficiently learned online using incremental $\ell_2^2$ norm PCA [4]. Nevertheless, the $\ell_2^2$ norm enjoys optimality properties only when image noise is independent and identically distributed (i.i.d.) Gaussian; for data corrupted by outliers, the estimated subspace can be arbitrarily skewed. On the other hand, robust reformulations of PCA [5–7] typically cannot be extended for efficient online learning.

3D shape information can be used to produce algorithms which are able to handle many challenges such as inaccurate face alignment, pose variations, measurement noise, missing data, facial expressions and partial occlusion. Many different approaches were proposed for dealing with the aforementioned problems [8,9,1,10–12]. Early approaches, such as [1], use specific face regions that are not affected by the presence of facial deformations caused by facial expressions, such as the nose and the area around it. Subspace learning algorithms for 3D mesh normals, such as Principal Component Analysis (PCA), employ low-dimensional representation of surfaces [13–16]. In its simplest form, PCA on surface normals has been applied on the concatenation of normal coordinates [14]. One attempt to exploit the special structure of normals (i.e., that lie on a sphere) was conducted in [15].