Pattern Recognition 47 (2014) 2783–2793

Contents lists available at ScienceDirect

Pattern Recognition

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

Feature fusion for facial landmark detection

Panagiotis Perakis [a,b,*], Theoharis Theoharis [b,a], Ioannis A. Kakadiaris [c]

[a] Computer Graphics Laboratory, Department of Informatics and Telecommunications, University of Athens, Ilisia 15784, Greece
[b] Department of Computer and Information Science, NTNU, 7491 Trondheim, Norway
[c] Computational Biomedicine Lab, Department of Computer Science, University of Houston, TX 77204, USA

ARTICLE INFO

Article history:
Received 10 March 2013
Received in revised form
18 September 2013
Accepted 8 March 2014
Available online 20 March 2014

Keywords:
Facial landmarks
Feature extraction
Feature fusion
Landmark detection

ABSTRACT

Facial landmark detection is a crucial first step in facial analysis for biometrics and numerous other applications. However, it has proved to be a very challenging task due to the numerous sources of variation in 2D and 3D facial data. Although landmark detection based on descriptors of the 2D and 3D appearance of the face has been extensively studied, the fusion of such feature descriptors is a relatively under-studied issue. In this paper, a novel generalized framework for combining facial feature descriptors is presented, and several feature fusion schemes are proposed and evaluated. The proposed framework maps each feature into a similarity score and combines the individual similarity scores into a resultant score, used to select the optimal solution for a queried landmark. The evaluation of the proposed fusion schemes for facial landmark detection clearly indicates that a quadratic distance to similarity mapping in conjunction with a root mean square rule for similarity fusion achieves the best performance in accuracy, efficiency, robustness and monotonicity.

1. Introduction

Facial landmark detection is a crucial first step in facial analysis for biometrics and numerous other applications. However, it has proved to be a very challenging task due to the numerous sources of variation in 2D and 3D facial data. These variations can be environment-based (illumination conditions and occlusions), subject-based (pose and expression variations) and acquisition-based (image scale, distortion, noise, spikes and holes). Both 2D and 3D facial landmark detection suffer from occlusion and expression variations. In addition, 2D facial landmark detection suffers from pose and illumination variations.

2D and 3D facial landmark detectors have to possess the properties of robustness to data variations, repeatability and distinctiveness. To fulfill these properties and constrain the detection process, landmark detectors use trained landmark classifiers or 2D/3D appearance landmark models/templates and 2D/3D geometry models for global topological consistency. 2D landmark detectors use view-based 2D geometry and appearance models or 3D geometry models. 3D landmark detectors use solely 3D geometry and 3D appearance models. Fused 2D/3D landmark detection methods use 3D geometry and 2D+3D appearance models. 2D and 3D landmark detection is based mostly on variations of the seminal work on Active Appearance Models of Cootes et al. [1–4]. Fused 2D/3D landmark detection is presented in Boehnen and Russ [5], Jahanbin et al. [6], Lu and Jain [7], Passalis et al. [8] and Perakis et al. [9,10].

Although many 2D/3D descriptors of facial features are used in the literature, a crucial issue has not been answered yet. How can these facial features be fused together in order to exploit their individual strengths and create a robust and accurate landmark detector?

Different feature descriptors can have complementary strengths and weaknesses, so combining them can increase system *accuracy*, *efficiency* and *robustness*, featuring *monotonicity*. Accuracy can be increased by exploiting data content from multiple sources (3D/2D) or the strengths of different data descriptors. In addition, using multiple descriptors can improve efficiency by limiting the landmarks' likelihood area. Finally, fusion can increase system robustness by limiting deficiencies inherent in using a single descriptor. For example, a corner/edge detector is very sensitive in illumination variations, but the shape index is not. Thus, using multiple descriptors is a form of uncertainty reduction, since one descriptor may pick up what the other misses.

A landmark detector has four important levels (Fig. 1). At the *acquisition level* a sensor acquires the facial data. At the *feature extraction level* the data are transformed into features that represent the landmark classes. At the *matching score level* the extracted features are compared with feature templates that represent each landmark class in order to detect candidate landmarks with an

* Corresponding author.
E-mail addresses: p.perakis@di.uoa.gr (P. Perakis), theotheo@di.uoa.gr (T. Theoharis), ikakadia@central.uh.edu (I.A. Kakadiaris).

http://dx.doi.org/10.1016/j.patcog.2014.03.007
0031-3203/

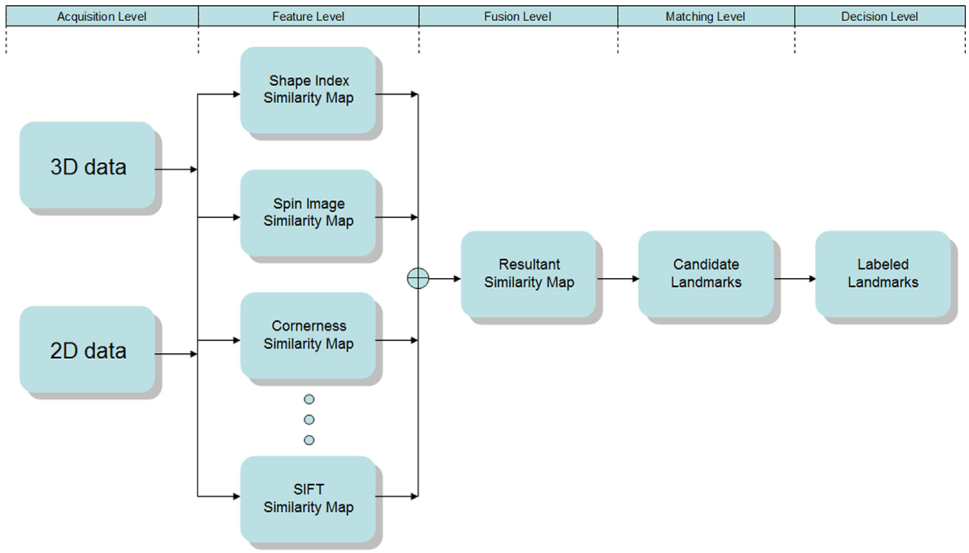

Fig. 1. Pipeline of feature fusion procedure for landmark detection.

associated matching score. Finally, at the *decision level* the matching scores (or ranks) are used to select a candidate landmark as the optimal solution for the queried landmark class.

Fusion can be applied at the acquisition or feature extraction level (pre-classification fusion) and at the matching score or decision level (post-classification fusion) [11,12]. Fusion at the matching score level can be viewed in two distinct ways. In the first, fusion is approached as a *classification* problem, while in the second, it is approached as a *combination* problem [11,13]. In the classification approach, a composite feature vector (by weighted concatenation) is constructed using the values of the fused features, which is further classified by a composite classifier (e.g., Neural Network, K-NN, Decision Trees, SVM). In the combination approach, the matching scores of the fused features are combined to generate a single resultant feature score which is used for the final decision. The common characteristic of all combination techniques is that the individual feature classifiers are separately trained and the combination relies on simple fixed rules [13]. These rules are the *sum rule*, *product rule*, *max rule*, *min rule*, *median rule* and *majority voting* [14]. The various schemes for combining classifiers can be grouped into three main categories according to their architecture: (i) *parallel*, (ii) *cascading* (serial), and (iii) *hierarchical* (tree-like) [15].

For landmark detection, although the construction of a composite feature classifier might be a potential solution, the combination method can be more easily applied to features whose values can be mapped to images, is more transparent (having also the strength of visualization), and possesses all the other fundamental properties required by a fusion scheme [16].

Feature fusion techniques have been proposed in the past (see Section 2), but in an entirely different context, that of multimodal biometrics or that of abstract feature fusion. The problem that is investigated in this paper is the behavior of fusion schemes under the strict context of landmark detection on facial datasets, which is an entirely different problem, since fusion techniques for landmark detection have to be also "locally consistent", which means that they have to boost results on a constrained area on facial surfaces. This problem has not yet been investigated.

This paper provides a novel generalized framework of fusion methods and their application to landmark detection and comes as an extension to our previous work for landmark detection [10]. The proposed framework fills a gap in existing research, which is dominated by methods that use single landmark descriptors of 3D or 2D appearance of the face, without combining them (see Section 2). The fusion scheme proposed acts after the "feature extraction level", transforms features to similarities and then combines them to generate a resultant feature similarity, which is considered as the matching score, and is used at the "matching level" for the detection of the queried landmarks (Fig. 1). The proposed approach of feature fusion offers significant dimensionality reduction and is easily extendable by adding new feature-components in feature space and changing the resultant similarity appropriately. This approach works equally well for any feature extracted either from 3D or 2D facial data. The only prerequisite is the availability of a common (u,v) parameterization so that the 3D and 2D data can be combined at the "acquisition level".

The rest of this paper is organized as follows: Section 2 presents related work in the field, Section 3 details the theoretical background of the proposed method, Section 4 presents its application to the detection of facial landmarks, Section 5 presents our results, and Section 6 summarizes our method.

2. Related work

A number of studies showing the advantages of information fusion in pattern recognition and especially in multimodal biometrics have appeared in the literature.

Xu et al. [12] grouped different fusion methods into categories and proposed methods for classifier fusion at different levels (measurement, rank and abstract) for recognizing handwritten numerals. They reported a significant improvement over the performance of individual classifiers.

Kittler et al. [14] have developed a theoretical framework for the combination approach to fusion at the matching score level of multimodal biometric applications. In their approach the matching scores of individual classifiers are interpreted as posterior probabilities and the resultant scores are the outcome of simple fixed fusion rules (sum rule, product rule, max rule, min rule, median rule and majority voting). They have experimented with face and voice biometrics and found that the sum rule outperformed the others.

Jain et al. [15] conducted experiments concerning the characteristics of combining twelve different classifiers using five different combination rules and six different feature sets generated from handwritten numerals (0–9). Reported results show that