International Conference on Computer Vision and Graphics ICCVG 2022

https://iccvg.sggw.edu.pl

Monday, 19 September 2022 - Wednesday, 21 September 2022

Warsaw University of Life Sciences – SGGW

Book of Abstracts

Important remark

This booklet contains abstracts as they were submitted by the Authors in the preliminary round of abstract submission, so the contents may be outdated. These abstracts are intended only as an aid during the conference, before the conference proceedings are published.

The Proceedings will be published in Lecture Notes on Networks and Systems series, Springer.

The abstracts are sorted alphabetically according to titles.

Contents

A Deep Multi-Layer Perceptron Model for Automatic Colourisation of Digital Grayscale Images

Authors: Wande Shokunbi[1]; Joseph Akinyemi[1]; Olufade Onifade[1]

[1] *University of Ibadan, Nigeria*

Corresponding Author: wandeshokunbi@outlook.com

Colour images tend to be more visually appealing to humans compared to grayscale images, as colour images are closer in representation to the natural way we perceive our environment. While obtaining grayscale images from colour images is relatively trivial, the reverse process is not. More so, colourisation techniques tend to be very intensive on either the human or the machine resources. In this paper, a machine learning method inspired by the Bayer filter of digital colour cameras and the demosaicing process for the colourisation of grayscale images is proposed. The proposed method involves training a multilayer perceptron model on colour images that are semantically similar to each other. The model can, henceforth, colourise grayscale images that are semantically similar to those in the training set. The success of our method is dependent on an image data representation model developed for this purpose. The proposed model gives impressive results despite requiring no human intervention and fewer machine resources for training when compared with existing deep learning models.

Adaptive Binarization of Metal Nameplate Images Using the Pixel Voting Approach

Authors: Hubert Michalak[1]; Krzysztof Okarma[1]

[1] *West Pomeranian University of Technology in Szczecin*

Corresponding Author: okarma@zut.edu.pl

In the paper, an application of the recently proposed approach to hybrid image binarization based on pixel voting is considered for industrial images. Since such images typically contain the text embossed or engraved in metal nameplates, often non-uniformly illuminated, a proper binarization of such images is usually much harder than for scanned document images, or even for the photos of text documents. Assuming that no single method would be the best solution for such images, a hybrid solution, based on the combination of multiple algorithms using pixel voting, has been recently proposed for document images. The obtained experimental results for the dataset of "industrial" images confirm the usefulness of this approach and the proposed combinations of previously developed algorithms outperform the other methods, making it possible to increase the OCR accuracy also for demanding images containing light reflections and shadows.

An Algorithm for Automatic Creation of Ground Level Maps in Two-Dimensional Top-Down Digital Role Playing Games

Authors: Dariusz Frejlichowski[1]; Krzysztof Kaczmarzyk[1]

[1] *West Pomeranian University of Technology, Szczecin*

Modern digital Role Playing Games are often played on some sort of a map. Creating those maps by artists can take a long time. In this paper we propose a solution for automatic creation of such maps, that can then be used as a support tool for creating a baseline map to work on by artists, or as an automatic map generator. The general approach of creating maps using our solution is briefly explained, and then a closer look at the ground part of the map creation process is taken. Our algorithm takes as an input a list of tiles to be used and a rough sketch of a map to be generated, prepared with specific colours.

It then creates a tile grid from that sketch using simple downscaling. That tile grid is then corrected using methods proposed in this paper, so that it does not have sharp edges and mismatched tiles. Finally, using tiles from input, the grid is transformed into an output map that closely resembles the sketch in its structure. Our solution is tested in practice by means of several exemplary sketches, each accompanied with comment explaining the significance of obtained result.

Carotid artery wall segmentation in ultrasound image sequences using a deep convolutional neural network

Authors: Guillaume Zahnd[1]; Hervé Liebgott[2]; Nolann Lainé[2]; Maciej Orkisz[2]

[1] *Institute of Biological and Medical Imaging, Helmholtz Zentrum München, Neuherberg, Germany*

[2] *CREATIS, Université Lyon 1*

Intima-media thickness (IMT) of the common carotid artery is routinely measured in ultrasound (US) images and its increase is a marker of pathological changes due to atherosclerosis.
As manual measurement is subject to substantial inter- and intra-observer variability, automated methods have been proposed to find the contours of the intima-media complex (IMC) and to deduce IMT from the distance between them.
Most of these methods explicitly seek smooth curves passing through points with strong gradients expected between artery lumen and intima, and beetween media and adventitia layers.
These assumptions may not hold depending on image quality and arterial wall morphology.
We therefore have relaxed these explicit assuptions and developed a region-based segmentation method that learns the appearence of the IMC from data annotated by human experts.
This deep-learning method uses the dilated U-net architecture and proceeds in two steps.
First, the shape and location of the arterial wall of interest are identified in full-image-height patches using the original image resolution.
Then, the actual segmentation of the IMC is performed in multimple patches distributed around thus identified location to cope with the morphological variability of the arteries. This step uses a finer spatial resolution to achieve sub-pixel accuracy.
Eventually, the predictions from these patches are combined by majority voting and the contours of the segmented region are extracted.
In a recently published comparative study using a large common publicly-available database this method outperformed state-of-the-art algorithms in accuracy and robustness, but not in processing time (>2s).
Nevertheless, the details of the method have not yet been published in a peer-reviewed support.
Here we describe the methodological details and report main results. We also show that satisfactory processing time, four times shorter, can be achieved by modifying the parameter setting of the method without degrading the accuracy.

Digital Wah-Wah guitar effect controlled by mouth movements

Authors: Adam Nowosielski[1]; Przemysław Reginia[1]

[1] *West Pomeranian University of Technology, Szczecin*

Corresponding Author: anowosielski@wi.zut.edu.pl

The Wah-Wah is a guitar effect used to modulate the sound while playing. This is an unusual effect in that the guitar player, having his hands on instrument, controls it in real time with the foot. The digital equivalent proposed in this paper transfers this control to mouth movements by capturing an image from a computer camera and then applying computer vision algorithms. The paper analyzes the applicability and studies the effectiveness of using mouth movement to control a Wah-Wah type guitar effect.

Energy Efficient Hardware Acceleration of Neural Networks with Power-of-Two Quantisation

Authors: Dominika Przewlocka-Rus[1]; Tomasz Kryjak[1]

[1] *AGH University of Science and Technology, Kraków*

Deep neural networks virtually dominate the domain of most modern vision systems, providing high performance at a cost of increased computational complexity. Since for those systems it is often required to operate both in real-time and with minimal energy consumption (e.g., for wearable devices or autonomous vehicles, edge Internet of Things (IoT), sensor networks), various network optimisation techniques are used, e.g., quantisation, pruning, or dedicated lightweight architectures. Due to the logarithmic distribution of weights in neural network layers, a method providing high performance with significant reduction in computational precision (for 4-bit weights and less) is the Power-of-Two (PoT) quantisation (and therefore also with a logarithmic distribution). This method introduces additional possibilities of replacing the typical for neural networks Multiply and ACcumulate (MAC - performing, e.g., convolution operations) units, with more energy-efficient Bitshift and ACcumulate (BAC). In this paper, we show that a hardware neural network accelerator with PoT weights implemented on the Zynq UltraScale + MPSoC ZCU104 SoC FPGA can be at least $1.4x$ more energy efficient than the uniform quantisation version. To further reduce the actual power requirement by omitting part of the computation for zero weights, we also propose a new pruning method adapted to logarithmic quantisation.

Error analysis and graphical evidence of randomness in two methods of color visual cryptography

Authors: Leszek Chmielewski[1]; Mariusz Nieniewski[2]; Arkadiusz Orłowski[1]

[1] *Warsaw University of Life Sciences – SGGW*

[2] *University of Lodz*

Corresponding Author: leszek_chmielewski@sggw.edu.pl

Analysis of errors in two methods of color visual cryptography with random shares introduced by us in previous publications was performed. In both methods the shares are random, and consequently, errors occur in the decoded image. In one of the methods, where the coding is done by unhiding the pixels, there are two types of errors: missing color errors and hiding failure errors. In the other method, where the coding is done by unhiding, only the missing color errors occur. Probabilities of the missing color errors were modelled mathematically and frequencies of two types of errors were tested experimentally for both methods. Tests demonstrate that the model is correct. The results show in which cases the considered methods exhibit more accurate decoding results. Also, the extended results of randomness tests conducted with the NIST randomness test suite on the results of coding for a set of typical benchmark images are presented in the form of histograms of p-values. These graphical results indicate that the shares are indeed truly random.

Fast Triangle Strip Generation and Tunneling for Different Cost Metrics

Authors: Jonas Treumer[1]; Lorenzo Neumann[1]; Ben Lorenz[1]

[1] *Technische Universität Bergakademie Freiberg*

Triangle strips provide a memory-efficient representation of triangle meshes. In the ideal case, they allow to encode a mesh of n faces with n + 2 vertices. All modern graphics interfaces offer a primitive type for triangle strips to disburden the host/device bottleneck and save graphics memory. Although modern 3D rendering pipelines mainly utilize triangle lists, strips still offer significant performance benefits in two-dimensional applications, e.g. the visualization of polygons in orthographic maps.

Encoding a regular mesh into an optimal triangle strip representation is an NP-complete problem. Therefore, multiple heuristic techniques for different cost metrics have been proposed and implemented. In this paper, we provide an overview of related approaches and propose a unified cost model to rate their quality. Furthermore, we introduce a novel, flexible implementation for triangle strip generation. It extends an established key technique, the tunneling operator, with a new algorithm for circle detection. A concluding benchmark compares our code with existing solutions and outlines its superior performance in most use cases.

Fuzzy approach to object-detection-based image retrieval

Authors: Marcin Iwanowski[1]; Aleksei Haidukievich[1]; Bartosz Wnorowski[1]; Maciej Leszczyński[1]

[1] *Warsaw University of Technology*

In the paper, a method for image search is proposed. It allows for finding images starting from a text query that contains names of object classes and their expected spatial relations. It is based on image object detection and fuzzy scene description. The input query containing the description of desired objects and their spatial relations are used to select and score relevant images. The score is next used to order the output images putting more relevant ones on the top of the list. The proposed approach is based on object detection methods and a fuzzy approach to describe the position of objects in relation to other ones. The method may be used to retrieve images from databases containing images with metadata produced by object detectors.

Influence of Step Parameterisation on the Results of the Reidentification Pipeline

Authors: Damian Pęszor[1,2]; Konrad Wojciechowski[1,2]; Łukasz Czarnecki[3]

[1] *Silesian University of Technology*

[2] *Polish-Japanese Academy of Information Technology*

[3] *Kar Tel Sp. z o.o. Spółka Komandytowa*

In this paper, research on the influence of parameters' values in the pipelines of facial-based reidentification systems is presented. It was assumed that the solution should operate in real time, in conditions typical for the reidentification system to be employed. Such conditions were obtained as part of research regarding the re-identification of aggressively acting people during sports events. Typically, such a pipeline consists of many steps, including facial region detection, frontalisation, embedding, and classification, which are usually evaluated separately. This paper focuses on the parameters of facial alignment and classification in the context of systems based on well-established solutions based on Multi-task Cascaded Convolutional Networks coupled with Inception Resnet embedding. The authors propose evaluating the results of the entire pipeline as a way to identify the optimal set of parameters for each step, thus producing a pipeline where the subsequent steps are best fitted to each other rather than giving the best results on their own.
The results indicate that the correct selection of parameters of the steps of the pipeline depends on further steps used and vice versa. It is therefore suboptimal to select parameters based on a separately evaluated set of steps, as it is usually presented in the literature. The reidentification pipeline must therefore be evaluated as a whole, disregarding the results achieved by any single part of the pipeline, as they are not an indicator of overall system performance.

Novel Co-SIFT Descriptor for Scanned Images Differentiation

Authors: Paula Stancelova[1]; Zuzana Cernekova[1]; Andrej Ferko [1]

[1] *Comenius University Bratislava, Slovakia*

Corresponding Authors: zuzana.cernekova@uniba.sk, paula.stancelova@fmph.uniba.sk

We describe experiments with the hi-tech contactless scanner CRUSE CS 220ST1100 in the digitization of originals of natural and cultural heritage. The 2D scans guarantee high accuracy both in geometry and radiometry (48 bits for RGB colors). However, an inexperienced customer needs support in selecting the appropriate scan mode. To distinguish similar CRUSE scans, we proposed an image descriptor based on Harris corners and a topological structure embedding a planar subgraph. For some use-cases, the Harris approach did not perform well. We report on a novel SIFT type detector using concurrent color channels, hence the proposed name. We put our solution into the context of previous research and compare, on selected use-cases, the solution quality and/or disadvantages.

On Formal Models of Interactions between Detectors and Trackers in Crowd Analysis Tasks

Authors: Andrzej Śluzek[1]; M. Sami Zitouni[2]

[1] *Warsaw University of Life Sciences – SGGW*

[2] *Khalifa University of Science and Technology*

Corresponding Author: andrzej_sluzek@sggw.edu.pl

In crowd analysis tasks (regardless the crowd nature, e.g. humans, cattle, birds, insects or drones) the low-level vision tools remain the same, i.e. detection and tracking of targets (i.e. either individuals or groups). The required results of analysis are, however, more complicated (e.g. patterns of group splitting/merging, changes in group sizes and membership, group formation and disappearance, etc.). For completing those tasks, raw-data results of detectors/trackers are converted into data associations representing crowd structure/evolution.
Standard data associations in this area are based on target labeling (i.e. they are deterministic) while performances of even *state-of-the-art* detectors/trackers are non-perfect, which effectively makes such results non-deterministic.
In this paper, we discuss mathematical models of interactions between (possibly multiple) detectors and (possibly multiple) trackers so that data associations can be represented non-deterministically (in the matrix form) and further processed in a similar way. Then, substitutions/switches between algorithms can be formally performed to maximize the overall accuracy of the label-based associations, since such associations are eventually needed as the final outcomes of the crowd analysis.
Apart from mathematical details, the paper presents examples (both synthesized and derived from real visual data) illustrating feasibility (and advantages) of the presented approach.
The paper is a continuation, extension and (in many aspects) improvement of our recent results (March 2022) published in a journal paper (Int. J. AMCS).

On multi-stream classification of two person interactions in video with skeleton-based features

Authors: Paweł Piwowarski[1]; Sebastian Puchała[1]; Włodzimierz Kasprzak[1]

[1] *Warsaw University of Technology*

We propose two methods based on skeleto data and deep neural networks for two-person interaction classification in video clips. These are two different ideas of using multi-stream networks. The first method is an ensemble of *weak* pose classifiers, where every classifier is trained on a different time-phase of an interaction, while the overall classification result is a weighted combination of their

results. An advantage is the simplicity of models and of their learning process, while a disadvantage is in the need first to generate explicit motion-relevant information. In the second approach, there are three feature streams created from the skeleton data and they fed a tripple-stream of LSTM networks for classification. In a feature preprocessing step, the skeletons are tracked over time allowing to correct the skeleton data, i.e. to properly reassign the same skeleton to appropriate person and to approximate the missing joints. Contrary to previous method, no explicit motion data need to be included in the feature vector generated from the tracked skeletons, as the classifier is based on LSTM networks. Both models were learned, optimized and evaluated on the interaction subset of the NTU RGB+D data set, showing comparable performance with the best reported CNN- and GCNN-based classifiers for this dataset.

On the Influence of Image Features on Performance of Deep Learning Models in Human-Object Interaction Detection

Authors: Grzegorz Sarwas[1]; Marcin Grząbka[1]; Marcin Iwanowski[1]

[1] *Warsaw University of Technology*

In recent years Human Object Interaction (HOI) detection has experienced rapid performance growth mainly due to the development of various deep learning-based methods and algorithms. One of the most popular approaches to this task is three-stream architecture consisting of three processing paths: human-stream, object-stream, and relation-stream. They all gather information on the scene using features extracted from particular image elements. The total amount of features depends on the detailed processing schemes and applied hyperparameters. Their number may vary from tens of thousands to hundreds of millions. Still, the increase in the number of features does not necessarily increase the final efficiency of the HOI detection method. This paper focuses on the relation between the number of features used to detect HOI and its predictive efficiency. We investigate the influence of the quality and quantity of features on the final detection results. Several experiments were conducted to validate various model configurations of different features considered.

PointPillars backbone Type Selection For Fast and Accurate LiDAR Object Detection

Authors: Konrad Lis[1]; Tomasz Kryjak[1]

[1] *AGH University of Science and Technology, Kraków*

3D object detection from LiDAR sensor data is an important topic in the context of autonomous cars and drones. In this paper, we present the results of experiments on the impact of backbone selection of a deep convolutional neural network on detection accuracy and computation speed. We chose the PointPillars network, which is characterised by a simple architecture, high speed, and modularity that allows for easy expansion. During the experiments, we paid particular attention to the change in detection efficiency (measured by the mAP metric) and the total number of multiply-addition operations needed to process one point cloud. We tested 10 different convolutional neural network architectures that are widely used in image-based detection problems. For a backbone like MobilenetV1, we obtained an almost 4x speedup at the cost of a 1.13% decrease in mAP. On the other hand, for CSPDarknet we got an acceleration of more than 1.5x at an increase in AP of 0.33%. We have thus demonstrated that it is possible to significantly speed up a 3D object detector in LiDAR point clouds with a small decrease in detection efficiency.
This result can be used when PointPillars or similar algorithms are implemented in embedded systems, including SoC FPGAs. The code will be made available after approval of the paper.

Real time intersection queue length estimation system

Author: Kamil Bolek[1]

[1] *Polish-Japanese Academy of Information Technology, Warsaw*

Many systems which check queue length at intersections are inaccurate. Information about small/-medium/large queues are not enough for intelligent transport systems which could modify an actual intersection program in a way which may allow for optimising the traffic flow in the city. Not only does the presented system provide information about queue length in metres on every lane, but also it is based solely on camera focal length, sensor size and number of lanes in camera view, which minimizes the involvement of traffic operators in time-consuming camera setup. The system consists of several submodules, the first of which detects license plates and uses them to create a configuration of the camera. Subsequently, the second module that performs detection of the type of vehicle can determine what the current length of the queue on every lane is. All this information is sent to the traffic management system in Wroclaw which modifies the traffic lights controller programs, optimising the traffic flow in the city.

Traffic Sign Detection Using Deep and Quantum Nerual Networks

Authors: Sylwia Kuros[1]; Tomasz Kryjak[1]

[1] *AGH University of Science and Technology, Kraków*

Quantum neural networks (QNN) are an emerging technology which can be used in many application, also in computer vision. In this paper we presented a traffic sign detection system implemented using a hybrid convolutional and quantum neural network. Experiments on the German Traffic Sign Recognition Benchmark dataset indicate that currently QNN do not outperform “pure” DCNN (Deep Convolutuional Neural Networks), yet sill provide an accuracy about over 90% and are a definitely promising solution for advanced computer vision.