Neurocomputing 280 (2018) 23–31

Contents lists available at ScienceDirect

Neurocomputing

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

Pose recognition using convolutional neural networks on omni-directional images

S.V. Georgakopoulos [a], K. Kottari [a], K. Delibasis [a], V.P. Plagianakos [a], I. Maglogiannis [b,*]

[a] Department of Computer Science and Biomedical Informatics, University of Thessaly, Lamia, Greece
[b] Department of Digital Systems, University of Piraeus, Piraeus, Greece

ARTICLE INFO

Article history:
Received 21 February 2017
Revised 1 July 2017
Accepted 16 August 2017
Available online 21 November 2017

Keywords:
Computer vision
Convolutional neural networks (CNNs)
Zernike moments
Geodesic correction
Omni-directional image
Fisheye camera calibration
Pose classification
Transfer learning

ABSTRACT

Convolutional neural networks (CNNs) are used frequently in several computer vision applications. In this work, we present a methodology for pose classification of binary human silhouettes using CNNs, enhanced with image features based on Zernike moments, which are modified for fisheye images. The training set consists of synthetic images that are generated from three-dimensional (3D) human models, using the calibration model of an omni-directional camera (fisheye). Testing is performed using real images, also acquired by omni-directional cameras. Here, we employ our previously proposed geodesically corrected Zernike moments (GZMI) and confirm their merit as stand-alone descriptors of calibrated fisheye images. Subsequently, we explore the efficiency of transfer learning from the previously trained model with synthetically generated silhouettes, to the problem of real pose classification, by continuing the training of the already trained network, using a few frames of annotated real silhouettes. Furthermore, we propose an enhanced architecture that combines the calculated GZMI features of each image with the features generated at CNNs' last convolutional layer, both feeding the first hidden layer of the traditional neural network that exists at the end of the CNN. Testing is performed using synthetically generated silhouettes as well as real ones. Results show that the proposed enhancement of CNN architecture, combined with transfer learning improves pose classification accuracy for both the synthetic and the real silhouette images.

1. Introduction

Several computer vision and Artificial Intelligence applications require classification of segmented objects in digital images and videos. The use of object descriptors is a conventional approach for object recognition though a variety of classifiers. Recently, many reports have been published supporting the ability of automatic feature extraction by convolutional neural networks (CNNs) that achieve high classification accuracy in many generic cases, without the need for user-defined features. This approach is often referred to as deep learning. More specifically, CNNs have been suggested as state of the art methodology for pattern recognition [1], object localization [2], object classification in large-scale databases containing real world images [3], and malignancy detection on medical images [4–6]. The success of the CNNs on the aforementioned fields relies on their inner ability to exploit the local translational invariance of signal classes over their domain [7], allowing the transformation of long range interaction in terms of shorter localized interactions.

CNNs are trainable multistage architectures [8]. Each stage may be one of three types of layers, or an arbitrary combination of them: convolution, pooling or classic neural network layer. The last is commonly refereed as fully connected layer. The trainable components of a convolutional layer are mapped as a batch of kernels and perform the convolution operator on the previous layer's output. The pooling layer performs a subsampling to its input. The most common pooling function is the max pooling, which takes the maximum value of the local neighbourhoods. Finally, the fully connected layer can be treated as a special case of kernel with size 1×1. To train this network, the Stochastic Gradient Descent is usually utilized with the usage of mini-batches [9]. A drawback of the CNNs is the requirement of long training times, due to the amount of trainable parameters. However, the inherent parallelism of their structure, allows the utilization of Graphics Processing Units (GPUs) for performing the training phase [3].

The trained CNNs construct proper local features to discriminate between different image classes. This is a convenient property that removes the burden of extracting user-defined image features, while also often outperforms traditional classifiers based on the user provided features. However, in certain problems the statisti-

* Corresponding author.
E-mail addresses: spirosgeorg@dib.uth.gr (S.V. Georgakopoulos), kkottari@uth.gr (K. Kottari), kdelimpasis@dib.uth.gr (K. Delibasis), vpp@dib.uth.gr (V.P. Plagianakos), imaglo@unipi.gr (I. Maglogiannis).

https://doi.org/10.1016/j.neucom.2017.08.071
0925-2312/

cal properties of local stationarity and multi-scale compositional structure may not hold. Consequently, the performance of CNNs may deteriorate when compared to global image descriptors. This was observed in our previous work [10], where the Zernike image descriptors, custom designed using the calibration of the acquiring fisheye camera, outperformed CNNs in human poses recognition (in a limited number of real videos), while GZMI exhibited slightly inferior performance when applied to the synthetic data.

A limited number of the previously described works utilize CNNs with binary images, while, the combination of CNN features and other explicitly defined external features (at the feed-forward neural network (FNN) layer) is not extensively explored in literature. In this work, we present an enhancement of the well-established CNNs' architecture with the utilization of geodetically corrected Zernike moments (GZMI) [11] for human pose recognition in binary images. Synthetic human silhouettes were used to produce the extensive dataset required for the training of the CNN. The silhouettes were generated by rendering 3D human models through a calibrated omni-directional (fisheye) camera. The testing of CNNs is performed on a subset of the synthetically generated silhouettes, as well as on automatically segmented, manually labelled real videos of humans performing similar poses. Furthermore, transfer learning from training with synthetic to real data has been implemented by continuing the training of the CNNs with few frames of real human silhouettes, using the weights of synthetically trained CNNs as initial weights. Comparisons between CNNs, GZMI, CNNs enhanced with GZMI, with and without transfer learning are provided for synthetic and real data.

The rest of the paper is structured as follows: in Section 2 related work and background information are provided, while in Section 3 the proposed methodology is described. The corresponding experimental results are reported in Section 4 and the conclusions in Section 5.

2. Related work and background information

Two basic algorithmic classes span the literature for the problem of human pose estimation. The algorithms belonging to the first class rely on the leveraging of images descriptors, such as histogram oriented gradient (HoG) [12], SHIFT [13], Zernike [11,14–16] in order to extract features and subsequently constructing a model for classification. In [17], Poppe used the HoG descriptor on human figure images for pose estimation, after background subtraction and shadow suppression. The second algorithmic class is based on model fitting processes [18,19]. Only recently, methods that use deep neural networks have emerged for the problem of pose estimation. Their main disadvantage is the requirement of additional information, such as joint nodes, depth image, etc. In [20,21] the use of CNN as a regressor for rough joint locator hand-annotations of the Frame Labelled In Cinema (FLIC) and BBC TV sign language broadcast datasets is proposed. In [22] an approach for 3D human pose estimation is presented, using 2D pose information. The proposed methodology expands the classic CNNs architecture at loss function stage to multi-loss functions, in order to predict the joints of the 3D human and estimate the pose. Hand-annotated root node is required to estimate the rest joint nodes. Other recent implementations of CNN present combination of information that has been produced by independent convolution operators into the hidden layer stage. In [23], an approach of training CNNs' architecture that consists of two independent sequential convolutional stages interpreted as feature extractors is presented. The two convolutional stages are fed from parts of the same input image and outputs are multiplied using outer product at each location of the image and pooled to obtain an image descriptor. In [24], an RGB image and its optical flow are combined in the first and second sequential convolutional stage, respectively. This method is intended to utilize both spatial and temporal image information. The idea of spatio-temporal information was also expanded in classic CNNs, in an attempt to capture the motion information in 3D, encoded in multiple adjacent frames, for activity recognition [25].

A very popular technique for CNNs training is the transfer of knowledge from one problem named A to another related problem named B. This methodology is also known as transfer learning (TL). Due to the internal structure of CNNs, their first layer features tend to learn features that resemble either Gabor filters or colour blobs [26]. This observation leads to using the first n number of layers, of a network pre-trained to a specific task, to initialize a CNN for fine-tuning the solution of a different task. TL is very useful in cases where the dataset is not adequate to train a full network (small dataset, missing values, difficulties on annotations, etc.). Such pre-trained networks have been applied on ImageNet dataset [27,28].

In our previously published works [10,11], we suggested a modification of the classic Zernike moments adjusted to omni-directional camera. For this reason, we replaced the Euclidian distance metric by the geodesic distance metric for fisheye lens. The human poses' classification accuracy, using GZMI on synthetic and real human poses, was higher compared to other image descriptors [18]. In [10], CNNs have achieved better classification results than GZMI on synthetic human poses, but lower performance at human poses in real video. This is probably caused by poor segmentation quality, which resulted in fragmented silhouettes. Thus, we may conclude that the GZMI, which are applied globally to the part of the image containing the silhouette, have tolerance on segmentation error compared to the CNNs, which extract structures locally from the images. Considering the above, in this work we extend the information that the CNNs construct by exploring the applicability of the combination of CNN with GZMI features. Furthermore, we investigate the usefulness of transfer learning between CNN-training with synthetic and real silhouettes.

3. Methodology

3.1. Overview of the method

The main goal of this work is the enhancement of the popular CNN method with the GZMI descriptors to recognize different poses of binary human silhouettes from indoor images acquired by a roof-based omni-directional camera. Fisheye cameras are dioptric omni-directional cameras, increasingly used in computer vision applications [29,30], due to their 180° field of view (FoV). In [31–33] the calibration of fisheye camera is reported to emulate the strong deformation introduced by the fisheye lens. In [34] a methodology for correcting the distortions induced by the fisheye lens is presented.

An extensive synthetic dataset of binary silhouettes was constructed for training the CNN. The generation of synthetic poses is shown in Fig. 1(a): 3D human models of selected poses [35–37] were placed in different positions and at different rotations round the Z-axis in the real world room. The calibration of the camera was employed to render the synthetic binary frames. In the case of real videos, the foreground was segmented and the binary silhouette was cropped and resized. This process is explained in Fig. 1(b). The employed architecture of the CNN is shown in Fig. 1(c). The CNN consists of a series of convolutional/pooling layers, followed by a traditional feed-forward neural network (FNN). The GZMI descriptors were fused into the FNN layer of the CNN architecture to increase the discriminating power of the system. Furthermore, a small number of real videos were acquired, automatically segmented and manually annotated. The real videos were