Journal of Computational Science xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Journal of Computational Science

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

An approach for behavior analysis using correlation spectral embedding method

Deepak Kumar Jain[a,*], Neha Jain[b], Shishir Kumar[b], Amit Kumar[b], Raj Kumar[c], Haoxiang Wang[d,e]

[a] Institute of Automation, Chinese Academy of Sciences, Beijing, China
[b] Department of Computer Science &Engineering, Jaypee University of Engineering and Technology, Guna, India
[c] SGTBIM&IT, Affiliated to GGSIPU, Delhi, India
[d] Department of ECE, Cornell University, NY, USA
[e] R&D Center, GoPerception Laboratory, NY, USA

ARTICLE INFO

Article history:
Received 3 November 2016
Received in revised form 17 June 2017
Accepted 12 July 2017
Available online xxx

Keywords:
Correlation
Encoding
Embedding approach

ABSTRACT

Automatic identification of various facial movements and expressions with high recognition value is important for human computer interaction as the facial behavior of a human can be treated as an important factor for information representation as well as communication. A high deviation of human appearance and existence of noisy contextual background makes the human pose analysis is hard to achieve. A number of basic factors such as cluttered background, occlusion, and camera movement and illumination variations degrade the image quality resulting in poor performance for identifying different facial expressions. Moreover, the identification of the automatic feature detection in facial behavior requires high degree of correlation between the training and test images. Our proposed work tries to address the mentioned problems and resolve to some extent. In this methodology, a Decision-based Spectral Embedding approach combining appearance and geometry based features for head pose estimation and facial expression recognition by minimizing the objective function which leads to selection of optimal set of fiducial points. The method preserves the local information from different facial views for mapping neighboring input to its corresponding output, resulting in low dimensional representation for encoding the relationships of the data. The proposed methodology is validated with benchmark datasets for analyzing the performance of recognition of facial behavior.

1. Introduction

Human action recognition has effortlessly gained a lot attention of the researchers due to its inevitable contribution in video surveillance, sports video analysis,human computer interaction, video analysis and many more. The application of depth cameras for capturing the human-object interaction has given a new dimension for better analysis. The temporal relationship among different actions performed by human to accomplish a task requires continuous monitoring of intra-class variations which is challenging task. Tracking human actions kinematically is another important task. The challenge is to differentiate the poses and appearance that appears to be similar while performing different tasks. Another challenge involves where active appearance is sensitive to illumination and view.

Head pose estimation is an effortless orientation of human gesture for means of communication. The different orientation of head can represent different kind of information. A number of gestures are formed representing different messages for example quick head movements can be treated as a sign of surprise or alarm, visual attention of communication among group of people can be interpreted. Speech recognition is one of the major and vast areas of head pose estimation that includes identification of the messages from the people who are impaired of speaking and listening. Such applications may neutralize the gap between the common people to the handicapped one. Moreover, this approach can be applied for vehicular safety measure so that accidents occurrence due to unawareness of the vehicle pilot can be rectified by alarming uncertain situations. Head pose estimation establishes a relation between the vehicle movement and gaze of the driver. The

* Corresponding author at: Institute of Automation, Chinese Academy of Sciences, Beijing, China.
E-mail address: deepak@cripac.ia.ac.cn (D.K. Jain).

http://dx.doi.org/10.1016/j.jocs.2017.07.011
1877-7503/

applications alarms in the situation where the relation among the movement of the vehicle and the driver gaze is not accurate.

2. Literature survey

A numerous research work has been carried out in the field of action recognition and head pose estimation. A few models have been discussed below along with their advantages and disadvantages.

The Slow Feature Analysis (SFA) [1] is an unsupervised technique was proposed by Jones and Viola for estimating the slow variation of features. The position and identity of person variates over time continuously, this variation is a trace optimization problem with orthogonality constraints is obtained from discrete time input signal and low dimensional output signal as a linear transformation of a non-linear expansion. SFA is identified by solving an eigen value problem for the joint diagonalization of the data covariance matrix. As a result, the data neighborhood structures are represented using their temporal variations. In SFA, the action recorded for different temporal phases are: Neutral, when the face is relaxed, Onset, when the action initiates, Apex, when the muscles reach the peak intensity and Offset when the muscles begin to relax. The action finally ends with Neutral. The algorithm identifies linear projections to extract the common slowest varying features of two or more sequences and Expectation Maximization (EM) algorithm is hybridized along with this approach to perform inference in a probabilistic formulation of SFA and similarly extend it in order to handle two and more time varying data sequences.

In 2010, Wang et al. [2] developed the system to recognize the human action. The system initially performs the detection of person from each frame with the sliding window based on the Histogram of Oriented Gradients (HOG) over time. The tracking of the person is combined in the bounding-box over temporal continuity of the same person. The similarity among the bounding box of the person on each frame is computed and grouped to detect the human parts. This similarity integrates the information of location, scale and variation of time in each frame. For human pose estimation connections between the parts form tree structure. To reduce the searching space, the part detectors are used to detect possible parts. The part detector assigns part to prototypes that are created by clustering large number parts extracted from detected pose of the training data. The histogram of the part is used as the descriptor computed with Euclidean distance for classifying the upper parts and lower parts. However, the method is dependent on the number of cluster prototype. The more will be the cluster better representation of the parts can be done. But this limits to a range that the increase in the cluster beyond the limit will imprecise of part detection that will disturb the action recognition.

Most of the methods capture the location and size of the action region from a video frame which can be unimportant for unknown classification. The Sparse Representation (SR) [3] method takes the advantage of the information about location and size of the action region by classifying the same with the dictionary constructed from the training videos. The SR method performs classification of human actions consisting of two split dictionaries namely, action-region dictionary and context region dictionary. The former characterizes human action region in non-segmented test video while the later characterizes non-human action region in test video. The procedure involves finding the sparse representation by estimating the contextual information required for classification. The resulting contextual information is the concentration of sparse coefficient that helps for recognition of human action. As the concentration of sparse coefficients associated with the context region is higher than the concentration of sparse coefficients associated with the action region, the sparse methodology utilizes more information from the context region resulting in better classification accuracy.

The local feature tracking encounters a problem in low resolution conditions. Such methods require large sets of training data to store the image of eye with respect to the person that variate over time [4]. The specific appearance, head pose, scale, illumination and eyelids movements involve training data results. The geometric gaze model handles the head pose and gaze directions in a unified framework. The model computes the likelihood of the eye in the training phase that is maximized by gaze annotated in training samples. The entire procedure is done in following manner

1) The 3D head pose is achieved by fitting, frame-by frame and applying 3D mesh template to depth data with iterative closest points, resulting in 3D head rotation and translation for each frame.
2) The obtained frame is transformed into textured 3D mesh followed by the inverse head pose parameter on the 3D data surface and the crop eye images from the frontal looking facial texture, resulting in pose-rectified eye images.
3) The gaze direction is estimated from the resulting pose-rectified image of eye with the geometric transformation of the input gaze image.

For recognition of human posture, head pose estimation is a key issue as it conveys significant information regarding behavior and intentions of a person from its 3D structural image data provided by depth cameras [5]. The proposed methodology performs in the following phases: At the initialization stage, a reference depth image of a person is acquired. During runtime, the methodology looks for the 6-dimensional pose space in order to obtain a resulting pose in which the appeared head is identical to the reference view. The exploration procedure is formulated as a Particle Swarm optimization (PSO) approach whose objective function quantifies the divergence of the depth measurements between the hypothesized views to the reference view.

The erroneous estimation occurs due to absence of person from the scene. In such frames, the obtained score of the objective function is below the defined threshold value resulting in inaccurate measurement of the head pose estimation. To overcome this problem, a valid pose is applied in order to determine the center of the search range for all subsequent frames until a valid pose estimate is computed again.

The methodology proposed by Fanelli et al. [6] solves the classification and regression problems by extending the regression forests approach through discriminating depth patches which belong to head i.e. classification and utilizes the patches to predict the pose i.e. regression [6]. The method learns a mapping between depth features and real-valued parameters such as 3D head position and rotation angles. The methodology employs random regression forests in order to estimate the 3D head pose in real dataset from high quality depth data and make robustness to occlusions and poor signal-to-noise ratio. The problem is resolved as regression by estimating the head pose from depth data. The random forest need to be trained on labeled data and accuracy depends on amount of training, data obtaining becomes primary issue. The problem of training the data can be resolved by randomly generating number of examples.

The random forest consists of trained trees that are better than the decision trees. With this method the randomness is introduced in training process. The process is an ensemble process of classification and regression known as Discriminative Random Regression Forest (DRRF). The estimation of 3D position of head and different orientation can be obtained from low quality images with minimum cost sensors. The process involves the labeling of the orientation and location of the head construction of the training data.

Please cite this article in press as: D.K. Jain, et al., An approach for behavior analysis using correlation spectral embedding method, J. Comput. Sci. (2017), http://dx.doi.org/10.1016/j.jocs.2017.07.011